European Commission

JRC SCIENCE FOR POLICY REPORT

# Technology and Democracy

## Understanding the influence of online technologies on political behaviour and decision-making

2020

**Coordinating lead authors:**
Stephan Lewandowsky
Laura Smillie

**Lead authors:**
David Garcia
Ralph Hertwig
Jim Weatherall
Stefanie Egidy
Ronald E. Robertson

**Contributing authors:**
Cailin O'Connor
Anastasia Kozyreva
Philipp Lorenz-Spreen
Yannic Blaschke
Mark Leiser

Joint Research Centre

EUR 30422 EN

This publication is a Science for Policy report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Abstract
Drawing from many disciplines, the report adopts a behavioural psychology perspective to argue that "social media changes people's political behaviour". Four pressure points are identified and analysed in detail: the attention economy; choice architectures; algorithmic content curation; and mis/disinformation. Policy implications are outlined in detail.

Acknowledgements
Grateful thanks are extended to the following external reviewers: Profs. Chris Bail, Ullrich Ecker, Dan Larhammar, Sune Lehmann, Jason Reifler and Jane Suiter as well as the following internal JRC reviewers: Gianluca Misuraca, Paulo Rosa, Mario Scharfbillig and Lucia Vesnic Alujevic. Grateful thanks are also extended to the stakeholder participants who joined the foresight workshop on 11 March 2020.

Manuscript completed in August 2020

Contact information
Name: Laura Smillie
European Commission, Joint Research Centre, Brussels - Belgium
Email: JRC-ENLIGHTENMENT2@ec.europa.eu

EU Science Hub
https://ec.europa.eu/jrc

JRC122023
EUR 30422 EN

| | | | |
|---|---|---|---|
| PDF | ISBN 978-92-76-24088-4 | ISSN 1831-9424 | doi: 10.2760/709177 |
| Print | ISBN 978-92-76-24089-1 | ISSN 1018-5593 | doi: 10.2760/593478 |

Luxembourg: Publications Office of the European Union, 2020

How to cite this report: Lewandowsky, S., Smillie, L., Garcia, D., Hertwig, R., Weatherall, J., Egidy, S., Robertson, R.E., O'connor, C., Kozyreva, A., Lorenz-Spreen, P., Blaschke, Y. and Leiser, M., Technology and Democracy: Understanding the influence of online technologies on political behaviour and decision-making, EUR 30422 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-24088-4, doi:10.2760/709177, JRC122023.

# Preface

This report is the second output from the Joint Research Centre's (JRC) *Enlightenment 2.0* multi-annual research programme. The work started with the classical Enlightenment premise that reason is the primary source of political authority and legitimacy. Recognising that advances in behavioural, decision and social sciences demonstrate that we are not purely rational beings, we sought to understand the other drivers that influence political decision-making. The first output "*Understanding our political nature: how to put knowledge and reason at the heart of policymaking*" published in 2019[1], addressed some of the most pressing political issues of our age. However, some areas that we consider crucial to providing an updated scientific model of the drivers of political decision-making were not fully addressed. One of them is the impact of our contemporary digital information space on the socio-psychological mechanisms of opinion formation, decision-making and political behaviour.

The JRC, together with a team of renowned experts addresses this knowledge deficit in a report that synthesises the knowledge about digital technology, democracy and human behaviour to enable policymakers to safeguard a participatory and democratic European future through legislation that aligns with human thinking and behaviour in a digital context. It is hoped that this report will prove useful as policymakers reflect upon the forthcoming European Democracy Action Plan, the Digital Services Act, the EU Citizenship Report 2020, as well as on how to legislate against disinformation.

The report has been written in spring/summer of 2020 when the COVID-19 pandemic took hold of Europe and the world. During this time, our democracies suffered while technology played a crucial role in keeping societies functioning in times of lockdown. From remote distance education to teleworking, religious services to staying in touch with family and friends, for many but not all, everyday activities moved online. Additionally, technological applications and initiatives multiplied in an attempt to limit the spread of the disease, treat patients and facilitate the tasks of overworked essential personnel.

Conversely, however, significant fundamental rights questions have been raised as unprecedented initiatives to track, trace and contain the pandemic using digital technologies have proven controversial. Governments invoking emergency measures in support of public health decision-making, used advanced analytics to collect, process and share data for effective front-line responses that lacked transparency and public consultation.

When used as an information source, social media have been found to present a health risk that is partly due to their role as disseminators of health-related conspiracies, with non-English language speakers being at greater risk of exposure to misinformation during the crisis. It is likely that these technologies will have a long-lasting impact beyond COVID-19. Yet despite the immediacy of the crisis, the authors invite the reader to take a longer perspective on technology and democracy to get a deeper understanding of the interrelated nuances. In dark times, we seek to bring light to the importance of understanding the influence of online technologies on political behaviour and decision-making.

[1] https://ec.europa.eu/jrc/en/facts4eufuture/understanding-our-political-nature

## Executive summary

The historical foundation of the European Union lies in the ideal of democracy as a mode of governing social, political and economic relations across European states with the objective of ensuring peace. This has led to an unprecedented period of peace across the Union.

Yet today some of the institutions, norms and rules that underpin this structure are experiencing major pressure. **A functioning democracy depends on the ability of its citizens to make informed decisions**. Open discussions based on a plurality of opinions are crucial; however, the digital information sphere, which is controlled by few actors without much oversight, is bringing new information challenges that silently shape and restrict debate.

In terms of understanding the online environment, **there are three key vectors that deserve consideration by policymakers: actors, content and behaviours.** For the most part, ongoing policy reflections have concentrated on understanding the actors and the nature of content. In the absence of behavioural reflections, policymakers may feel that they are constantly playing catch-up with technological advances. Taking a behavioural approach, this report seeks to help policymakers regain agency. **Essential components of human behaviour are governed by relatively stable principles that remain largely static even as the technological environment changes rapidly.**

Before getting into the details, we provide an answer to the basic question **"Do we behave differently online? If so, why?"** The web is cognitively unique, resulting in specific psychological responses to its structure and functionality and differences in perception and behaviour. Structural factors in the design of online environments can affect how individuals process information and communicate with one another. **Importantly, there is scientific evidence that social media changes people's political behaviour offline; this includes the incitement of dangerous behaviours such as hate crimes.**

Based upon an in-depth scientific analysis, four pressure points are identified that emerge when people and the online environment are brought into contact without much public oversight or democratic governance: i) Attention economy; ii) Choice Architectures; iii) Algorithmic content curation; iv) Misinformation and disinformation. Each pressure point is tackled in terms of its specific characteristics and how it affects behaviour. A dedicated chapter looks at the implications for policy.

***Attention economy*** — human behaviour unfolds online in an economy in which human attention is the predominant commodity. The digital sphere is designed so that people give their valuable resources of time, attention and data without considering the costs — for themselves and others. This exploits certain features of human behaviour, which makes it hard to address at the individual level. On a societal level, coordination is needed to assure privacy and autonomy as a public good, otherwise there are deep conflicts with the principles of democracy, freedom and equality.

Social media poses a risk for the fundamental rights to data protection and privacy of users, and even for non-users that extends far beyond what individuals explicitly share with social media sites, because of how much can be inferred from users' activity. **Ensuring online privacy preserves three core components of democratically empowered voters: freedom of association, truth-finding and opportunities to discover new perspectives**. Effective privacy online means a strengthened democracy offline.

The effects of highly personalised advertisements directed at users based on personal behavioural characteristics — the practice referred to as microtargeting — are nuanced and difficult to assess. However, there is enough evidence of (at least potential) harm to concern policymakers. **The microtargeting of political messages has considerable potential to undermine democratic discourse — a foundation of democratic choice.** Furthermore, research shows that the public are opposed to microtargeting about certain content (including political advertising) or based on certain sensitive attributes (including political affiliation).

Despite ongoing discussions about further online regulation, the web experience is uniquely subjective and largely influenced by the algorithms of private actors designed to maximise profits by capturing our attention without any public accountability. Consequently, **business models prevalent in today's online economy constrain the solutions that are achievable without regulatory intervention**.

***Choice architectures*** — are an important determinant of online behaviour. **Companies use defaults, framing and dark patterns to shape user behaviour**. These prompt lenient privacy settings to increase user engagement. These design features **limit freedom of association, truth-finding, opportunities to discover new perspectives, creating challenges for democratic discourse and the autonomous formation of political preferences**.

Importantly, users are generally unfamiliar with what data they produce, provide to others and how that data is collected and stored when they perform basic tasks on social media platforms.

***Algorithmic content curation*** — algorithms are an indispensable aspect of digital technologies which can be used or abused to impact user satisfaction, engagement, political views and awareness. **Curated newsfeeds and automated recommender systems are designed to maximize user attention by satisfying their presumed preferences, which can mean highlighting polarising, misleading, extremist or otherwise problematic content to maximize user engagement.** The ranking of content — including political messages — in newsfeeds, search engine ordering and recommender systems can causally influence our preferences and perceptions.

While the evidence on filter bubbles is ambiguous, there is legitimacy to the societal concerns raised about echo chambers. Scientific findings suggest that there is an ideological asymmetry in the prevalence of echo chambers, with people on the populist right being more likely to consume and share untrustworthy information.

***Misinformation and disinformation*** — misinformation generally makes up a small fraction of the average person's "media diet", but some demographics are disproportionately susceptible (advanced age, some cognitive attributes). The problem of misleading online content extends far beyond strict "fake news" and when misleading content is considered in its entirety, the problem is extensive and concerning.

Two core attributes from the attention economy and human psychology create the perfect conditions for the spread of misinformation: **algorithms that promote attractive, engaging content and people's strong predisposition to orient towards negative news, as most "fake news" tends to evoke negative emotions such as fear, anger and outrage.**

**The shape and spread of misinformation is governed by social media network structures**; they can give rise to significant distortions in perceived social signals that in turn can affect the entrenchment of attitudes.

There are asymmetries in how false or misleading content and genuine content spread online, with misinformation arguably spreading faster and further than true information. Some of this asymmetry is driven by emotional content and differing levels of novelty.

Related to this, the interpretation and classification of misleading content often turns on subtle issues of intent and context that are difficult for third parties — especially algorithms — to ascertain, making it difficult to distinguish legitimate political speech from illegitimate content.

***Taking democracy online*** — this chapter looks at the pros and cons of encouraging democracy online. Some self-governed online fora have been identified as contributing to radicalisation and toxic extremism. **Secluded online spaces can function as laboratories that develop extremist talking points that then find entry into the mainstream**. Importantly, however, **online spaces can also provide voices to marginalised and disadvantaged communities**.

Current social media platform architectures are not primarily designed for democratic discourse, yet they are heavily used for political purposes and debates. The platforms may, for example, provide social signals that can lead to misperceptions about relative group sizes. This has consequences for social movements who can come to believe that their ideas have broader penetration than they actually do.

Importantly, government-supported platforms have been shown to allow large-scale public consultation with existing research in online deliberative spaces suggesting that **when properly designed and managed well, online deliberation may match the success of offline deliberative processes**.

***What does this mean for policy?*** — this chapter translates the impact of the four pressure points into implications for policymakers. Given the integrated nature of these pressure points, it is not meaningful to recommend individual policy actions. Instead, the three fundamental democratic principles of equality, representation and participation are used as a framework to shape the proposals formulated in this chapter.

***Future Research Agenda*** — of all current and future human behaviours, online political behaviours are perhaps the most important ones for our collective future. However, a mix of platform reticence and a lack of regulatory clarity have hampered a full scientific understanding of these behaviours. This chapter proposes a collectively operated, publicly funded European alternative to commercial platforms that would see the research community and citizens jointly pursue a research agenda to understand the influence of digital technology on democracy.

## Table of contents

In the 5 minutes it took you to get to this page… There have been 20 million Google searches, 6.5 million Facebook logins, 95 million WhatsApp messages sent, 12.5 million Snaps created on Snapchat, 23.5 million videos viewed on YouTube, 8 million Tinder swipes, 1 million Tweets tweeted, 7,000 TikTok downloads and 3.5 million Instagrams scrolled. That's a lot of posting, swiping, tweeting, scrolling, liking, sharing, downloading, viewing and snapping but what does it all mean? Who controls that data, what are they doing with it, and by what authority?

# Chapter 1

Introduction

Methodology

Understanding the basics: Cognition in context

# Chapter 1: Introduction

The historical foundation of the European Union lies in ensuring peace in Europe by means of democracy as the ideal way of governing social, political and economic relations across the Union. This ideal has been put into practice within and across Member States through a set of institutions, norms, rights and rules that have regulated the relationship of trust and legitimacy between governments and citizens, giving rise to democracy as arguably one of the most stable forms of political system and collective living.

Yet today some of these institutions, norms and rules are witnessing major pressure to keep apace with the evolving character of societies as well as with their ways of constituting themselves as a political community. A functioning democracy empowered by fundamental rights depends on the ability of its citizens to make informed decisions. Open discussions based on a plurality of opinions are crucial to identify the best arguments, exchange diverse viewpoints and build consensus. Therefore, freedom of discussing and exchanging ideas is of essential importance.

However, the digital information space is bringing new challenges on a different level. In an online "marketplace of ideas" [1], where attention is limited and information is sorted by algorithms developed by powerful platforms, there is a deeper power structure shaping and restricting debate. Online platforms allow and enable the marketplace of ideas to fail, for example through interference in democratic processes and elections or other votes. This threatens to manipulate the opinion formation upon which democracy depends and exerts undue influence on democratic decision-making. Of course, biased forces have always tried to influence political decision-making in pursuit of their own interests. But today, the affordability of online communication, its lack of transparency as well as the scope and gravity of influence take a much more threatening form. In particular, the digital sphere offers tools that make targeted manipulation on a global scale very easy, without offering any transparency, meaningful regulation of the actors in the advertising ecosystem or insights into the underlying proprietary processes.

In terms of understanding the online environment, there are three key vectors that deserve regulatory consideration; actors, content and behaviours. For the most part, ongoing policy reflections have concentrated on understanding the actors and the nature of content. In the absence of behavioural reflections, policymakers may feel that they are constantly playing catch-up with technological advances. Taking a behavioural approach, this report seeks to reduce such uncertainties as — notwithstanding its variability and diversity — human behaviour is governed by stable principles that remain relatively unchanged even as the actors, contents and environments may change rapidly. Even though people adapt easily to new contexts and environments, that adaptation involves relatively stable cognitive processes that scientists are beginning to understand well.

> *"This is a coalition of democracies founded on the principle of freedom. That is our bastion, that is our platform, that is our struggle."*
>
> — Alcide de Gasperi, 1952

So can ever-evolving digital technologies be regulated? If so how and why? Is there proof that we behave differently online from offline? While mindful of the rights-based society in which we live, how can the regulatory toolbox be strengthened to reduce the chance of minor technical tweaks (e.g. Facebook adding different reaction emojis other than the ‘like' function) having large unanticipated consequences at a societal level?

These are just some of the questions EU policymakers wanted answers to when they were approached to discuss the scope of this report that subsequently determined the parameters of the scientific literature review. This report is therefore not a “systematic review”, but it responds systematically to the scoping questions put to the authors by the European Commission.

The influence of the digital world can only be understood by joint consideration of behaviour and cognition on the one hand and the full range of socio-political, philosophical, economic, regulatory and design contexts in which it unfolds on the other. This interdisciplinary report recognises and explores this tension at all levels of analysis; from the macro level of the “attention economy” and how it shapes global streams of human behaviour, to the micro level of the design of newsfeeds and defaults and how they affect cognition in the moment.

The solutions offered in this report will draw on the recognition that human cognition while inextricably tied to context, is also governed by stable principles that remain largely unchanged even as the technological environment changes rapidly. Understanding those principles and how they are leveraged by context will enable policymakers to strengthen the regulatory toolbox with instruments that can transcend changes in technology.

Despite substantial legislation already applying to the online world and several regulatory initiatives currently taking shape at the European level, this report is intended to help policymakers identify frameworks and policies that can remain meaningful in a rapidly changing world.

In this context, the European Commission's Joint Research Centre (JRC) also recognised the importance of incorporating a strategic foresight element into this work to enable policymakers to “see” and consider alternative possible futures. The annex of the report presents different scenarios for the “European information space in 2035”, created in collaboration with a broad range of stakeholders, from industry to civil rights groups, from academia to media regulators and policymakers.

## Methodology

This report is a state-of-the-science review based upon a solid interdisciplinary critical analysis and a synthesis of the relevant peer-reviewed scientific literature.

As the European Commission's knowledge and science service, the JRC used innovative knowledge brokerage techniques to produce this report; embedding European Commission staff in a team of international scientific experts spanning different disciplines. Renowned cognitive psychologists and philosophers who contributed to the first study under the Enlightenment 2.0 multi-annual research programme were joined by specialists from the fields of Complexity Science, Computational Social Science, Constitutional Law, Fundamental Rights, Mathematics and Network Science as well as specialists in the ethical and societal implications of Artificial Intelligence.

The report is firmly embedded in two principles of enquiry:

- First, the authors are committed to the idea that truth is not just a construct in the eye of the beholder but something that exists independently and that should, in democratic societies, be a common goal of political debate[2]; and

- Second, the report is based on the balance of evidence rather than the balance of opinions and the report foregrounds evidence irrespective of whether it aligns with a preferred balance of opinions.

Where normative judgements were required, the experts used the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities, as laid down in Article 2 of the Treaty on European Union, to guide all recommendations.

Despite the thoroughness of the scientific review herein, the authors acknowledge three important methodological considerations:

> *"If you want to have the right balance of governance measures, you need to have very clear and strong values and in Europe we have these values. If you understand how we are building our continent on these values, you understand how you need to behave."*
>
> — Thierry Breton, European Commissioner for the Internal Market in a live-streamed debate with Mark Zuckerberg, Facebook CEO
>
> 18 May 2020

[2] United States, United Kingdom, Canada, Germany and Australia; https://www.scimagojr.com/countryrank.php?category=3201

1. Although human cognition is studied the world over, the fields of behavioural science and psychology are disproportionately Anglophone. Of the top five countries in psychological research, four are either exclusively or predominantly Anglophone and none of those four are members of the EU. This imbalance is necessarily reflected in this report and it must be acknowledged. Fortunately, although cognition is remarkably flexible and adapts to the prevailing context, within western industrialized nations its basic principles have been found to be largely invariant. People's basic cognitive apparatus in Canada or the US does not differ qualitatively from that of people in Finland or Italy. Moreover, although a large share of new technology emerges from Silicon Valley, those new modes of interacting and communicating almost invariably find global penetration [3]. From July 2019 to July 2020, 98.5% of social media use in the EU was on 5 platforms, all of which are American (Facebook: 75.66%; Pinterest: 8.78%; Twitter: 7.61%; Instagram: 4.47%; YouTube: 1.14%).[3]

   The reliance on non-European sources therefore does not undermine the significance of the findings outlined in this report. However, in light of the possibility that European and American cultures may continue to drift further apart, this reliance on non-European research in a culturally-sensitive arena is not sustainable. The report therefore concludes with a strong call for further European research into cognition within its cultural setting (Chapter 9).

2. The report does not address the wider context of the contemporary European political landscape but instead distils — in as much as is it is meaningful and possible — the specific digital layer added by information technology to previously existing means of exerting political influence. This approach does not mean that we assume this digital layer to exist in isolation from offline communication, traditional media or larger societal trends.

3. The report mainly focuses on human political behaviour online. Although we touch on automated processes, algorithmic decision-making and artificial intelligence, we mainly exclude from consideration non-authentic or non-human actors such as "bots", "avatars" and "sock-puppets", which are polluting the information landscape with manipulative messages on behalf of hidden political interests. Although these artificial entities play an influential role online [4, 5, 6], their control is a matter of cybersecurity rather than understanding human cognition online. The European Commission's recent report on Cybersecurity[4] addresses those threats. Additionally, the JRC's report "Artificial Intelligence: A European perspective" provides many different perspectives of the developing technology and its possible impact in the future[5]. This report touches on artificial entities only when they have unique cognitive or behavioural implications.

[3] https://gs.statcounter.com/social-media-stats/all/europe

[4] https://ec.europa.eu/jrc/en/news/put-cybersecurity-at-centre-of-society

[5] https://ec.europa.eu/jrc/en/publication/artificial-intelligence-european-perspective

## Understanding the basics: Cognition in context

Human cognition is context dependent. No decision is ever made in an informational void.

- When people make decisions about matters of money, health or entertainment, they are considerably more likely to accept preselected choice options, so-called “defaults” [7].

- When shopping online, we are more likely to click on items at the beginning of a list of options or at the very end, irrespective of other aspects of our preferences [8].

When the context changes, decisions change. For example, people’s support for climate mitigation policies increases considerably if identical economic consequences are presented as a foregone gain (reduction in future wealth increases) than a loss (reduction in wealth) [9].

Calls for greater “media literacy” or “critical thinking” are, by themselves, therefore likely to be insufficient to counteract any adverse effects on democracy from political online behaviour. Context matters and it can override people’s best intentions, in particular in a rapidly changing environment where existing skills may rapidly become obsolete.

Nevertheless, humans are not absolute slaves to their environment. They can be “boosted” to exercise their own agency in specific contexts and they can become more skilled consumers of information [10]. However, even though people can be empowered to become better decision-makers, in many cases boosting cannot be achieved without relying on platforms to provide the (informational) basis and not to distract.

To understand digital influence, we must explore the tension between context and cognition at all levels of analysis, from the macro level of the “attention economy” and how it shapes global streams of human behaviour, to the micro level of the design of newsfeeds and defaults and how they affect cognition in the moment.

### Levels of context: The macro context

At the broadest level, we must recognize that we live in an attention economy [11] in which competition is becoming increasingly fierce. Whenever we venture online, our attention is a precious commodity that platforms vie for in pursuit of profit. We pay for a “free” service online by selling our attention and personal data to advertisers. At present, the attention economy is the inescapable driving-force of online behaviour and no understanding of the influence of online technologies on political decision-making is possible without appreciation of this context.

We must also recognise that most of the information we consume is presented to us shaped and curated by algorithms whose design and operations are proprietary and not subject to public scrutiny. Every newsfeed and every search result represents output from algorithms that, ultimately, are designed to satisfy the demands of the attention economy. This creates an inherent asymmetry in the power of platforms and citizens: while the platforms know much about their users—and even people who are not on their platforms— and deploy that knowledge to shape our information diets, citizens know little about what data the platforms hold and how they are used to customise our online experience.

### Levels of context: The micro context

At the micro level, seemingly trivial platform features can have far-reaching consequences. To illustrate, in India in 2018, false rumours about child kidnappers shared via WhatsApp's unlimited forward facility were implicated in at least 16 mob lynchings, leading to the deaths of 29 innocent people [12]. The power that digital architectures have to shape individual actions and to turn those actions into collective behaviours, has an important corollary: The converse also holds and minor technological revisions can result in significant collective behaviour changes. For instance, curtailing the number of times a message can be forwarded on WhatsApp (thereby slowing large cascades of messages) may have contributed to the absence of lynch killings in India since 2018 [13].

More than a decade into the online attention economy, it is difficult to imagine an online environment that is designed not to influence and manipulate, but to accurately inform citizens in the interest of civil democratic discussion. The first challenge for the future, therefore, is to imagine what a better online environment would look like. The annex to this report contains the results of a foresight exercise that describes possible alternative futures for the European Information Space in 2035. It is intended to help readers reflect in more depth about what they would consider a better future online environment.

The next six chapters of this report summarise the current state of the science of how online technologies interact with human political behaviour and decision-making.

# Chapter 2

Why do we behave differently online?

How social media can stir up hate crimes

The distinct cognitive attributes of the web

Differences in Structure and Functionality

Differences in Perception and Behaviour

## Chapter 2: Why do we behave differently online?

Technological innovations have a long history of evoking a mixture of Utopian euphoria and Dystopian fears. Socrates, for example, was deeply troubled by the detrimental consequences of writing (Plato, ca. 370 B.C.E/1997, pp. 551–552).

Some 2,000 years later, we accept that writing has redeemed itself. Heeding this lesson from history, we must not lose sight of the immense benefits of the digital revolution. Arguably, the COVID-19 pandemic did not wreak unmitigated havoc because digital technologies permitted the economy to continue to function during "lockdown."

Digital technologies, including social media, also made physical distancing more bearable because it enables friends and family to stay in touch in ways that would have been unthinkable without the web and its multitude of communication apps.

Social media has also been heralded as "liberation technology" [14], owing to its role in the "Arab Spring", the Iranian Green Wave movement of 2009 and other instances in which it mobilised the public against autocratic regimes. A review of protest movements in the United States, Spain, Turkey and Ukraine found that social media platforms (e.g. Twitter and Facebook) serve as vital tools for the coordination of collective action, mainly through spreading news about transportation, turnout, police presence, violence and so on [15]. Social media were also found to transmit emotional and motivational messages relating to protest activity [15].

> *"Your invention will enable them to hear many things without being properly taught, and they will imagine that they have come to know much while for the most part they will know nothing."*
>
> — Socrates on writing

However, at the same time, there is evidence that political behaviours — and consequently our democracies — may be adversely affected by events on the web. Some analysts have identified social media as a tool of autocrats [16], with empirical support provided by the finding that the more autocratic regimes aim to prevent an independent public sphere, the more likely they are to introduce the Internet [17]. In Western democracies, recent evidence suggests that social media can cause problematic political behaviours and developments [18, 19, 20, 21].

Establishing causality is crucial because it offers opportunity for intervention and control. If social media were found to cause social ills, then it would be legitimate to expect that a change in platform architecture might influence society's well-being. In the absence of causality, this expectation does not hold: For example, if certain people were particularly prone to express their hostilities by anti-social behaviours and by hostile engagement on social media, then any intervention targeting social media would merely prevent one expression of an underlying problem while leaving the other unaffected.

Establishing causality is, however, notoriously difficult, measurements can only establish an association or correlation but not causation. One approach to establishing causality that has gained popularity through the availability of “big data”, is known as instrumental variable analysis. The key idea of this technique is to find events in the world that are not associated with the outcome but are associated with the potential predictor variable. For example, it is unlikely that the availability of broadband internet, which is driven by considerations such as terrain and local regulations [22], would be directly associated with people’s voting behaviour. However, broadband availability would be expected to be associated with internet use. This identifies broadband availability as a good instrumental variable because it is expected to determine internet usage without affecting the outcome variable (voting behaviour in this case) directly. Thus, if the variation in internet usage that is due to broadband availability were found to predict voting behaviour, then this relationship would be identified as causal. A recent study conducted in Germany and Italy used broadband availability at the level of municipality as an instrumental variable. Reliance on the web for political information was found to predict the share of votes for populist parties [21]. In both countries, reliance on the web as a source of political information strongly predicted voting for populist but not for mainstream parties. Because this relationship was due to the variation in web use associated with broadband availability, a causal interpretation is possible.

Several recent studies have established causality in this manner, including for the role of social media in triggering ethnic hate crimes [19, 20] and the role of misinformation in voting for populist parties [23].

## How social media can stir up hate crimes

It is troubling that social media have been causally linked to hate crimes and ethnic violence by two studies that used the instrumental-variable approach. To illustrate, a recent study in Germany [20] examined the association between anti-refugee posts on the Facebook page of Germany’s far-right AfD party and hate crimes against refugees at the level of municipalities. The analysis revealed a strong relationship between the number of online posts and attacks on refugees. Municipalities with AfD Facebook users were three times as likely to experience refugee attacks than municipalities without. This association alone, however, would not warrant a causal interpretation for the reasons mentioned earlier. To isolate the *causal* effect of social media posts on hate crimes, local internet and Facebook outages were used as the instrumental variable. The association between Facebook posts and attacks was found to disappear in localities in which outages (e.g. internet services unavailable due to technical faults) prevented access to Facebook for limited time periods [20]. The study estimated that a 50% reduction in anti-refugee sentiment on social media would result in 421 fewer anti-refugee hate crimes (a reduction of 12.6%) [20].

Russian researchers have found similar results with the social-media platform *VKontakte* [19]. The fact that social media usage can have measurable *causal* effects on politically adverse behaviours such as hate crimes must give rise to concern.

## The distinct cognitive attributes of the web

The digital world differs from its offline counterpart in ways that have profound consequences for individuals as well as society. A more systematic and extensive review of the psychologically-unique properties of the internet was recently provided by Kozyreva and colleagues [24]. We leverage their analysis to provide a conceptual overview of the cognitive attributes of the web. Many of these are taken up at length in later chapters. The researchers identified two systematic differences between online and offline environments, one relating to structure and functionality and another relating to differences in perception and behaviour.

### Differences in Structure and Functionality.

***Network size***. On the one hand, the structures of communities and the number of close friends people have online can resemble their offline counterparts [25]. It appears that the cognitive and temporal constraints that limit face-to-face networks, such as attention and information processing, also limit online social networks. On the other hand, social media permit messages to be broadcast to a potentially very large audience. The number of followers (as opposed to followees) on a platform with a directed network structure such as Twitter is not limited and can far exceed any offline social reach [26]. When viral content travels through these large networks, it can accumulate social reactions (likes, shares, comments, etc.) in huge numbers that have no offline equivalent.

***Permanence***. On the one hand, the web does not forget. Information can be stored more or less indefinitely. This situation prompted the European Union to codify in Article 17 of the General Data Protection Regulation (GDPR) what is commonly referred to as the “right to be forgotten” which provided European citizens with a legal mechanism for requesting, under certain conditions, the removal of their personal data from online databases. On the other hand, platform outputs like Google Search rankings or Facebook newsfeeds are ephemeral. It is currently impossible to reproduce what a search for “Brexit” looked like during the UK referendum in June 2016.

***Personalisation***. Search engines and recommender systems collect and infer users’ preferences to deliver personalised results or recommendations. This technology has led to a gradual relinquishing of public control. Algorithms are both complex and non-transparent — sometimes for designers and users alike [27].

***Power of design***. The web cannot be accessed without interacting with choice architectures that constrain, enable and steer user behaviour. While physical environments such as cities or supermarkets can also be engineered, interventions are limited by physical factors and the original purpose of the infrastructure (e.g. streets for transport or supermarket shelves for storage). Online, by contrast, these constraints largely disappear. This has allowed platforms to evolve into sophisticated choice architectures whose main purpose is to engage user attention and persuade users to take certain actions. Moreover, while it might take several years to make a city bike-friendly (e.g. by building new bike lanes), adjusting powerful default settings of online choice architectures can occur almost instantly and at low costs.

### Differences in Perception and Behaviour.

***Social cues and communication***. On the one hand, compared to face-to-face interactions, online communication provides several additional opportunities, such as: (a) the potential for anonymity; (b) the ability to broadcast to multiple audiences; and (c) availability of extensive audience feedback. On the other hand, online communication eliminates many non-verbal or physical cues (e.g. body language or facial expressions). This elimination originally elicited much concern that computer-mediated communication might lead to impoverished social interaction [28]. However, it has now been recognised that users can replace non-verbal cues in digital communication with verbal expressions and graphical elements such as emoticons and “likes” [29]. Nonetheless, there is a large literature arguing that the distinctive features of online interactions — such as anonymity, invisibility and lack of eye contact — can reduce inhibitions, possibly increasing people’s tendency to express aggression in online fora [30, 31, 32, 33]. The lack of eye contact has been identified as having the greatest disinhibiting effect, being more important than anonymity [32].

***Cues for epistemic quality****.* Much web content now bypasses traditional gatekeepers such as professional editors. Content can nonetheless look professional and authoritative. Traditional cues for epistemic quality — e.g. quality of branding or typesetting — have therefore become less useful. New markers are emerging, such as crowd-sourcing (e.g. Wikipedia), but social-media feeds are largely curated without regard to epistemic quality [34].

***Social calibration****.* The internet has radically changed social calibration — that is, people’s perceptions about the prevalence of opinions in their social environment or the population. Offline, people gather information about how others think based on the limited number of people they interact with, most of whom live nearby. In the online world, physical boundaries cease to matter; people can connect with others around the world. One consequence of this global connectivity — which is usually heralded as a positive feature — is that small minorities can form a seemingly large, if dispersed, community online. This in turn can create the illusion that even extreme opinions are widespread, a phenomenon known as the false-consensus effect [35]. It is difficult to meet people in real life who believe the Earth is flat, whereas online, among the billions of those active on social-media, there are some who do share this belief and they can now easily find and connect with each other. The existence of an epistemic community provides perceived legitimacy for a person’s belief and renders them more resistant to changing their mind [35].

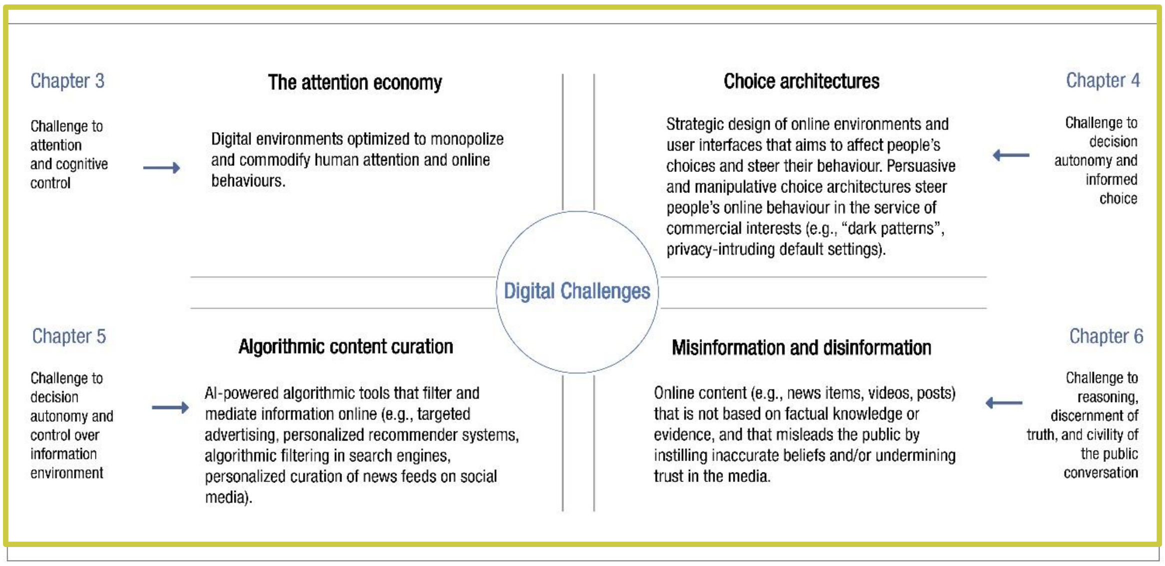
Social media has created a further source of miscalibration when multiple people are sharing information that is partially based on the same source. For example, if a single news article is retweeted by different individuals each of whom adds a comment in the tweet, a common recipient would receive messages that are correlated (because they rely on one article) but appear to be independent (because different individuals retweet). In those circumstances, people discount the correlation between messages, thus “double-counting” the underlying common source and being more sensitive to the information than is advisable [36].

***Self-disclosure and privacy behaviour****.* People’s attitudes and behaviours relating to privacy online are characterised by several paradoxical aspects. There is some evidence that people tend to be more willing to disclose sensitive information in online communications [37] and in online — as opposed to face-to-face — surveys [38, 39]. People are typically also highly permissive in their privacy settings when using the web. However, when their attitudes are probed, people profess to put a lot of weight on privacy [40]. This divergence between the importance people place on privacy in surveys and their actual behaviour when it comes to acting on those opinions has been identified as the “privacy paradox” [41].

***Norms of civility****.* Behavioural disinhibition is observed in many contexts online. Disinhibition can express itself in a behaviour known as “trolling”, a practice defined as “behaving in a deceptive, destructive, or disruptive manner in a social setting on the Internet with no apparent instrumental purpose” [42, p. 97]. Trolling can be used strategically to disrupt the possibility of constructive conversation. Trolling and other forms of incivility and harassment are pervasive: For example, among young Finnish people, approximately 47% reported encountering online hate in 2013 and this proportion had risen to 74% at the end of 2015 [43]. Women and minorities are disproportionately subject to online incivility and hostility [44]. An important dimension of the discussion about online incivility involves the distinction between incivility *per se* (i.e., rudeness) and anti-democratic intolerance [45]. The latter should be of far greater concern — even if expressed in seemingly civil language — than mere lack of politeness. The problem of online incivility and anti-democratic intolerance may be compounded by the recent finding that online moral outrage is experienced as being greater than in conventional media or in person [46].

***Dissolution of shared perceptions***. The web offers nearly unlimited choice. A result of this abundance of choice is that audiences are increasingly segmented. The segmentation of audiences has two related consequences for democracy: First, it creates an incentive for extremism because a politician may gain more voters on the extreme margins of their "base" than they repel in the moderate middle if they can selectively target extreme messages to their followers [47]. Second, when segmentation is accompanied by public polarisation, it becomes possible for politicians to create their own "alternative facts" [48] that they present as an ontological counter-measure to accountability [49].

***Pressure points: citizens vs. the internet***. Based on this analysis of the unique cognitive attributes of the web [24], four pressure points were identified that emerge when people and the online environment are brought into contact without much public oversight and democratic governance. Figure 1 summarises these four challenges. Each challenge is taken up in a chapter in this report.



Chapter 3

Challenge to attention and cognitive control

The attention economy

Digital environments optimized to monopolize and commodify human attention and online behaviours.

Choice architectures

Strategic design of online environments and user interfaces that aims to affect people's choices and steer their behaviour. Persuasive and manipulative choice architectures steer people's online behaviour in the service of commercial interests (e.g., "dark patterns", privacy-intruding default settings).

Chapter 4

Challenge to decision autonomy and informed choice

Chapter 5

Challenge to decision autonomy and control over information environment

Algorithmic content curation

AI-powered algorithmic tools that filter and mediate information online (e.g., targeted advertising, personalized recommender systems, algorithmic filtering in search engines, personalized curation of news feeds on social media).

Misinformation and disinformation

Online content (e.g., news items, videos, posts) that is not based on factual knowledge or evidence, and that misleads the public by instilling inaccurate beliefs and/or undermining trust in the media.

Chapter 6

Challenge to reasoning, discernment of truth, and civility of the public conversation

*Figure 1* - Map of challenges in the digital world. Adapted from [24].
Chapter number refers to chapters in this report that take up each challenge.

***The attention economy***. We can only consume a finite amount of information. We must therefore spread our attention between the multitude of competing sources offered by the web [50]. This has created an entire economy and its supporting technological apparatus to compete for our attention. As we will show in Chapter 3, the attention economy has several consequences for understanding how political behaviour unfolds online. For example, it has been argued that the zero-sum race for finite human attention explains why Internet technologies are designed to be appealing, addictive and distractive [51].

Perpetual information overload results from the attention race. Information overload has been associated with impoverished decisions about what to look at, spend time on, believe and share [52]. For example, longer-term offline decisions such as choosing a newspaper subscription (that then constrains one's information diet) have evolved into a multitude of online micro-decisions about which individual articles to read from a scattered array of sources. The more sources crowd the market, the less attention can be allocated to each piece of content and the more difficult it becomes to assess their trustworthiness — even more so given the demise and erosion of classic indicators of epistemic quality (e.g. name recognition, reputation, print quality, price). When quality ceases to be accessible for the end user, it disappears as a focal point of competition in the attention economy. The explicit goal is quantity, screen time and clicks, independent of the content itself. We take up the challenges arising from the attention economy in Chapter 3.

***Choice architectures***. Supermarkets are carefully designed to maximize shoppers' spending. In-store marketing can draw attention to products that shoppers had no intention of purchasing before they entered the store [53]. Conversely, stores can be redesigned to facilitate purchases of items recommended by nutritionists over other foods [54, 53]. Are those design decisions acceptable? Do they represent legitimate influence or persuasion or are they manipulative or even coercive?

Online architectures are behaviourally far more powerful than the physical options available to supermarket designers. Accordingly, some online choice architectures are ethically problematic because they stray into coercion or manipulation. Coercion is a type of influence that does not convince its targets, but rather compels them by eliminating all options except for one (e.g. take-it-or-leave-it choices). Manipulation is a hidden influence that attempts to interfere with people's decision-making processes in order to steer them toward the manipulator's ends. Manipulation does not persuade people and it may not technically deprive them of their options; instead, it exploits their vulnerabilities and cognitive shortcomings [55]. Not all choice architectures are manipulative [56] — only those that exploit people's vulnerabilities (e.g. hidden fears) in a covert manner. There are at least two cases where persuasive online design borders on manipulation: dark patterns and hidden privacy defaults. We devote Chapter 4 to an exploration of choice architectures, with a particular emphasis on instances of manipulation and coercion.

***Algorithmic content curation***. Without algorithms the utility of the web would be severely curtailed. Information is useful only to the extent that we can access it — and any search of the web inevitably involves algorithms that curate and personalise information.[6] Algorithmic filtering and personalisation are not inherently malign technologies — on the contrary, instead of showing countless random results for search queries, personalisation aims to offer the most relevant results. Googling "Newcastle" in Sydney, Australia, *should* prioritise information about the city that is 200 km to the north, not its distant British namesake.

In a similar vein, newsfeeds on social media strive to show information that is expected to be interesting to users. Recommender systems offer content suggestions based on our past preferences and the preferences of users with similar tastes (e.g. video suggestions on Netflix and YouTube). Algorithms can also filter out information that is harmful or unwanted, for example by filtering spam or flagging hate speech and disturbing videos. There are countless examples of why algorithms are indispensable and can be useful for human decision-making [58].

[6] We restrict consideration here to algorithms that members of the public are likely to encounter on the web. This excludes a class of important and powerful algorithms that serve as decision aids for experts, for example when predicting recidivism [57].

However, algorithms, like any other technology, come with their own set of problems. Those problems range from lack of oversight and transparency and consequent loss of autonomy, to computational violations of privacy and targeted political manipulation. Such problems can then be compounded by the use of biased data, which reproduce inequalities reflected in historical data. We explore these issues mainly in Chapter 5.

***Misinformation and disinformation***. Disinformation and conspiracy theories have been implicated in a number of recent political tragedies around the world. In Myanmar, the military orchestrated a propaganda campaign on Facebook that targeted the country's Muslim Rohingya minority group. The ensuing violence forced 700,000 people to flee the country [59].[7]

Most recently, the worldwide COVID-19 pandemic gave rise to multiple conspiracy theories and misleading news stories that have found considerable traction. For example, 29% of Americans believe that COVID-19 was created in a laboratory [60]. In the UK, the belief that 5G mobile technology is associated with COVID-19 has led to vandalism of infrastructure, with numerous cell phone masts being set alight by arsonists [61]. About one quarter of the British public consistently endorses some form of conspiracy related to COVID-19 [62] and endorsement of conspiracies has been found to be negatively associated with health-protective behaviours [63]. These developments are considered in detail in Chapter 6.

[7] At the time of this writing Facebook rejected requests to release Myanmar officials' data to the World Court (https://www.reuters.com/article/us-myanmar-facebook/facebook-rejects-request-%20to-release-myanmar-officials-data-for-genocide-case-idUSKCN2521PI ).

## Key scientific findings

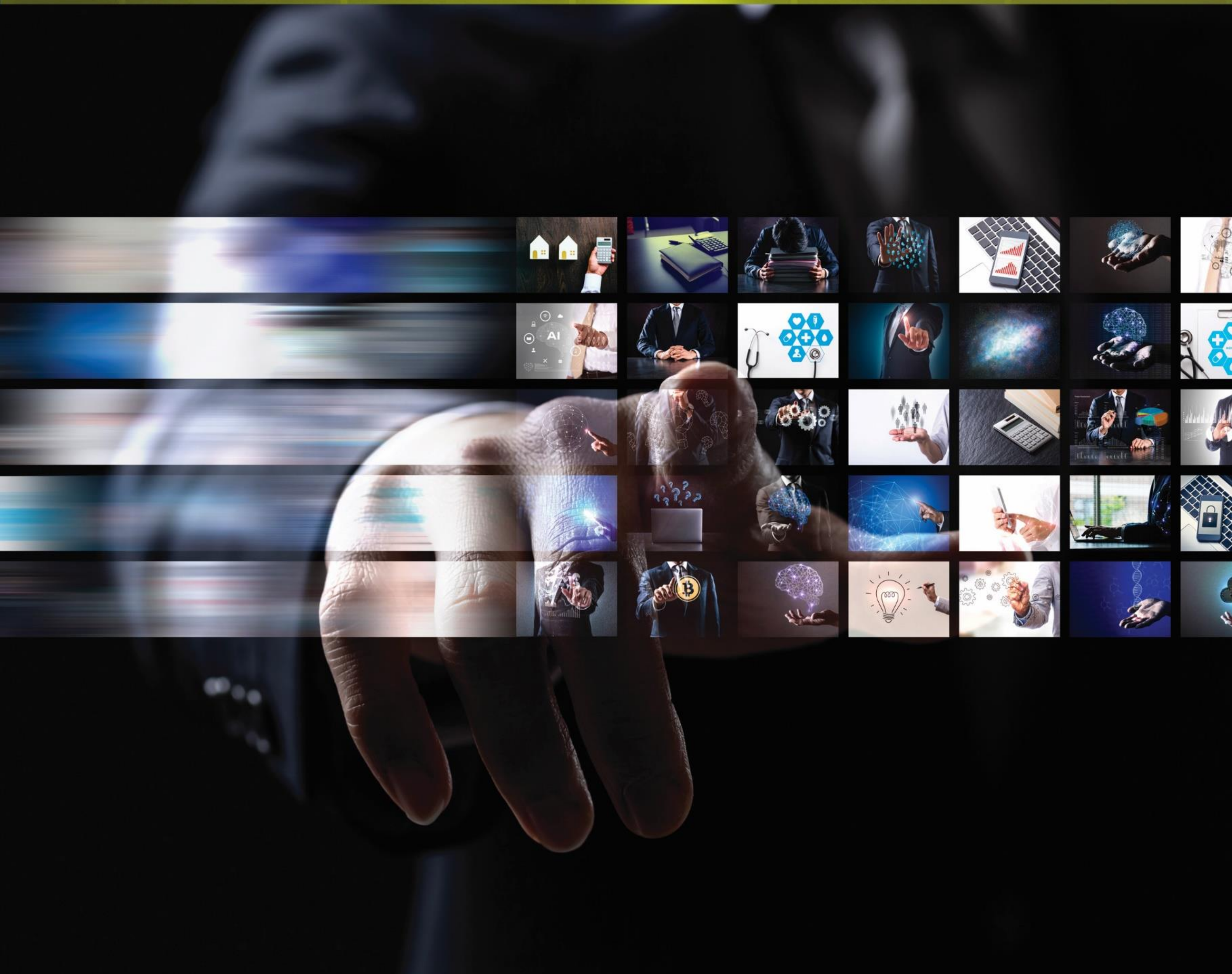
- Social media can have a causal effect on people's political behaviours, including inciting dangerous behaviour such as hate crimes.
- The web is cognitively unique, resulting in specific psychological responses to its structure and functionality as well as differences in perception and behaviour compared to the offline world.
- There are 4 pressure points when people and online systems interact: the attention economy; choice architectures; algorithmic content curation; and misinformation and disinformation.
- Structural factors in the design of online environments can affect how individuals process information and communicate with one another.

# Chapter 3

The attention economy

Specific characteristics

How this affects our behaviour



# Chapter 3: The attention economy

The philosophy of the internet has always been one of empowerment [16, 64] and this is echoed in EU policy. A recent example is the European Commission's Communication on *Shaping Europe's digital future*[8] that underscores citizens' empowerment as a goal of European digital policy.

These ambitions stand in contrast to the stark reality that the overabundance of available information has rendered people cognitively more impoverished than ever before [50]. As the informational capacity of the web increases, more issues can be considered, but the public's available attention span for each issue decreases. This is not mere speculation. Analysis reveals that whereas in 2013 a hashtag on Twitter was popular on average for 17.5 hours, by 2016 this time span had decreased to 11.9 hours [65]. The same declining half-life has been observed for Google queries [65]. The limitations of human attention and its exploitation by the attention economy have given rise to several interlinked consequences.

## Specific characteristics

Human attention has become the most precious resource in the online marketplace [11] and one that online platforms can steer by organising and curating content [66]. The business model of all leading platforms is to capture user attention for the benefit of advertisers. This commercial imperative may risk users' autonomy and the public good. For example, YouTube's recommender algorithm has the primary purpose to increase viewing time [67] and YouTube itself has claimed that 70% of viewing time on YouTube results from recommendations of its AI system, rather than purposeful consumer choice.[9] This raises questions about how much personal autonomy has been supplanted by recommender systems. Moreover, there is evidence that YouTube's recommendations are drawing viewers into increasingly extremist content [68, 69]. This raises questions about whether algorithms foster or undermine the public good. Crucially, users' attention and their behaviour are products being sold even when they are unaware that a commercial transaction is taking place — our time and attention are products while we watch videos on YouTube.

> *"Europe's digital transition must protect and empower citizens, businesses and society as a whole."*
>
> — Ursula von der Leyen,
> President of the
> European Commission

The limitations of attention resources and the resulting demand for algorithmic curation of information has created a relationship between platforms and their users that is profoundly asymmetric: Platforms have deep knowledge of users' behaviour and even intimate aspects of their lives [70]. Whereas, users know little about how their data are collected, how it is exploited for commercial or political purposes and how it

[8] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/shaping-europe-digital-future_en

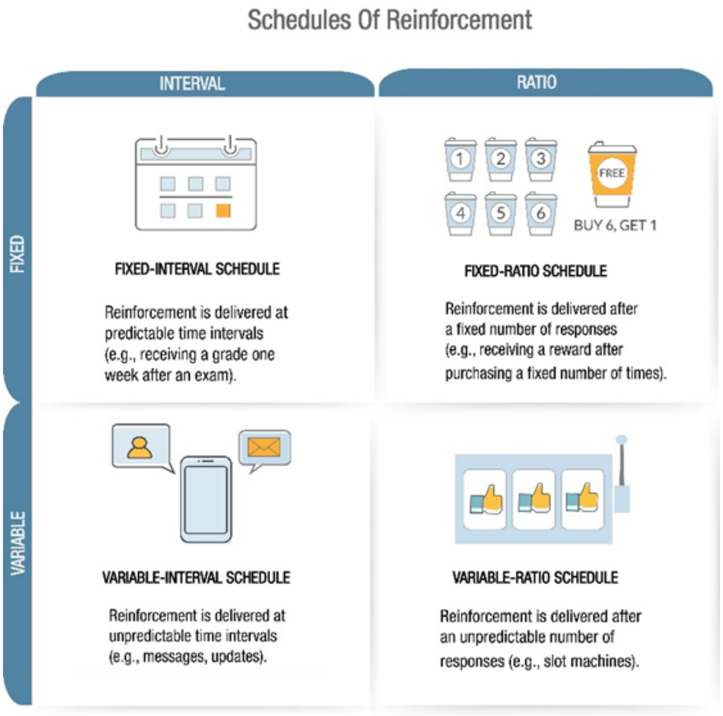
[9] https://www.cnet.com/news/youtube-ces-2018-neal-mohan/

and the data of others are used to shape their online experience. This asymmetry in knowledge also translates into an asymmetry of power: To keep others under surveillance while avoiding equal scrutiny oneself is the most important form of authoritarian political power [71, 72]. To know others while revealing little about oneself is the most important form of commercial power in an attention economy.

***Reinforcement architectures***. It is a known fact in psychology that the strength of behaviour depends upon reinforcement and in particular on the intervals or schedules of reward delivery. If one's goal is to maximise user attention, reinforcement schedules provide a powerful tool to pursue this goal. Scientists have identified two major classes of such schedules that are summarised in Figure 2 below:

- Fixed schedules deliver rewards at predictable time intervals (fixed-interval) or after a predictable number of attempts (fixed-ratio schedules).

- Variable schedules deliver reinforcement with less predictability.



Schedules Of Reinforcement

FIXED-INTERVAL SCHEDULE

Reinforcement is delivered at predictable time intervals (e.g., receiving a grade one week after an exam).

FIXED-RATIO SCHEDULE

Reinforcement is delivered after a fixed number of responses (e.g., receiving a reward after purchasing a fixed number of times).

VARIABLE-INTERVAL SCHEDULE

Reinforcement is delivered at unpredictable time intervals (e.g., messages, updates).

VARIABLE-RATIO SCHEDULE

Reinforcement is delivered after an unpredictable number of responses (e.g., slot machines).

*Figure 2* – Schedules of reinforcement: Social media or online gaming offer their users rewards (e.g. "likes" or reaching another level in a game) to reinforce and maintain the desired behaviour — namely, time on platform. See text for details about schedules. Equivalent schedules can also be found offline.

Variable schedules are further differentiated into variable-interval and variable-ratio schedules. In a variable-interval schedule, reinforcements are delivered at time intervals that are independent of a person's behaviour and unpredictable from their perspective, even though the underlying dynamics are known to the experimenter or designer. Variable-ratio schedules, by contrast, involve reinforcement after an average (but variable) number of responses (e.g. winning a prize after a variable number of attempts).

> *Once Big Data systems know me better than I know myself, authority will shift from humans to algorithms. Big Data could then empower Big Brother.*
>
> — Yuval Noah Harari

Both variable schedules are known to create a steady rate of responding, with variable-ratio schedules producing the highest rates of responding and variable-interval schedules producing moderate response rates. It seems that if rewards are difficult to predict, people — just like other organisms studied in the laboratory — tend to increase the rate of a particular behaviour, perhaps hoping to eventually attain the desired reward.

Although people are indubitably capable of analytic thought, we do not always engage in careful deliberation. In those circumstances, people respond to social rewards online in much the same way as any other species responds to reinforcement in the laboratory. To illustrate, Facebook provides users with rewards in the form of "likes" and shares, social reinforcements in messages, comments and friend requests. A recent analysis of four large social media datasets (Instagram and three topic-specific discussion boards) revealed that reward learning theory, originally developed to explain the behaviour of non-human animals in conditioning environments, can also model human behaviour on social media [73]. People calibrate their social media posts in response to rewards (likes) as predicted by reward learning theory [73].

Jonathan Badeen, cofounder of the online dating app Tinder, recently acknowledged that its algorithms were inspired by this behaviourist approach [74]. Reinforcements constitute messages, likes, matches, comments or any desirable content that is delivered at irregular intervals and prompts users to constantly refresh their feeds and check their inboxes.

Attracting attention is only a first step to successful advertising: a further necessary step is to persuade the recipient to engage with content. The success of persuasion can be enhanced by personalising message content.

***Personalisation and audience segmentation***. The "Cambridge Analytica scandal" of 2017 created much public concern about "microtargeting" [75]. Microtargeting is an extreme form of personalisation that exploits intimate knowledge about a consumer to present them with maximally persuasive advertisements. Cambridge Analytica was implicated in using microtargeting during the Brexit referendum campaign.[10]

Microtargeting is particularly problematic when it exploits people's personal vulnerabilities. For example, according to a 2017 report, Facebook (in Australia) had the technology to allow advertisers to target vulnerable teenagers at moments when they feel "worthless" and "insecure." Facebook did not dispute the existence of the technology although it claimed that it was never made available to advertisers and only used in an experimental context [76]. Facebook apologised at length and reassured the public that "Facebook does not offer tools to target people based on their emotional state.".[11] Facebook was, however,

[10] https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy

[11] https://about.fb.com/news/h/comments-on-research-and-ad-targeting/

awarded a patent[12] based on technology that allowed one "to predict one or more personality characteristics for the user. The inferred personality characteristics are stored in connection with the user's profile and may be used for targeting, ranking, selecting versions of products and various other purposes."

There is considerable evidence to suggest that data collected about people online can be used to make inferences about highly personal attributes. Kosinski and colleagues analysed how the Facebook user likes could be used to infer private attributes, including sensitive features such as religion and political affiliation [77].

The predictive power varied across attributes, from nearly perfect for race to slightly better than chance for "whether an individual's parents stayed together until they were 21 years old" [77]. Algorithmic personality judgements based on information extracted from people's digital fingerprints (specifically, Facebook likes) can be more accurate than those made by relatives and friends [70]. Knowledge of 300 likes is sufficient for an algorithm to predict a user's personality with greater accuracy than their own spouse [70].

A review of 327 studies revealed that numerous demographics could be reliably inferred from digital fingerprints, including for example sexual orientation [78]. Other research concluded that online architecture inferred personality (defined as the "Big 5" attributes) from digital fingerprints with greater accuracy than human judges [79]. Recent empirical research has shown that Facebook might be inferring sensitive attributes of European users, such as sexual orientation, even after the GDPR was implemented [80].

On balance, there is little doubt that access to people's digital fingerprint permits inference of their personality. Inferences of other attributes, such as personal values and moral foundations, is also possible albeit at best with modest accuracy [81].

The power afforded by such inferences into intimate details of people's lives is considerable. A recent analysis warned of the dangers that the "personality panorama" offered by big-data analysis could all too readily turn into a "personality panopticon", a dystopia in which each person's behaviours are "ceaselessly observed and regulated" [82, p. 6]. There is therefore a direct and strong link between the data that permit personalisation and the implications for people's privacy.

*Privacy in the attention economy. Privacy as a public good.* The conventional view of privacy is as a private good: My data are mine, your data are yours and each of us is entitled to choose whether or not to relinquish these data to government, corporations and other entities or persons. At its core, the individual can decide whether or not to grant access to and allow use of their data. In the realm of the EU Charter of Fundamental Rights (Article 7), the right to privacy protects "private" information from any unjustified state interference. Independently, the Charter contains a fundamental right to data protection (Article 8) that applies only to natural (not legal) persons, but covers all (not just private) personal data [83]. It demands that this data "must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law." The corresponding EU data protection regime imposes concrete obligations on private parties. It is prohibitive in nature as no-one is allowed to process personal data unless one of several conditions is satisfied (pertaining to grounds of processing and data protection principles).

[12] US Patent No 8,825,764, with Michael Nowak and Dean Eckles as inventors; see https://patents.google.com/patent/US8825764B2/en

The regime thus empowers data subjects with a set of rights they can exercise against data controllers. While the regime expressly permits the processing of personal data without consent (if certain conditions are met, i.e. a lawful basis exists, it serves legitimate interests and is necessary for the performance of a contract, etc.) and the GDPR's objectives focus on the uninhibited transfer of personal data around the European Union, transfer is only permitted when processing adheres to the "fundamental rights and freedoms of natural persons."

Thus, the role of the data protection regime is limited in both its personal (data controllers and processors) and material scope (personal data *relating* to an identifiable natural living person) and is insufficient in an attention economy in which personal data are used to infer intimate characteristics not just of the individual user *but also of others* [84]. Exercising one's right to make personal data available or public therefore has negative externalities (i.e. negative side effects on innocent bystanders not involved in the decision): Individuals are vulnerable merely because *others* have been careless with their data.

This turns privacy into a public good [85, 86], with far-reaching ramifications for democracy. One recent illustration involves the exercise app Strava, which published a "heat map" showing where its users were jogging or cycling. The map was found to inadvertently reveal the location of US military installations around the world, including some whose existence had not been made public [87]. It follows that empowerment of individual citizens, for example through the GDPR, may be insufficient—privacy also requires coordination between individuals [85].

*Beyond the raw data.* The power to exploit digital fingerprints [77, 70] implies that protecting users' data has to go beyond considering the data that are collected from them — we must also take into account how that data is processed and what inferences are drawn. Users are unlikely to be fully aware of what data is being collected — few may realise that the text of Facebook comments is analysed even if the user decides not to post the comment [88]. People are also unlikely to recognise what is inferred from their data, as revealed by the anecdote of the department store Target inferring from purchasing behaviours that a teenager was pregnant before her parents knew [89].

Beyond individual inferences, the persistent and networked nature of online information systems creates further problems for individuals to control the use of their data [90]. Data are shared across systems and services, which curtails users' ability to understand what can be inferred from their data and what they might be disclosing about others. One example of the complexity of digital privacy is the possibility to build shadow profiles with information on individuals who do not have an account on the platform [91]. Information on these individuals can be inferred from the data that users voluntarily provide, which can be combined with contact lists and other kinds of relational data to make inferences of personal attributes of people without an account. This inference builds on statistical patterns of social interaction e.g. the preferential congregation of people with shared political affiliation or sexual orientation [92].

Field research has shown how shadow profiles can be built. When empirical data on social networks are combined with simulations of the spreading of their adoption, it can be illustrated how a network can infer the friendship between two non-users [93]. This has far-reaching consequences, particularly for circumspect citizens who choose to stay away from social network platforms in the belief that this will help protect their privacy, as this is not the case. Using data from the now disbanded social network "Friendster", research has shown that Friendster user data was predictive of people who did not have an account, with particular respect to marital status and sexual orientation [92]. Analysis of Twitter data shows that the location of a non-user can be predicted from their friends on Twitter, showing increasing accuracy with the number of friends who are on Twitter [94].

This complexity of privacy also threatens the right to be forgotten that is enshrined in the GDPR. Even after deleting a user account, the traces remaining in the data of other users can be used to infer attributes of a person [95]. Several personal attributes, including religion and personality traits, can be estimated using only the tweets written by the *friends* of a Twitter user, without the need to use the text of the user itself [96]. If digital traces are left behind without a user's awareness and ability to delete them, furtive political manipulation is facilitated.

A recent scientific study compared the predictive ability of analysing a user's past tweets to predict the text of their future tweets, to the ability of predicting that person's future text from the text of their close contacts [97]. The analysis revealed that 95% of the potential predictive accuracy for an individual can be achieved without access to that person's previous texts, by using their social ties only (with as few as 8 or 9 contacts being sufficient) [97]. As user text is a powerful tool to infer behaviour and personality [82], this kind of social inference opens the door to potential manipulation, beyond the control or awareness of an individual.

**Manipulative targeting**. Marketers have always segmented audiences. Motorcycle magazines are unlikely to contain advertisements for cosmetics. What, then, is the boundary between justifiable audience segmentation and manipulation? Philosopher Daniel Susser and colleagues published a significant analysis relevant to this report. They examined what it means to manipulate someone online and how manipulation can be systematically differentiated from other forms of influence that are seen to be more legitimate, such as persuasion, as well as forms that are clearly considered unacceptable, such as coercion [55]. The core concept in the analysis is that of users' autonomy; that is, their ability to know what they desire and to act on reasons they think best [55, p. 36]. That autonomy is threatened by manipulators hiding their intentions and actions, this threat increases in proportion to the amount of knowledge held by the manipulator about their targets [55, 38]. A similar analysis was provided by B. J. Fogg [56].

A recent analysis by Lorenz-Spreen and colleagues derived three dimensions that together determine whether algorithmic targeting is ethically and politically problematic or acceptable [98]:

- granularity of the target group (i.e., the fewer people are targeted, the more personal data of each individual are used to permit narrowing of the target);
- domain (political vs. non-political, to the extent that a clear differentiation is possible); and
- how personal data were obtained (provided by the user or, e.g. personality data inferred from digital fingerprints, such as Facebook "likes").

Information that microtargets just a few users for political purposes based on inferred personal information is maximally problematic: It is non-transparent and manipulative irrespective of content. In contrast, information targeting a broad audience based on user-provided data for commercial purposes is least problematic: It represents conventional market segmentation. Although these distinctions are relatively easy to make at a conceptual level, in practice the differentiation may be more difficult. To illustrate, whereas advertisements for facemasks would have attracted no political attention a year ago, the COVID-19 pandemic has rendered this product highly political in at least some societies.

This analysis is echoed in public opinion data. In Germany, a recent representative survey probed the public's attitudes towards artificial intelligence online, in particular the use of machine learning to exploit personal data for personalisation of services [40]. Attitudes towards personalisation were found to be domain-dependent: Most people find personalisation of political advertising and news sources unacceptable. For instance, 61% oppose customised political campaigning and 57% object to personalised newsfeeds on social media. At the same time, a majority approves of personalised entertainment (77%), shopping (78%) and search results (63%). A majority of respondents objected to personalisation based on sensitive information (e.g. religious or political views, personal events, personal communications) [40].

Additionally, a 2018 Special Eurobarometer Report on democracy and elections[13] found broad support amongst Internet-using respondents for applying the same pre-election rules for traditional media to social networks, Internet platforms and the actors that use them. More than six in ten Internet-using respondents in each Member State shared this opinion. At least eight in ten of these respondents were in favour of transparency about political advertised content including who is paying for it.

The analysis in the study by Lorenz-Spreen and colleagues [98] also meshes well with other researchers' conclusions that microtargeting carries considerable risks in the political domain. At least six harms can arise from microtargeted political advertising [75]:

- Microtargeting is harmful because it exploits personal data without the user's consent.
- Microtargeting is harmful because it conceals its intent and true nature.
- Microtargeting is harmful because claims made in targeted messages cannot be corrected or debated in the free marketplace of ideas.
- Relatedly, microtargeting is harmful because it permits disinformation to spread without opportunity for correction.
- Microtargeting is harmful because it potentially allows politicians to make mutually incompatible promises to different segments of the electorate.
- Microtargeting is harmful because it permits foreign actors to influence domestic political campaigns.

A further possible avenue to limit the harms associated with malicious advertising consists of content regulation. Facebook offers a "custom audiences" feature that allows the social media platform to match users to an email address provided to them by an advertiser. The provision of these addresses usually takes place through a consent mechanism, satisfying data protection obligations. The platforms undertake minimal content checking with much of the oversight executed by AI with limited human supervision. In consequence, "issue-based" advertising has thrived on platforms like Facebook. Actors interested in influencing democratic processes and elections purchase some of these advertisements, which could be designed to stoke civil unrest and spread hate speech. Yet, they might escape the remit of regulators concerned with safeguarding the integrity of electoral processes. At the same time, because these advertisements can amount to dark posts (only visible for individual targeted users and not accessible to a common audience), their content is not subjected to the corrective marketplace of ideas as is other (user-generated) content.

Similarly, platforms offer demographic marketing; for example, an ad can be delivered to all users living in Amsterdam over the age of 50 (or any multiple of variations). During the COVID-19 pandemic, the European Union's Consumer Protection Authorities spent considerable time and resources tracking and

[13] https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/ResultDoc/download/%20DocumentKy/84538

ordering the removal of malicious advertisements; for example, selling high dose Vitamin C to combat COVID-19 “after scientific breakthroughs”. In the latter example, it is not a certainty that the advertiser ever comes into contact with personal data, but the advertisement’s harmful content could still be delivered to a considerable number of users with detrimental effects on society [99]. Other critical analyses of microtargeting have also been conducted [100][101].

With this in mind, it is noteworthy that a recent Eurobarometer study focusing on Artificial Intelligence,[14] found that 80% of the representative EU population sample think that they should be informed when a digital service or mobile application uses AI.

Whatever the political, legal or ethical implications of microtargeting may be, the magnitude of the issue is tied to pragmatic considerations: Is it really possible to target people’s selective vulnerabilities and, if so, does this lead to successful—but illegitimate— manipulation? We explore these pragmatic questions next.

## How this affects our behaviour

**Microtargeting: Hype or dystopian manipulation?** Microtargeting is technically legal. The direct effects of microtargeting on behaviour are nuanced and difficult to assess. There is evidence of (at least potential) harm, but benefits may also exist.

***The potential harms of targeting*.** Facebook claimed credit for their targeting abilities after the Conservative Party in the UK won the election of 2015, boasting that “the party was able to reach 80.65% of Facebook users in the key marginal seats. The party’s videos were viewed 3.5 million times, while 86.9% of all ads served had social context—the all-important endorsement by a friend.”[15]

Based on the testimony of a former Cambridge Analytica employee, the firm used Facebook personality profiling to target fear-based messages (e.g. “Keep the terrorists out! Secure our Borders!”) to people high on neuroticism during Donald Trump’s 2016 presidential campaign.[16] Researchers have subsequently confirmed those claims by identifying, for example, three attack vectors that could be exploited to target specific *individuals* via Facebook [102]. Attributes such as income, net worth, relationship status, clothing size, housemates and many others could be obtained from a single person on Facebook, even if their privacy setting precluded sharing of that information with people who are not “friends” [102]: In response to the research Facebook closed this loophole. The same study also showed that single individuals and single households could be targeted with messages using Facebook’s ad delivery services. Facebook did not alter their policies in response to the research, suggesting that such targeting of specific individuals is still possible [102].

[14] https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/%20instruments/STANDARD/surveyKy/2255

[15] https://www.facebook.com/business/success/conservative-party

[16] *The Great Hack Documentary* / Netflix

One concern about targeted political advertising is the fear that it may promote polarisation. There are multiple ways polarisation can emerge and is maintained [103]. One pathway to polarisation involves separate information streams for subgroups within a society. It seems possible, even likely, that targeted information campaigns can promote polarisation [104]. Showing ads across partisan lines on Facebook (i.e. liberal ads to conservatives or vice versa) can cost three times more than showing an "aligned" ad to the same audience [104]. Even when an advertiser is explicitly trying to reach a diverse audience, Facebook's delivery system has been found to preferentially present the ad to users who Facebook predicts to be more interested and hence likely to be of the same political stripe [105]. Russian efforts to polarise members of the American public provide an example of this process.

> *"The trick to a successful campaign lies less in a powerful overarching cause than in crafting different messages for different constituencies, focusing on the issues that matter to each of them. For example, the most successful message in getting people out to vote had been about animal rights. Vote leave argued that the EU was cruel to animals because, for example, it supported farmers in Spain who raise bulls for bullfighting. And within the "animal rights" messaging, Vote Leave could focus (sic) even tighter, sending graphic ads featuring mutilated animals to one type of voter and more gentle ads with pictures of cuddly sheep, to others. A country of 20 million people requires between 70 and 80 types of targeted message on social media"*
>
> — Thomas Borwick, Chief technology officer for the UK's Vote Leave campaign in an interview with Peter Pomerantsev

The campaigns have used online platforms to reach different groups of users with distinct messages intended to drive division and hatred between these groups. For instance, largely liberal groups (e.g. LGBTQ+ and Black Lives Matter groups) were presented with information vilifying conservatives, while conservative groups (e.g. gun rights advocates) were presented with the opposite sort of content [103].

There is debate, however, about the degree to which this sort of advertising actually drives polarisation. At least one analysis of large data sets found no evidence of a link between political advertising and increased polarisation [106]. However, other authors have argued that advertising can cause polarisation [107]. The latter position is supported by a recent report [108] that Facebook was aware of the polarising effects of its algorithms but shelved potential countermeasures.

The impact of microtargeting is exacerbated by the lack of transparency in political campaigning online. Notwithstanding recent transparency measures by Facebook (e.g. the "ad library"), it is nearly impossible at present to trace how much has been spent on microtargeting and what content has been shown [109]. This difficulty is likely to persist because ads on Facebook are delivered by a continually-evolving algorithm, known as *AdTech*, that auctions off ads on a second-to-second basis based on live analysis of user data [104].[17] A recent investigation of AdTech by the UK's Information Commissioner's Office expressed a number of serious concerns, mainly relating to the non-consensual use of sensitive personal data and the complexity of the data supply chain.[18] In 2018, the European Data Protection Supervisor concluded that AdTech is an ecosystem that "has now been weaponised by actors with political motivations, including those wishing to disrupt the democratic process and undermine social cohesion. Opaque algorithmic decision-making rewards content which provokes outrage, on the basis that greater engagement generates revenue for the platforms in question. This poses obvious risks to fundamental values and democracy."[19]

***The psychology of targeting***. Several laboratory experiments have shown targeting to be effective. An early study from 2012 showed that "adapting persuasive messages to the personality traits of the target audience can be an effective way of increasing the messages' impact" [110]. Similar results were reported in a large-scale field study on Facebook [111]. The study used the platform to deliver cosmetic ads to subsets of (female) Facebook users that were identified as being either introverted or extraverted based on their "likes" [111]. Ads that were matched to people's introvert/extrovert score "resulted in up to 40% more clicks and up to 50% more purchases than their mismatching or unpersonalised counterparts" [111, p. 12714].[20]

***Benefits of targeting***. Social media can amplify the reach of public information campaigns and mass media communication, activating public discussion and setting up political agendas. Digital reach amplification has been used in several domains to inform citizens for their own good or to motivate behaviour change towards better well-being. The best documented examples of such digital campaigns are often related to health issues. For example, advocacy organisations on Facebook have raised awareness about autism spectrum issues by building "cultural bridges" between communities that can be empirically quantified [116].

Health promotion campaigns can combine the additional reach of social media with microtargeting techniques to increase their effectiveness. For example, cancer prevention messages are more effective when their text is personalised to match cognitive traits of the recipient [117]. However, campaigns for good can also meet resistance if they are considered invasive or patronising. A recent example is the EAT-Lancet report [118], which proposed a healthy and sustainable diet. Despite being promoted by several organisations and news media, the EAT-Lancet social media campaign met a significant backlash on Twitter [119].

***How do people understand and manage privacy***. A recent Eurobarometer study[21] on "Attitudes towards the impact of digitalisation on daily lives", found that 59% of the representative EU population sample

[17] See also https://www.theatlantic.com/technology/archive/2020/04/how-facebooks-ad- technology-helps-trump-win/606403/

[18] https://ico.org.uk/media/about-the-ico/documents/2615156/adtech-real-time-bidding- report-201906.pdf

[19] https://edps.europa.eu/sites/edp/files/publication/ar2018_en.pdf

[20] The study by [111] has been subjected to critiques [112, 113] which were rebutted by the original authors [114, 115].

[21] https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/survey/getsurveydetail/%20instruments/special/surveyky/2228

would be willing to share some of their personal information securely to improve public services. In particular, most respondents were willing to share their data to improve medical research and care (42%), to improve the response to crisis (31%) or to improve public transport and reduce air pollution (26%). However, importantly in 13 EU countries, more than one third of the respondents would not be willing to share any of their personal information for any purpose, with more than four in ten saying this in Bulgaria (43%), Poland (43%), France, Hungary and Latvia (41%).

In general, however, people value their privacy, with 82% of respondents in a recent German survey claiming that they are very or somewhat concerned about their data privacy [40]. In line with the "privacy paradox", in the same survey significantly fewer respondents reported taking steps to protect their privacy online: Just 37% adjust privacy and ad settings on online platforms and 20% do not use any privacy-enhancing tools [40]. There are several potential reasons for this discrepancy between what people say about online privacy and what they actually do. One reason is the lack of transparency and understanding of how online platforms collect and use people's data and what can be inferred from that data. For example, more than 60% of the participants of a Facebook user study in a US university were not aware that the Facebook newsfeed selects information to be displayed based on their personal data [120] (this study was conducted in 2015 and public perceptions of Facebook may have changed since then).

The second reason is that platforms may use defaults that favour collection of data over users' privacy, making it difficult to choose privacy-preserving options. With this in mind, people are not incoherent, but rather, their attitudes towards privacy are difficult to translate into behaviour because the platforms have made privacy unnecessarily complicated to achieve.

A somewhat different view, proposed by some researchers, is that people engage in a privacy calculation, "which states that people will self-disclose personal information when perceived benefits exceed perceived negative consequences" [121, p. 369].

## Key scientific findings

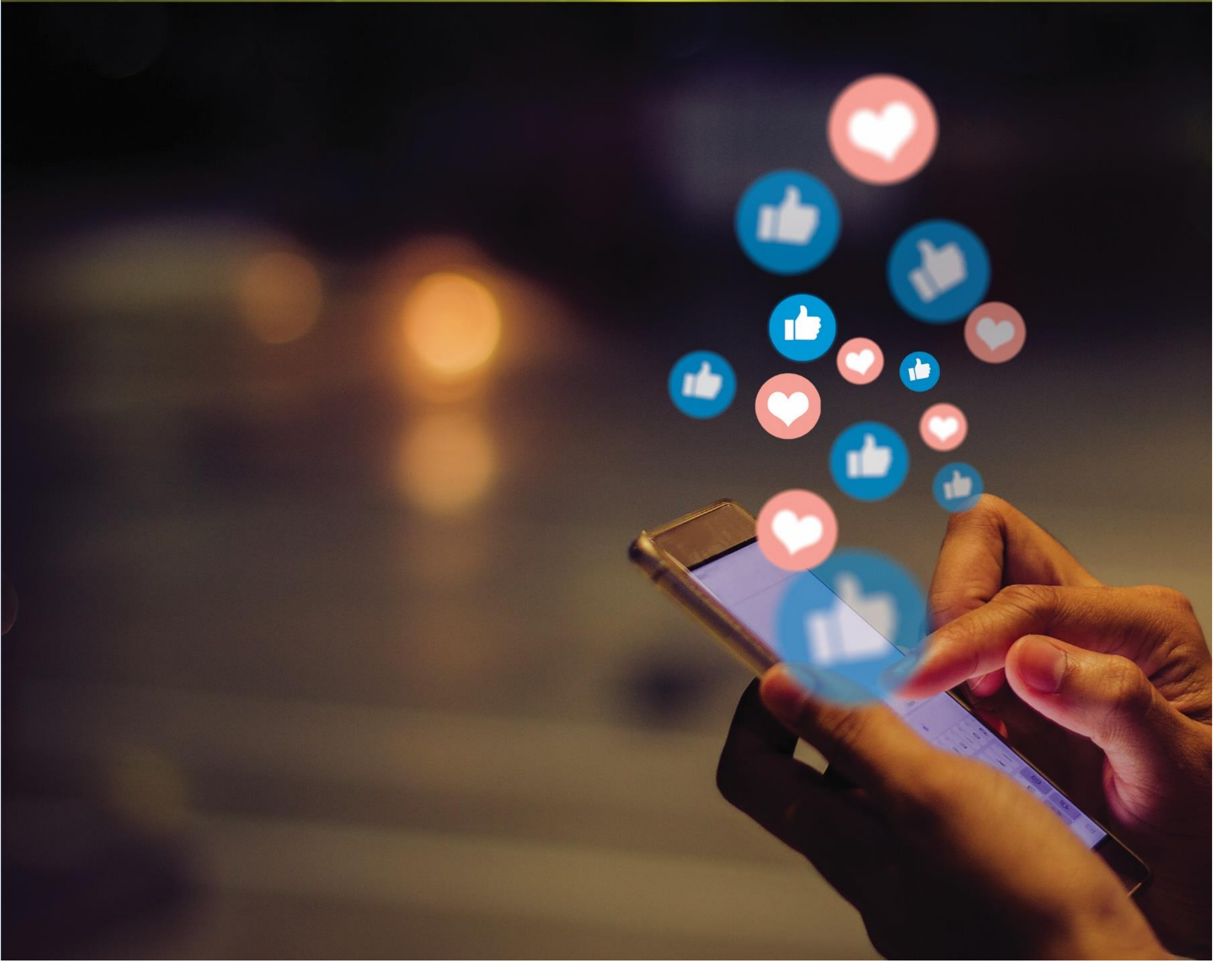
- The Internet is tightly controlled by private corporate algorithms designed to maximise profits by capturing our attention without public accountability.
- Human attention is the main object of the online economy and is considered as a commercial "product" especially on social media.
- The business model of the online economy constrains the space of possible solutions that are achievable without regulatory intervention.
- Microtargeting of political messages has considerable potential to undermine democratic discourse, which is a foundation of democratic choice and is being used to this end.
- Most people are opposed to microtargeting of certain content (e.g. political advertising) or based on certain sensitive attributes (e.g. political affiliation).
- Corporate social media pose a privacy risk for users and non-users alike, that extends far beyond what individuals explicitly share with social media sites, because of how much can be inferred from users' activity.
- The digital sphere is designed so that people give their valuable time, attention and data without considering the costs — for themselves and others. This exploits certain features of human behaviour, which makes it hard to address at the individual level. On a societal level, coordination is needed to assure privacy as a public good and to buttress democracy, freedom and equality.

# Chapter 4

## Choice architectures

Specific characteristics

How this affects our behaviour



# Chapter 4: Choice architectures

Choices almost always require an enabling architecture. When we order a beverage in a restaurant, we choose from a set of options (e.g. “small”, “medium” or “large”) rather than expressing our preference based on an assessment of our thirst (should this even be possible) and translating that into a desired quantity (“362 millilitres please”). Choice architectures are ubiquitous in digital settings. The necessity for a choice architecture creates a powerful opportunity for platforms to guide users’ choices — be they commercial or political — in the platform’s interests.

As choice architectures are ubiquitous, our analysis is necessarily limited to a few key domains that are the most significant from a democracy perspective. Our selected domains focus on choices about privacy and associated issues, such as knowledge and management of data sharing.

## Specific characteristics

***Defaults***. Default settings have a strong effect on the choices of users of online platforms. A simple change from an opt-out to an opt-in default has been found to raise participants’ consent to be informed about future health surveys from 48% to 96% [122]. Default settings can also lead users to unwittingly share sensitive data, such as location information that can be used to infer attributes like income or ethnicity. On Twitter, sharing the GPS location of a user was, at one point, activated by default (opt-out format) when posting a tweet that included a place tag, for example “New York City” [123]. In April 2015, the default was changed to opt-in, leading to a decrease in the number of geocoded tweets in the US from more than 2.5 million per day in 2014 to less than half a million per day in 2015 [124]. In 2019, the option to add precise GPS coordinates to tweets was removed [125].

The design of defaults can also allow platforms to get around regulations. For example, the GDPR stresses the importance of privacy-respecting defaults and insists on a high level of data protection that does not require users to actively opt out of the collection and processing of their personal data (GDPR Article 25). However, according to a report by the Norwegian Consumer Council [126], tech companies such as Google, Facebook and, to a lesser extent, Microsoft, use design choices in “arguably an unethical attempt to push consumers toward choices that benefit the service provider” [126, p. 4]. Thus, default settings, serving the interests of the service providers, tend to be privacy-intrusive (e.g. Google requires that the user actively go to the privacy dashboard in order to disable personalised advertising).

Compliance issues are not limited to defaults. A quantitative study of privacy policies from 248 US companies in privacy-sensitive markets such as social networks and dating sites revealed multiple problems [127]. For example, online contracts often missed critical information on privacy-relevant issues, contained unclear language and frequently did not include the privacy standards the firms claimed to adhere to (such as the EU Safe Harbour Agreement; [127]).

***Framing and commercial nudging***. Framing and wording may also be used to nudge users towards a choice by presenting the alternative as risky (e.g. on Facebook, users are encouraged to keep face recognition turned, because it ostensibly helps “protect you and others from impersonation and identity misuse, and improve platform reliability.”[22]). Choice architectures may also require a take-it-or-leave-it

[22] https://www.facebook.com/help/122175507864081

decision (e.g. a choice between accepting the privacy terms or deleting an account) or they may be designed such that the privacy-friendly option requires more effort and knowledge from users. These design features may contribute to the privacy paradox previously discussed.

***Dark patterns***. Dark patterns are particularly extreme forms of manipulation and are defined as design choices that "benefit an online service by coercing, steering, or deceiving users into making unintended and potentially harmful decisions" [128, p. 2].[23] They are a coercive and manipulative design technique used by web designers when some sort of action is needed from a user — typically to begin the processing of personal data or indication of agreement to a contract [129]. One notorious example of dark patterns is the "roach motel" which makes it easy for users to get into a certain situation, but difficult to get out. For instance, creating an Amazon account requires just a few clicks, but deleting it involves 12 steps that are difficult to achieve without instructions.[24]

Dark patterns are pervasive. A recent search for dark patterns on 11,000 shopping websites identified 1,818 dark patterns [128]. The patterns relied on various techniques such as misdirection, applying social pressure, sneaking items into the user's shopping basket and inciting a sense of urgency or scarcity (a strategy often used by hotel booking sites and airline companies) [128]. It is unknown how dark patterns are being used by political influencers, but the potential for users being involuntarily retained by political campaigns must give rise for concern.

Yet, under Article 38 of the EU's Charter of Fundamental Rights, "public authorities shall guarantee the protection of consumers and users and shall, by means of effective measures, safeguard their safety, health and legitimate economic interests". The EU consumer protection *acquis* provides for an extensive framework of consumer rights. The EU has recently initiated a "New Deal for Consumers" initiative, ultimately adopting the "Directive on better enforcement and modernisation of EU consumer protection" (Directive (EU) 2019/2161),[25] which also adapts existing rules to the modern conditions of the online marketplace. Both the data and consumer protection regimes stand alone but can be used to supplement and complement each other.

As noted recently by Leiser [129], "analysing the overall system architecture, user experience, and interfaces for fairness would permit consumer law to gauge whether the totality of the consent process was fair on consumers. This requires moving away from proposals for remedying the consent fallacy that focus on the panacea of consent simplification and reflects recent acknowledgements found in the European Union's 'New Deal for Consumers' that personal data is increasingly seen as having economic value." Therefore, the EU's consumer protection regime could be used to determine whether the process that led to processing was fair on users [130].

[23] See also https://darkpatterns.org/
[24] https://www.wikihow.com/Delete-an-Amazon-Account
[25] https://eur-lex.europa.eu/eli/dir/2019/2161/oj

## How this affects our behaviour

***Defaults and information cost***. Defaults clearly matter. It is also clear that defaults are exploited by platforms to elicit user behaviours that are in the platform's interest. However, people's interaction with defaults is more complex and nuanced than simple compliance.

Much research has focused on the design of default rules, often focusing on the costs to declare an opt out. These costs include mechanical costs as well as information costs: That is, the user must manually opt out and acquire the necessary information in order to take this decision. One method to increase the "stickiness" of defaults is to increase the mechanical effort users have to invest in order to opt out.

***Understanding data***. People are generally unaware of what personal data is being collected and potentially shared. A particularly pertinent example involves photos, which can be accompanied by a very large number of "tags", including the date and time a photo was taken, the name of the camera owner, a camera's unique ID, geolocation information and even uncropped preview images [131]. Because there is bewildering variety in how those tags are stored, how they are stripped (if at all) during upload, the privacy implications of photo metadata are difficult to discern, even for experts [131]. In a survey conducted among members of a university community in Germany, only 61% indicated familiarity with the concept of metadata accompanying photos [132]. Of that informed subset, over half (58%) did not know what happened to metadata upon uploading to their preferred platform [132].

***Interacting with privacy choice architectures***. A recent study examined the effects of different consent management "pop-up" windows, using a browser extension that displayed different privacy control pop-ups in real websites [133]. The designs were mimicking the consent management pop-ups that we commonly find in the EU following the introduction of the GDPR. The experiment was conducted with US participants (to avoid familiarity with current European consent practice) and measured the effect of pop-up design on the final privacy decisions of participants. In the experiment, the basic layout of the consent notification had no effect on the final privacy decisions, but other components of the design mattered [133]. Removing the 'reject all' button from the first page of the consent form increased the probability of consent by 22%. The display of granular consent choices on the first page also had effects on consent: Showing a list of granular choices that spelled out the purposes of data use decreased consent by 8%. Showing a list of vendor companies that would access the data decreased consent by 20% and showing the list of both purposes and vendors decreased consent by 11% [133]. These results suggest that the lack of accessible granularity and the absence of simple opt-out buttons in consent forms leads users to share more data than they would when given accessible control over their privacy.

## Key scientific findings

- Online privacy preserves three core components of democratically empowered voters: freedom of association, truth-finding and opportunities to discover new perspectives. More privacy online means a strengthened democracy offline.
- Users are generally unaware of what data they produce, provide to others and how that data is collected and stored when they perform basic tasks on social media platforms.
- Choice architectures are an important determinant of online behaviour.
- Defaults, framing of choices and dark patterns can substantially influence user choices, likely contributing to the privacy paradox and limiting opportunities to discover new perspectives.
- Companies use defaults, framing and dark patterns to prompt the choice of lenient privacy settings and to increase user engagement.

# Chapter 5

## Algorithmic content curation

## Specific characteristics: The dark side of algorithms

## How this affects our behaviour

# Chapter 5: Algorithmic content curation

Navigation of the web is nearly impossible without intelligent algorithms that curate content for us. Organising information and making it accessible is Google's official mission statement.[26] We benefit from those algorithms every time we search for information. However, algorithms frequently also act without our knowledge and involvement, to satisfy our "presumed" preferences. Preference satisfaction turns out to be a double-edged sword. On the positive side, intelligent recommender systems help us find movies or restaurants that we like [134]. YouTube by default continues to play videos automatically without our intervention, selecting one video after another based on what an algorithm deems to be of interest to us. On the more negative side, algorithms also open the door to manipulation and subterfuge. As we have seen in Chapter 3, "microtargeted" political messages that are based on extracting psychological characteristics from digital fingerprints can exploit people's personal vulnerabilities without their knowledge and without public scrutiny or opportunity for rebuttal. The benefits and harms of algorithms thus deserve to be explored in depth.

## Specific characteristics: The dark side of algorithms

Anne dislikes violent movies. Why shouldn't her internet movie provider withhold *Chainsaw Massacre* from a list of offerings? Bob likes to share current-event stories with his friends. Should his social media newsfeed provide him with exciting stories that are untrue but that fit his political views? Where should one draw the line between helpful algorithmic customisation and manipulation or algorithmic selections that run counter to the common good?

***The responsibility gap***. Algorithms make decisions without public oversight, regulation or a widespread understanding of the mechanisms underlying the resulting decisions. Most algorithms are considered proprietary trade secrets and therefore operate as black boxes where neither individual users nor society in general knows why information in search engines or social media feeds is ordered in a particular way [135]. The problem is compounded by the inherent opacity and complexity of machine-learning algorithms [136], such that even creators or owners of algorithms may not be aware of their functioning.

The delegation of choice from humans to algorithms under conditions of opacity and complexity raises questions about responsibility and accountability [137]. Since artificial agents are capable of making their own decisions and since no one has strict control over their actions, it is difficult to assign responsibility for the outcomes. Because the manufacturer or designer of the algorithm cannot predict its future behaviour, it is easy to claim that they cannot be held morally or legally liable for its behaviour. Although designers must be held to account for flaws that could reasonably have been detected by careful testing; the Association for Computing Machinery (ACM) has codified principles on fair use of personal data and fair use of algorithms [138]. This diffuse link between designers' intention and the actual behaviour of an algorithm creates a "responsibility gap" that is difficult to bridge with traditional notions of responsibility [139] and is subject to ongoing debate (see, e.g. the EU's recent statement on artificial intelligence by the Group on Ethics in Science and New Technologies[27]). This gap must be of concern because some algorithms are known to exhibit systematic biases.

[26] https://www.google.com/about/

[27] https://op.europa.eu/en/publication-detail/-/publication/dfebe62e-4ce9-11e8-be1d- 01aa75ed71a1

***Algorithmic discrimination and biases***. Notwithstanding their opacity and complexity, algorithms always have an input and an output. Those two openings can be leveraged into “reverse engineering” an algorithm’s functioning [137], that is, understanding its design based upon its observable behaviour. Reverse engineering can range from the relatively simple (e.g. examining which words are excluded from auto-correct on the iPhone [140]) to the highly complex (e.g. an analysis of how political ads are delivered on Facebook [104]). Reverse engineering therefore offers a tool to shine the daylight of transparency into what is otherwise an opaque but powerful ecosystem.[28]

Reverse engineering of algorithms has consistently identified biases in algorithmic design based upon a number of socio-demographic variables. For example, an early analysis by the *Wall Street Journal* revealed price discrimination by online vendors based on user geography (i.e. distance to a rival’s store) and various other variables [141]. A more disturbing set of examples concerns deeply rooted gender or racial biases that are encapsulated into data-processing algorithms. One study of personalised Google advertisements demonstrated that setting the gender to female rather than male in simulated user accounts, resulted in fewer ads related to high-paying jobs [142]. This finding was confirmed by another field study using Facebook, which found that an ad promoting information about careers in science and engineering was seen by fewer women than men, even though the ad was explicitly intended to be gender neutral in its delivery [143]. This discriminatory delivery occurs because younger women are a prized demographic, which increases the price of ad delivery compared to their male counterparts. Hence, any algorithm that optimises the cost-effectiveness of an advertising campaign will deliver ads that were intended to target both men and women equally in an apparently discriminatory manner. No matter how unbiased the advertiser, the algorithm will introduce a bias through optimisation of a variable (cost of delivery) that happens to differ between genders [143]. This can happen “even when advertisers set their targeting parameters to be highly inclusive” [105, p. 1].

Another study in the US found that online searches for “black-identifying” names were more likely to be associated with advertisements suggestive of arrest records (e.g. “Looking for Latanya Sweeney? Check Latanya Sweeney’s arrests”). Names such as Jill or Kristen did not elicit similar ads even when persons by that name did have an arrest record [144]. Such algorithmic racial biases can have significant consequences in society, for example when they exacerbate inequalities in health care between White and Black Americans [145].

Racial discrimination is also well established for facial recognition algorithms. These algorithms have been found to exhibit low accuracy on non-white faces. For example, in a gender classification task, darker-skinned females were misclassified by the algorithm up to 35% of the time, compared to a maximum error rate for lighter-skinned males of below 1% [146]. Similarly, a study by the National Institute for Standards and Technology[29] found that current facial recognition systems misidentify Blacks and Asians at 10 to 100 times the rate of Whites. Given the increasing use of automated facial recognition in law enforcement, this is a significant cause for concern.

Recent research from the US has revealed that even email is not immune to political fallout from algorithmic sorting. Google’s popular gmail facility automatically sorts incoming messages between different mailboxes. A primary inbox, which is the default focus, receives “the mail you really, really want”,

[28] Although the GDPR has numerous provisions regarding the right to an explanation and right to meaningful information, independent, external audits are the only way to ensure compliance.

[29] https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software

whereas other messages are sent to a promotions folder (for “deals, offers and other marketing emails”) or to spam.

A recent audit revealed that fundraising emails from politicians were frequently redirected from the primary inbox even if a user signed up to receive them. More troubling still, the redirection differed considerably between different political candidates, with 63% of one candidate’s email showing up in the primary inbox compared to 0% for many others (including Joe Biden and Elizabeth Warren).[30]

***Search engines.*** Web search engines have been the focus of studies on information coverage, diversity and bias since the late 1990s [147, 148, 149]. Early scholars recognised that search engines also raised important political questions regarding their control over the flow of information [150]. Indeed, search engine bias can have a substantial impact on important societal decision-making processes [66].

Early work suggested that diversity in user interests might mitigate any search engine biases [151]. More recent work, however, suggests that user interests can exacerbate search engine biases [152]. These problems are not easily rectified, even with training and the best intentions of the user. Pinning down a “correct” query is difficult because web content and search rankings change over time. A query that returns high quality results today might return low quality results next month.

The rapid shift in meaning of search terms arises particularly in the context of gaps in search coverage (“data voids”) that can be exploited by malicious actors [153]. To illustrate, few people ever searched for “Sutherland Springs” before 4 November 2017, when a shooter walked into a Baptist church in the small Texas town and killed 26 people. Because there was little competition for online content about Sutherland Springs at the time (barring weather information, a map and a Wikipedia entry), malicious actors were able to influence search rankings by posting a torrent of material that (falsely) blamed the shooting on the “Antifa” movement. These malicious actors succeeded in shaping the front page of search queries and even injected “Antifa” into auto-suggest. The fictitious link to “Antifa” was picked up by Newsweek and took valuable time to debunk. There is evidence that this is no isolated incident and that white supremacists systematically seek to exploit data voids which can then be filled with extremist material against little competition [153].

Another cause for concern are allegations that Google hard-coded rules in its algorithm to put its own products at the top of the page [154]. For example, Google has been shown to push YouTube videos over those hosted by rivals[31] and in an audit of autocompletions Google was found to add “YouTube” to queries of politicians’ names at twice the rate of Bing [155].

Independent efforts to audit search engines have examined them with respect to a wide range of topics, including personalisation, news, health and discrimination. Personalisation has been explored several times in recent studies, with researchers generally concluding that individual-level personalisation is relatively low in search and that location-based personalisation is a bigger factor [156, 157, 158]. Researchers have examined search results for news content, finding that exposure is generally concentrated among a small number of highly popular outlets [159, 160].

[30] https://www.theguardian.com/us-news/2020/feb/26/gmail-hiding-bernie-sanders-emails-google-inbox-sorting-consequences-2020

[31] https://www.wsj.com/articles/google-steers-users-to-youtube-over-rivals-11594745232

While searching for the names of political candidates, researchers have also found that queries (i.e. user agency), can account for a large share of the variance in the partisanship of the results returned [161]. An additional finding is that Google's web page preview snippets (the short paragraph of text that typically accompanies a link) often amplify the partisanship of the corresponding web page [162]. Further highlighting the importance of users' queries, recent work has found that people formulate queries that contain partisan signals reflecting their ideology [163, 164].

Similar to most other online platforms, search engines can interact with inputs related to race and gender in ways that reflect a society's existing stereotypes and power structures. For example, researchers have found discriminatory advertising practices and stereotypical representations of Black Americans in Google Search [144, 165]. Similar findings have arisen with respect to the autocomplete suggestions that Google provides [166] and auditing methods for interrogating such suggestions have been developed [155].

Even if a search engine can correct for these problems, they are still subject to ongoing attempts by external actors to manipulate their rankings. Such attempts can be seen in prior work on web spam in search, where third parties attempt to surface content for political motives or financial gain [167, 168]. One way to measure resilience to such manipulations is to measure the stability of results over time [169, 170], but without collaboration with a search engine itself, detecting gaming attempts is not possible.

Not all search engines are created equal. A 2019 study by the Stanford Internet Observatory compared the performance of Google to Microsoft's Bing search engine.[32] In general, Bing was found to return disinformation and misinformation at a significantly higher rate than Google; it directed users to conspiracy-related content without being prompted by specific search terms; and it returned white-supremacist content in response to unrelated queries. There are, however, methodological limitations (e.g. a limited number of search queries) that suggest caution in interpretation of these results. Nonetheless, they are consistent with earlier research which found that searches related to suicide themes on Google would yield the phone number for the National Suicide Prevention Lifeline as the top item, whereas Bing would return methods for committing suicide among the top search results.[33]

***Recommender systems and algorithmic rankings***. YouTube boasts over 2 billion users[34], making it the second most visited website worldwide. At the heart of YouTube's architecture is a sophisticated recommender system that is designed to maximise viewing time on the platform [67]. The system learns approximately one billion parameters and is trained on hundreds of billions of cases. One consequence of the recommender system is that it tends to offer viewers more extreme content at every step. For example, users who viewed videos of Donald Trump during the 2016 presidential campaign were subsequently presented with videos featuring white supremacists and Holocaust denialists. After playing videos of Bernie Sanders, YouTube suggested videos relating to left-wing conspiracies, such as the claim that the US government was behind the September 11 attacks [171]. A recent preregistered study of the YouTube recommender system confirmed that it was liable to promote and amplify conspiratorial content even in response to relatively innocuous search terms [172].

Problems with the YouTube recommender system also arise outside the political arena. Recent research has shown that young children are likely to encounter "disturbing" videos (i.e. clips containing

[32] https://cyber.fsi.stanford.edu/io/news/bing-search-disinformation

[33] https://www.vice.com/en_au/article/nn97jk/how-google-searches-influence-suicides-511

[34] https://www.youtube.com/intl/en-GB/about/press/

inappropriately violent or sexual content) on YouTube when they randomly browse YouTube, starting from a benign video [173].

A particularly problematic feature of these recommendations is the "autoplay" setting on YouTube, which automatically starts playing the next video in a recommended sequence without requiring user input (turning off the autoplay feature requires a relatively complicated sequence of steps[35]).

There is now evidence suggesting that YouTube algorithms may have actively contributed to the rise and consolidation of right-wing extremists in the US [174] and Germany [175]. This concern about radicalisation was buttressed by a recent large-scale quantitative audit of the YouTube recommender system [176]. An alternative view on radicalisation argues that YouTube has other features besides the recommender system, such as monetisation of content, that facilitate content creation by fringe political actors [177] (see also, [178]). On balance, we are not aware of scientific dissent from the position that YouTube's design—i.e. mainly but not exclusively its recommender system — facilitates radicalisation and exposure to extremist content.

In response to mounting criticism, YouTube recently vowed to limit recommending conspiracy theories on its platform [179]. This move may be welcome in light of the demonstrably adverse effects of conspiracy theories on the public [180], but it also highlights industry's unilateral power to shape information diets, which at present is only challenged by academic research and investigative journalism. In this context, it is noteworthy that a former software engineer at YouTube has accused the company of shutting down an algorithm that had been designed in 2010 to insert more diversity into the recommendations because it reduced viewer time.[36]

**Curated social media newsfeeds**. Social media newsfeeds have become a ubiquitous feature of life. We inform ourselves about anything from family events and friends' adventures to political developments by checking Facebook, Instagram or Twitter (to name but a few platforms). In 2019, active social media penetration across the EU ranged from 88% in Malta to 46% in Germany. Overall, 48% of EU citizens used online social media networks every day or almost every day. The average daily length of time of social media use (via any device) ranged from 129 minutes per day in Portugal to 64 minutes in Germany. In Germany, more than 15 million Instagram users follow influencers Lisa & Lena. 54% of Finns use Snapchat several times a day, while in France 61% of 8-14 year-olds have a Snapchat account. In the Netherlands, from January-February 2020, the TikTok app was downloaded more than 600,000 times.[37] During the peak of the COVID-19 pandemic, a global survey found that 20% of respondents indicated that they would continue to spend more time on social media even after the pandemic [181]. Figure 3 (below) summarises internet penetration and associated social media use across all 27 EU Member States.

[35] https://www.lifewire.com/turn-off-autoplay-on-youtube-4178239

[36] https://www.nbcnews.com/tech/social-media/algorithms-take-over-youtube-s-recommendations-highlight-human-problem-n867596

[37] All data in this paragraph are from Statista's report *Social media usage in the European Union (EU)*; https://www.statista.com/study/32424/social-media-usage-in-the-european-union-eu-statista-dossier/

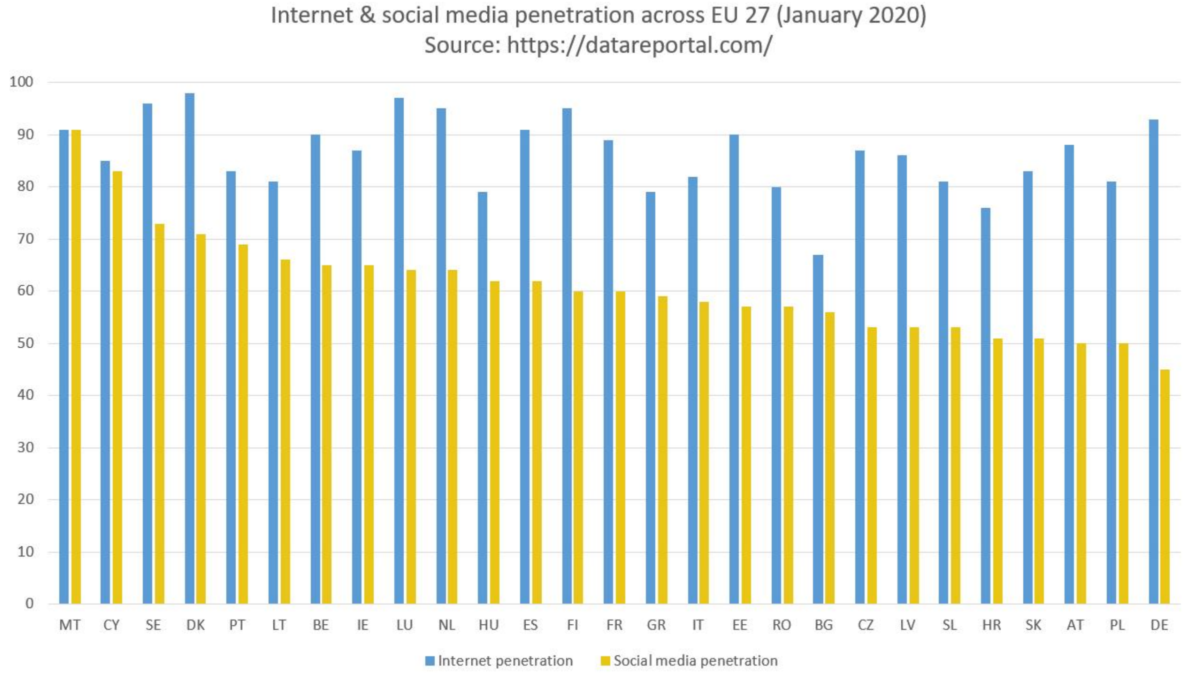


*Figure 3* - Internet and social media penetration across EU 27 (January 2020). Source: https://datareportal.com/

The public's habitual interaction with social media has likely dulled us to the fact that newsfeeds have become one of the most sophisticated algorithmically driven features of online platforms [182]. Most platforms order pieces of content according to attributes that are branded as "trending," "hot," "popular," or "new." Behind those branding labels is an algorithmic score that integrates some measure of popularity of a post, together with other variables such as its age and the proximity in the social networks of its originator. The exact combinations and algorithms are generally not public, even though sorting can be crucial for popularity dynamics [183].

Many sites, including Twitter, Reddit and Facebook, preferentially show content that has seen more engagement from other users or which the algorithm expects to yield high engagement. It follows that the algorithm will favour extreme, emotional and humorous content because it is often more engaging than generic news or shared personal news. Similarly, engagement (e.g. through disagreement) may amplify more polarising content over less polarising, community-building content. Newsfeeds consequently, may expose users disproportionately to polarising content or content emphasising disagreement.

In consequence, there has been much concern and public debate that social media platforms may be responsible for "echo chambers" [184] or "filter bubbles" [185]. Here, we use the term "echo chamber" to refer to environments in which people are only exposed to information from like-minded individuals (formed mainly by people's purposeful avoidance of opposing views). We use the term "filter bubbles", by contrast, to refer to algorithmic content selection according to a viewer's preferences as revealed by prior behaviour [186].

The adverse consequences that are thought to arise from echo chambers and filter bubbles are increasing political polarisation and radicalisation [187]. The existence of filter bubbles and echo chambers is subject to considerable academic debate. To put this debate on a solid footing requires a careful disentangling of the roles of algorithms that automatically curate a newsfeed (filter bubble) and people's choice to avoid (or denigrate; [188]) opposing views (echo chamber), as well as the interaction between human and algorithmic biases. For example, in principle it is possible for there to be little evidence for the existence of filter bubbles (i.e. people's news offerings may be less segregated than feared) even though evidence simultaneously reveals the existence of echo chambers (i.e. people may trust different content, sources or other users, so that even if people are exposed to a variety of content, uptake of that content may differ between people).[38] In support of this possibility it has been shown that even in completely connected networks with no curated newsfeeds (i.e. no filter bubbles), polarisation may appear because people treat evidence shared by some of their connections as less trustworthy than evidence shared by others [189]. These dynamics can lead to factions of users who share otherwise-unrelated polarised beliefs [190]. In such cases, echo chambers may be effectively present because users ignore or distrust some content to which they are exposed.

***The entanglement of algorithm and user***. Due to the complex interactions between algorithmic and human behaviour, a causal link to political preferences is difficult to establish [191, 192]. The complexity of interaction is best illustrated by considering a seemingly simple web search. When analysing potential search engine biases, it is difficult to tease apart confounding factors inherent to the scale and complexity of the web [193]. One factor involves the constantly evolving metrics "relevance" that search engines optimize for [194]: Google is known to change its algorithms hundreds of times each year.[39] Another factor relates to efforts to counter and prevent gaming by vested parties [169]. This Search Engine Optimization (SEO) can take many forms, ranging from the legitimate (e.g. linking to high quality sites) to manipulative (e.g. hiding unrelated high-traffic terms in the webpage to maximise its reach).

The problem is compounded by the observation that how users formulate and negotiate search queries can vary by political party [164, 152]. These behavioural differences, in turn, can have a large impact on the partisanship of the sources users encounter on the web. For example, query selection has been shown to substantially affect audience-based partisanship in both Google [161] and Twitter search results [195].

The process of generating queries — and the ensuing search results — thus involves not only the information retrieval algorithms at play, but also the cultural and political history of the human interacting with them [194, 152]. It appears that human biases are detected by algorithms and rewarded by a more biased offering of information.

Studies involving real user data are rarely conducted outside of corporate research labs [186], due in part to proprietary and privacy-related concerns [196]. The lack of up-to-date research on this topic is also partly due to the ever-evolving nature of users' information needs [197, 152] and the opaque interactions between users and autocomplete algorithms that influence the process of query selection [155], both of which require longitudinal study [198].

Evidence relating to selective exposure and engagement. Polarisation is on the rise in the US [184] and some (though not all) European countries [199]. Identifying the source of this trend is a difficult task, in light of the challenges involved in teasing apart the role of human agency and algorithmic curation. Do online environments really exacerbate polarisation or do they just reflect offline patterns of behaviour? Previous analyses have suggested that the evidence is inconclusive [200, 201].

A study using data from the American National Election Studies (ANES) found that polarisation has increased the most among demographic groups least likely to use the Internet and social media [202]. In

[38] Echo chambers may persist even if there is diversity of content, if non-aligned content is systematically denigrated [188]. Echo chambers may therefore be strengthened by the presence of oppositional viewpoints.
[39] https://moz.com/google-algorithm-change

a three-month study of 50,000 American online news consumers, the ideological distance between the news diets of a random pair of consumers was found to be small [203].

With respect to social media, several studies highlight the role of user choices and network structure as primary sources of polarisation. On Facebook, people tend to befriend people with similar politics and less than half of the content people are exposed to and engage with are from ideologically-different sources [186]. A 2015 study of retweet networks for 3.8 million Twitter users found that discussions on politically salient topics are often highly fragmented — occurring among users with the same ideological preferences — while discussions on other topics, like entertainment and sports, are more inclusive [204]. These results align with earlier studies of polarisation in blogs, where blogs with a similar ideological slant were more likely to link to each other than to blogs with the opposite slant [205].

Even when people are exposed to ideologically varied information on social media platforms [186, 204], the interactions resulting from such exposures are often vitriolic. For example, someone might post an article that is from a source coded as politically left-leaning, but add their own text harshly criticising it. It is important to note that most methods of measuring partisanship will miss this nuance. Moreover, exposure to ideologically-different information under such partisan criticism would result in increasing polarisation [206, 207] and would solidify existing echo chambers [188].

In controlled experiments, people generally do not choose information that conforms to their own views at the expense of information that contradicts held opinions. When presented with items that either confirm opinions held or are balanced, people are largely indifferent to the inclusion of opposing viewpoints. It is only when their own opinions are not represented at all, that they reject information [208]. By contrast, using real user data, Facebook researchers found that people are more likely to engage with ideologically consistent news sources, but concluded that this was due more to user choices (e.g. whom they friended) than Facebook's algorithms [186]. When the data are broken down by partisanship, liberals were more likely to encounter cross-cutting content in the newsfeed than conservatives [186]. These results were broadly replicated in a recent analysis [209], which additionally found that the share of Republican respondents whose media diets were much more conservative than those of the rest of the sample increased between 2015 and 2016.

This asymmetrical theme carries through other "big-data" analyses, with evidence for potential echo chambers of misinformation frequently emerging among particularly strong conservatives. For example, researchers reported evidence of substantial selective exposure to fake news, with Trump supporters consuming more news from untrustworthy websites than others [210]. The research team concluded that "echo chambers are deep (52 articles from untrustworthy conservative websites on average in this subset) but they are also narrow (the group in question represents only 20% of the public)" [210, p. 6]. Similarly, a recent large-scale study of Twitter users has found that sharing of misinformation was disproportionately concentrated among older Republicans [211]. This replicates a previous large-scale analysis of Twitter that also concluded that echo chambers exist or form as polarisation kicks in [204]. The same study also detected ideological asymmetry, with liberals being less likely to be caught in an echo chamber [204]. Nonetheless, non-political issues and events were found to be discussed without regard to partisanship [204].

Not all questions about echo chambers and filter bubbles have a clear answer. What is clear is that polarisation exists in society and that it exists online and is manifest in fragmented online spaces or highly selective media diets, that are justifiably called "echo chambers". Although those echo chambers may be narrow and limited in size [210], members of those chambers have been found to be most likely to vote, implying that “even if most Americans do not exist in online echo chambers, they are subject to the political influence of those who do” [209, p. 28].

What remains less clear is precisely how much of the blame for echo chambers can be attributed to algorithms or specific platforms. Human behaviour and algorithmic biases are entangled in a — frequently — reinforcing feedback loop that makes apportioning responsibility difficult. What is clear, however, is that algorithms do not militate *against* online polarisation—even though recent reports suggest that corrective technology exists but was rejected by Facebook [108].

## How this affects our behaviour

**The power of search engines**. The power of search engines to affect people’s perceptions has been demonstrated repeatedly in experiments. For example, Google search results have been shown to affect attitudes and beliefs about vaccinations [212]. Other experiments have shown that search ranking can impact attitudes about science controversies and that participants rated top-ranked pages as the most useful [213].

The results of such experiments are particularly pertinent and concerning in the political domain. In those studies, participants were presented with mock search engine results that were biased in favour of one politician at the expense of another. For example, it has been shown that (simulated) search engine rankings that favour a particular political candidate can shift voting preferences of undecided voters by 20% or more [66]. Even when participants are given very detailed warnings about possible ranking biases, the effects of the bias were reduced (to 14%) but not eliminated [214]. The only time the effects of rankings were eliminated in these experiments, was when the results alternated between the candidates, essentially an equal time rule.

Taking the US as an example, given that half of American presidential elections are decided by margins under 7.6% [66], the impact of potential search-engine biases should not be ignored.

## Key scientific findings

- Algorithms are an indispensable feature of the web that can be used or abused in relation to enhanced user satisfaction, engagement and awareness.
- Curated newsfeeds and automated recommender systems are designed to maximise user attention by satisfying their presumed preferences, which can mean highlighting polarising, misleading, extremist or otherwise problematic content to maximise user engagement.
- Newsfeed rankings, search engine ordering and recommender systems can causally influence our preferences and perceptions.
- The evidence for filter bubbles (i.e. algorithmic segregation of users' information content), is ambivalent, but there are legitimate serious concerns about echo chambers (i.e. formed through self-selection of content by users).

# Chapter 6

## Misinformation & disinformation

Specific characteristics: What is "post-truth"?

How this affects our behaviour: Receptivity to misleading information

14:59 PM
NEWS
FAKE
HEALTH DEPT.: WARNS OF DANGERS RELATED TO FLOODING
James Arches continues fighting to get his savings back
Kidnapping victims break silence on video
Egyptian leaders lay out plan for fast transition
The technological edge
MARKETS

# Chapter 6: Misinformation and disinformation

“Post-truth” was nominated word of the year by Oxford dictionaries in 2016, to describe “circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief.”40 A year later, Collins dictionaries declared “fake news” as word of the year, to refer to “false, often sensational, information disseminated under the guise of news reporting.”41 These choices illustrate the massive current level of concern, in the public sphere as well as in political and academic circles, about misinformation.

False information, political manipulation and online harassment are global public concerns. In the European Union, in every Member State, at least half of respondents in a large sample (N = 26, 576) say they come across fake news once a week or more [215]. In the US, 89% of adults (N = 6, 127) indicate that they come across made-up news intended to mislead the public at least sometimes [216]. In politics and policymaking, countless reports and recommendations have been issued about the threat to democracy arising from false information, although few have translated into policy. In the scientific literature, the increase in research attention has been equally striking. Whereas only 73 articles with “fake news” in the title had been published in all the years leading up to January 2017, since then 2,210 articles have appeared in the literature [217].

## Specific characteristics: What is “post-truth”?

Persistent concern about the widespread effects of misinformation are legitimately fuelled by a number of scientific results and present platform policies. These concerns are particularly serious if one considers the full spectrum of misinformation, from “fake news” to misleading statements by politicians, rather than just the fabrications emanating from a small number of “fake news” websites, which are consumed only by a small but arguably significant segment of the public [217, 211, 218, 219, 210].

***Platform policies.*** Turning first to platform policies, there are at least two reasons for concern: First, the *actual* platform policies and second, the volatility of those policies. Concerning the former, until recently Facebook had an explicit policy against fact-checking political advertisements.[42] Concerning the latter, at the time of this writing (July 2020), Facebook had just given customers the option to opt out of receiving political ads, at least in the United States.[43] Facebook has also taken down posts and ads for President Trump’s re-election campaign because they violated the platform’s policy against “organized hate,” marking the

> *“Those who can make you believe absurdities, can make you commit atrocities.”*
>
> — Voltaire

[40] https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016
[41] https://www.collinsdictionary.com/woty
[42] https://time.com/5762234/facebook-political-ads-election/
[43] https://www.nytimes.com/2020/06/16/technology/opt-out-political-ads-facebook.html

latest confrontation in an escalating battle over how tech companies handle controversial political content.[44]

Facebook is not the only platform exercising decisions about political content, Twitter has also recently begun to mark some of Donald Trump's tweets as false or misleading,[45] and has prevented retweets without comments, thus curtailing the President's ability to cause a social media cascade.

The theme that cuts across these two sources of concern is that we live in tumultuous times in which platform policies can appear not only arbitrary or inappropriate in the eyes of many, but in which they can also change at a moment's notice and without public conversation or accountability. This has not escaped key actors who are recognising the need for leadership and direction. For example, Facebook's CEO Mark Zuckerberg expressed an interest to "partner more closely with governments not just based on what is written into law but proactively understand — what they would like to see us do" in an interview with the European Commissioner for the Internal Market, Thierry Breton.[46]

***Agenda setting power of disinformation***. Even a small dose of fake news can set agendas through "its ability to 'push' or 'drive' the popularity of issues in the broader online media ecosystem" [220, p. 2043]. A recent study showed that although fake news did not dominate the American media landscape in 2014 – 2016, fake news was intertwined with partisan media (e.g. Fox News) and influenced news agendas across a wide range of topics, including the economy, education, the environment, international relations, religion, taxes and unemployment [220]. Similarly, extreme-right outlets such as Breitbart have been shown to alter the broader agenda of the media [221].

***Incentivising extremism***. The segmentation and polarisation of the media [222] (both online and offline) and the emergence of partisan outlets have also created a reward structure for politicians to engage in strategic extremism [47]. The agenda-setting power of fake — and often extremist — material is therefore further amplified by politicians who, quite rationally, seek to maximise their electoral success. Although conventional wisdom holds that vote-maximising politicians should cater to the middle by chasing the "median voter" [223], extremism is rewarded when a politician gains more from energising their own supporters and gaining supporters on the fringe, than they lose by alienating median or opposing voters. This relative benefit of extremism can only occur when awareness of a politician's message is higher among his or her supporters than it is among the opponent's supporters. The existence of influential, highly partisan media provides this opportunity for pragmatic extremism. We further examine the mechanisms by which extremist content can become part of mainstream discourse in Chapter 7.

Polarisation and strategic extremism may also be amplified by a striking feature of online disinformation in the past five years. Since the lead-up to the UK's Brexit referendum in 2016, sophisticated actors have honed the ability to create misleading content that is likely to be shared virally by ordinary users; often they are not aware of its source (or falsity) [107]. This dissemination strategy exploits existing social factors, including trust, affinities and conformity, to maximize the uptake of misleading content. Thus, disinformation may appear from known, trusted sources within one's social network, even though it was originally generated by unknown and often malicious third parties.

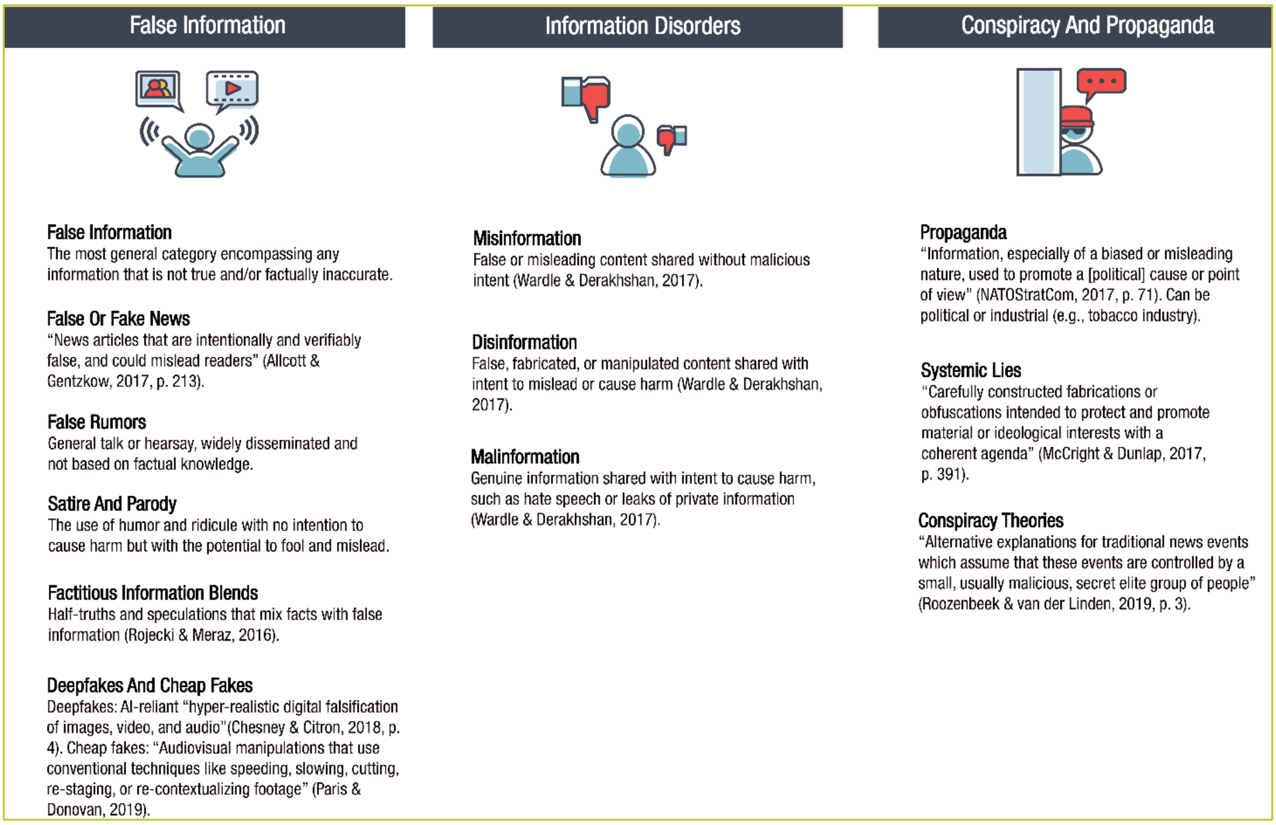
This means that even on platforms where (non-sponsored) content can be shared only by existing connections, such as Facebook, third-party disinformation may appear because another user has decided

[44] https://www.wsj.com/articles/facebook-removes-trump-campaign-posts-ads-for-violating-policy-11592504003
[45] https://techcrunch.com/2020/05/29/twitter-screens-trumps-minneapolis-threat-tweet-for-glorifying-violence/
[46] https://www.youtube.com/watch?v=uZfi6WkIfgU&feature=youtu.be

to share it. This “participatory propaganda” will be explored further below. The perception that one is interacting only with known contacts can make users more susceptible to disinformation, precisely because it does not directly originate with a malicious actor but gives the impression of a friend’s endorsement.

We now expand our focus to include the multitude of other forms of false or misleading information the public is exposed to. We begin by differentiating between different types of false or misleading information. This analysis focuses mainly on the supply-side of this information. It is followed by an examination of the demand side; that is, an analysis of the variables that render the public susceptible to the consumption of misinformation.

***Taxonomies of misleading information***. Not all false information is equal. Several different classifications of false or misleading information have been proposed, invoking several different attributes or dimensions (Figure 4 provides an overview).



False Information

**False Information**
The most general category encompassing any information that is not true and/or factually inaccurate.

**False Or Fake News**
“News articles that are intentionally and verifiably false, and could mislead readers” (Allcott & Gentzkow, 2017, p. 213).

**False Rumors**
General talk or hearsay, widely disseminated and not based on factual knowledge.

**Satire And Parody**
The use of humor and ridicule with no intention to cause harm but with the potential to fool and mislead.

**Factitious Information Blends**
Half-truths and speculations that mix facts with false information (Rojecki & Meraz, 2016).

**Deepfakes And Cheap Fakes**
Deepfakes: AI-reliant “hyper-realistic digital falsification of images, video, and audio”(Chesney & Citron, 2018, p. 4). Cheap fakes: “Audiovisual manipulations that use conventional techniques like speeding, slowing, cutting, re-staging, or re-contextualizing footage” (Paris & Donovan, 2019).

Information Disorders

**Misinformation**
False or misleading content shared without malicious intent (Wardle & Derakhshan, 2017).

**Disinformation**
False, fabricated, or manipulated content shared with intent to mislead or cause harm (Wardle & Derakhshan, 2017).

**Malinformation**
Genuine information shared with intent to cause harm, such as hate speech or leaks of private information (Wardle & Derakhshan, 2017).

Conspiracy And Propaganda

**Propaganda**
“Information, especially of a biased or misleading nature, used to promote a [political] cause or point of view” (NATOStratCom, 2017, p. 71). Can be political or industrial (e.g., tobacco industry).

**Systemic Lies**
“Carefully constructed fabrications or obfuscations intended to protect and promote material or ideological interests with a coherent agenda” (McCright & Dunlap, 2017, p. 391).

**Conspiracy Theories**
“Alternative explanations for traditional news events which assume that these events are controlled by a small, usually malicious, secret elite group of people” (Roozenbeek & van der Linden, 2019, p. 3).

*Figure 4* - Types of false and misleading information. Adapted from [24].

One important dimension characterising false information is whether or not it was intended to mislead. One analysis distinguishes between three types of “information disorders” [224]: misinformation (false or misleading content created and initially presented without malicious intent), disinformation (false, fabricated or manipulated content shared with intent to mislead or cause harm) and mal-information (genuine information shared with intent to cause harm, such as hate speech and leaks of private information) [224]. Intentional generation of false content is financially lucrative: The Global Disinformation Index recently estimated that online ad spending on disinformation domains amounted to $235 million a year [225].

Although intention can be challenging to discern, sometimes it is the only variable that differentiates fake news from satire. Most forms of satire should not be considered disinformation even if its content is technically false [226].

There is, however, one critical qualification to the endorsement of satire. There is evidence that irony, "in-jokes" and satire have been weaponised by extreme-right actors [227]. It has been argued that "the far-right exploits young men's rebellion and dislike of 'political correctness' to spread white supremacist thought, Islamophobia and misogyny through irony and knowledge of internet culture" [228, p. 11]. Under the umbrella of ironic ambiguity, many alt-right actions — such as the propagation of Nazi symbols, the use of racial epithets or spreading of racial slurs — effectively garner support for white supremacist ideologies.[47]

This is of concern for a number of reasons: First, machine-learning algorithms are typically unable to differentiate irony from sincere information, which renders automatic detection of extremist messages more difficult. Second, the European Commission's Code of Practice on Disinformation[48] explicitly excludes "misleading advertising, reporting errors, *satire and parody* or clearly identified partisan news and commentary" (emphasis added). Thinly-veiled abuse of the right to expression is therefore difficult to identify or regulate.

Another challenge to categorising false content according to intent is that the intentions of the content creator and those of the users who share the content may be different. For instance, content may be created by malicious actors intending to manipulate political beliefs; but shared by users who believe and are sincerely alarmed by the content, by users who are unsure what to think and hope to prompt discussion or commentary from others (e.g. because if it *were* true it would be interesting;

> *"If everybody always lies to you, the consequence is not that you believe the lies, but rather that nobody believes anything any longer. And a people that no longer can believe anything cannot make up its mind. It is deprived not only of its capacity to act but also of its capacity to think and to judge. And with such a people you can then do what you please."*
>
> — Hannah Arendt, German-American philosopher and political theorist

[47] https://dailystormer.su/a-normies-guide-to-the-alt-right/

[48] https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

[229]) or even by users whose principal goal is to fact-check or refute the content, but who thereby expose other users to it. In such cases, content that may begin as disinformation may morph into misinformation as the intentions of those who repackage it or share it change. A particularly pernicious form of this packaging of malicious intent in a seemingly innocuous format is the "just asking questions" strategy employed by conspiracy theorists [230]. This strategy creates legitimate space for a conspiracy narrative while maintaining some degree of respectability. It has been used by prominent American cable news provocateurs, for example when raising questions about the deaths of US troops in Niger [230].

Conversely, content created in a sincere effort at truth-telling, but which happens to be false or misleading when presented out of context, may be shared or promoted by malicious actors in a deliberate effort to mislead. This sort of strategy often occurs when political or economically-motivated actors cherry-pick legitimate scientific results to support their preferred position. Versions of this strategy have been widely used by industry groups for decades, such as when the tobacco industry sought to undermine the growing scientific consensus that smoking causes cancer [231, 232]. Various analyses have shown that the strategy can be effective even when it uses only legitimate scientific results [233].

Social media can make sharing legitimate content in misleading ways easier. During the COVID-19 pandemic, both traditional pro-business media and some social media influencers have disproportionately shared apparently-legitimate research purporting to show that COVID-19 has a substantially lower infection fatality rate than the World Health Organization has estimated, whereas those with different policy preferences have disproportionately shared studies apparently showing the disease is more deadly. In such cases, it would be wrong to say that the underlying scientific articles are misinformation, much less disinformation or mal-information, even if their conclusions are in fact false or their reported data are outliers. And yet, some sharers of that content do intend to mislead or manipulate.

Yet another class of false content that does not fall under the above categorisation are retracted or corrected articles that continue to propagate as if they were true even after they have been refuted or acknowledged as false by their authors. In the context of the COVID-19 pandemic, these articles have been called "information zombies" because they continue to spread on social media after they have been killed [234]. Prominent examples include a study by a group of Indian scientists claiming an "uncanny" similarity in genetic material between COVID-19 and HIV, which was withdrawn two days after it was posted on January 31, 2020, but which was tweeted more than 20,000 times as of May 2020; or an April 17 op-ed from the *Wall Street Journal* reporting results from a study of the prevalence of antibodies in a Santa Clara County population that were later substantially revised by the study authors, without any corresponding change to the op-ed—which in turn continued to be shared and promoted on Facebook for weeks after the revisions. In such cases, the content is not "misinformation" in the sense that it may well represent the best information available to the authors at the time; and yet as it continues to spread, it becomes misinformation or even disinformation.

Other analyses have taken for granted an intent to mislead the public and focused on the techniques and goals of the manipulator. For example, one study [235] classified misinformation along two dimensions, one pertaining to style and primary audience (which ranges from an informal, conversational style directed toward people's daily lives, to a formal, persuasive style aimed at institutions and systems) and one describing the underlying ontology of truth (which ranges from strong realism with the acceptance that truths exist and a respect for facts to strong constructivism where there is disbelief in the existence of external truths and a disrespect of facts). The latter dimension is of particular interest, because it neatly captures the evolution of misinformation during the last two decades. At one end, there are carefully curated deceptions whose purpose was to convince the public of a non-existent state of reality.

Perhaps the most famous example involves claims surrounding the existence of weapons of mass destruction (WMDs) in Iraq prior to the invasion of 2003. Although there is now much evidence [236, 237, 238] that the US and UK governments engaged in deception to construct those claims, there is little evidence of outright fabrication by UK officials [238]). The US and UK governments therefore displayed an ontological commitment to a form of realism. That is, they accepted that there was a ground truth and they relied on empirical notations, such as "evidence" or "intelligence," to contest the state of that ground truth in Iraq. The fact that Iraqi reality turned out to be different does not negate the fact that the WMD campaign contested a reality whose existence was acknowledged by all parties — governments, the public and U.N. weapons inspectors. Other examples of such carefully curated falsehoods involve climate change denial and other organised campaigns to convince the public of a state of affairs that is untrue but convenient for the campaigners. These curated campaigns can be highly successful; a poll from December 2014 found that around 40% of the American public continued to believe that WMD *were* found in Iraq.[49]

At the other end of the spectrum is a "shock and chaos" regime of falsehoods that seems to have given up on the notion of a shared reality and instead relies on an extreme form of constructivism in which "truth" is entirely in the eye of the beholder. This disinformation regime is characterised by indifference to the truth and a blizzard of erratic, often contradictory, messages. One striking example is the Russian government's response to the downing of Malaysian Airlines flight MH17 in 2014 by a Russian-made Buk missile. Sputnik, RT (Russia Today) and other pro-Kremlin websites first denied it was a Russian missile. Then they said the missile was fired by Ukrainians. Then they said the pilot had deliberately crashed the airliner and the plane had been full of dead bodies before impact. Finally they said it was all part of a conspiracy against Russia [239]. In the Western world, the shock and chaos regime is illustrated by Donald Trump and his supporting infrastructure [240]. According to the Washington Post, Trump has made in excess of 20,000 false or misleading statements during his presidency to date (July 2020[50]). One notable attribute of many of these false statements is that, unlike the more nuanced claims about WMDs based on government intelligence, they are readily and rapidly shown to be false. Indeed, some of Trump's claims, for example that people went out in their boats to watch Hurricane Harvey,[51] have an almost operatic quality and are not readily explainable by political expediency. This type of misinformation is not carefully curated but is showered onto the public as a blizzard of confusing and often contradictory statements. Incoherence and internal contradictions have also been shown to be rife with conspiracy theories relating to COVID-19 [241]. A RAND corporation report has referred to shock-and-chaos disinformation as the "firehose of falsehood" propaganda model [242].

This apparent shift over time in the predominant mode of misinformation, from tacit realism to extreme constructivism, has important consequences that must be understood for countermeasures to be successful.

*The changing ontology of truth and its public dimension*. The changing underlying ontology of truth can be illustrated by examining the responses to challenges. When no WMDs were found in Iraq after the invasion of 2003, this gave rise to multiple inquiries on both sides of the Atlantic and the absence of WMDs

[49] https://view2.fdu.edu/publicmind/2017/

[50] https://www.washingtonpost.com/politics/2020/07/13/president-trump-has-made-more-than-20000-false-or-misleading-claims/

[51] https://bit.ly/2A66Hc3

ultimately became acknowledged bipartisan reality in Washington. By contrast, in the current shock and chaos regime, challenges are not met with counter arguments, inquiries or even defences. Instead, Trump's spokespersons, for example, have repeatedly sidestepped accountability by postulating an explicitly constructivist view of the world.

These claims have quite explicitly repudiated the idea of external truths that exist independently of anyone's opinion. Thus, Trump's counselor Kellyanne Conway famously declared that she was in possession of "alternative facts." Such deflections are not isolated occurrences but arguably form a pervasive pattern that has been labelled "ontological gerrymandering" [48]. Ontological gerrymandering has culminated in Donald Trump's systematic invocation of labels such as "fake news" and "fake media" when describing news stories or organisations he dislikes. Arguably, any critical reporting of his actions are thus dismissed not by argument but by ontological gerrymandering [243].

Ontological gerrymandering is not confined to the US: When a British far-right personality's claim that a recent car accident in London had been a terrorist incident was challenged, she dismissed the correction as "blatant state propaganda" and added: "I have no belief in fact. Fact is an antiquated expression. All reporting is biased and subjective. There is no such thing as fact any more. There is no truth, only the truth of the interpretation of truth that you see".[52] The ontological gerrymandering is not coincidence, but lies at the heart of populism and its "black vs. white" view of the world as a binary conflict between "the people" and its enemies [244]. Those enemies may be the "elites" or other out-groups such as immigrants (or both). Resulting from this binary view is the affirmation of "common sense" truths against "elite" lies. This fundamental premise of populism of an eternal conflict between "the people" and "the elites" creates a self-sealing epistemic landscape in which "critics can never offer facts that question, challenge, or complement populist assertions. Populism's view of good people and bad elites is immune to factual corrections and nuances" [244, p. 26]. Instead, populists negate the possibility of truth-seeking as a shared goal of a society [244]. The disregard for facts exhibited by Trump and other populist politicians must therefore be understood as a necessary consequence, rather than an incidental by-product, of their ideology. Populist conceptions of truth and the ensuing ontological gerrymandering, are incompatible with liberal-democratic norms of truth-seeking.

One putative consequence of shock-and-chaos disinformation and the associated gerrymandering of the ontology of truth is that people become sceptical of truth itself [245, 246, 49]. As political activist Garry Kasparov put it, "The methodology of [shock-and-chaos disinformation] isn't to convince anyone exactly what the truth is, but to make people doubt that the truth exists, or that it can ever be known" [247]. Shock and chaos disinformation that undermines people's belief in truth appears to be custom-designed for the online attention economy.

***Dissemination of false information***. Two core attributes of the attention economy and human psychology combine to form the perfect conditions for the spread of false information. On the technological side, social media algorithms are designed to promote content that is most likely to attract user attention and engagement, and is most likely to be shared. As we showed in Chapter 5, algorithms are optimised with respect to those goals irrespective of whether the content benefits the user or the recipient of a share or whether it is true [46]. On the human side, there is strong evidence from around the world that audiences, on average, seek news that is predominantly negative [248] or awe inspiring [249]. This negativity bias may underlie the conventional lore among journalists that "if it bleeds, it leads." It is also known that people are more likely to share or retweet messages featuring moral-emotional

[52] https://bit.ly/2IOfRRi

language [250]. Moral-emotional language can thus increase the rate of spread of political messages, although this tends to be confined to communities of like-minded people and is less of an issue between such communities [250].

Those known psychological attributes appear to be amplified further by digital media: The degree of moral outrage elicited by reports of immoral acts online has been found to be considerably greater than for encounters in person or in conventional media [46]. Whether by design or coincidence, false online content appears to exploit this specific conjunction of technological and psychological factors. In a content analysis of 150 fake and real news items, fake news titles were found to be substantially more negative in tone than real news titles [251].

The text of fake news was found to be substantially higher in displaying negative emotions, such as disgust and anger that are known to elicit outrage. Fake news texts were also lower in positive emotions, such as joy [251]. Another hot button that fake news can press is the human attraction to novelty and surprise. Anything that is new, different or unexpected is bound to catch a person's eye. Indeed, neuroscientific studies suggest that stimulus novelty makes people more motivated to explore [252]. False stories have been found to be significantly more novel than true stories across various metrics [253]. There is evidence that people noticed this novelty, as indicated by the fact that false stories inspired greater surprise (and greater disgust) [253]. One interpretation of these findings is that falsehood's edge in the competition for limited attention is that it feeds on an — otherwise highly adaptive — human bias toward novelty.

## How this affects our behaviour: Receptivity to misleading information

Online disinformation has often been referred to as involving an "arms race". As users become savvy to various forms of disinformation, those seeking to mislead innovate, figuring out new, effective ways to do so. By implication, innovative media formats play key roles in successful, ongoing disinformation. Some current and emerging forms of media that are playing this role involve memes, especially humorous ones, conspiracy videos as well as "deepfake" images and videos.

Visual formats for misinformation dissemination may be particularly effective as they can be attention grabbing. In simulated social media environments, accompanying text with a photo increases sharing [254]. People have also been found to judge claims accompanied by photos more likely to be true [255]. Furthermore, the use of photographs and videos out of context is often an effective way to mislead [256]. Images, whether real or doctored, can also be very effective at eliciting emotional responses from viewers [257, 258]. Studies indicate that highly emotional content may "spread" more effectively on social media platforms [250]—although it is important to delve deeper into what it means for information to "spread."

Misinformation does not spread on its own. It is spread by people. A tweet by Donald Trump may reach millions of his followers, but it is the retweets, sometimes numbering in the thousands or more, that multiply the reach of the information. The architecture of social media thus permits the emergence of a misinformation ecosystem that has been referred to as participatory propaganda [259]. "While news is constructed by journalists, it seems that fake news is co-constructed by the audience, for its fakeness depends a lot on whether the audience perceives the fake as real. Without this complete process of deception, fake news remains a work of fiction. It is when audiences mistake it as real news that fake news is able to play with journalism's legitimacy" [226, p. 148].[53]

[53] See [260] for a further dissection of this relationship between originator and propagator of messages.

> *"We might have just handed a 4-year-old a loaded weapon"*
>
> — Chris Wetherell, one of the engineers who created the retweet button for Twitter in 2009

The involvement of an audience in the spreading of misinformation raises at least two important questions. First, what is the role of network architecture in disseminating information? Are some platforms more conducive to propagation of disinformation than others? Second, what is the underlying psychology of people's participation in spreading disinformation? Why do people engage in participatory propaganda? We take up those two questions in turn.

***Network structure and dissemination of disinformation***. Social networks share some fundamental structural characteristics. Most networks involve hubs (i.e. extremely well connected individuals) [261] and many have modular structures involving densely connected sub-groups [262]. In addition, the different connectivity features of online platforms can critically alter network structure [263]. One basic example is how the type of connections on a platform create either directed or undirected networks. Twitter or Instagram, for example, offer directed "follower" connections [264], which can be created so that information only travels in one direction. Facebook, on the other hand, provides undirected "friend" connections, which require reciprocity and symmetry (i.e. "friend" requests must be accepted and both see one another's content).

***Directed networks***. Directed networks have two independent measures for each individual; namely, the number of people an individual follows and the number of followers an individual has. These two measures can take on very different values as the former is ultimately bounded by cognitive capacities [265]. It is cognitively impossible to follow more than a certain number of sources without losing any benefit from the information. By contrast, the number of one's followers is not bounded, as followers do not require attention [266]. In consequence, on such platforms, a few individuals can reach huge audiences [267, 268]. For example, Lady Gaga is a major influencer and has a Twitter follower base roughly the size of the German population. Influencers gain their reach in a collective self-organised way (as opposed to, e.g. journalists), by followers choosing someone to attach themselves to based on their preferences. This preferential-attachment process can be further amplified by platform recommendations [264]. The distribution of the counts of followers on Twitter is very broad [269]. This type of structure is well known to facilitate large/global/viral outbreaks of simple contagion processes [270]. For complex contagion processes, that is those that require social influence from multiple sources to spread [271], the role of these influential users becomes particularly critical. At first, they act as gatekeepers, but once they are convinced to share something they become tipping points that can trigger global cascades [272].

Notably, such cascades can even be triggered by less connected users. A recent example of this phenomenon occurred in March 2020, when a relatively obscure user on Twitter tweeted at a more connected user about the benefits of the pharmaceutical chloroquine for treating COVID-19; this second user then tweeted about the drug, which in turn caught the attention of Elon Musk, the CEO of Tesla and a major Twitter influencer, who also began tweeting about the drug. Within days of Musk tweeting about it, influential Fox News anchor Tucker Carlson brought one of the links in this tweet chain onto his prime-time show.[54] Later

[54] https://www.huffingtonpost.co.uk/entry/chloroquine-coronavirus-rigano-todaro-tucker-carlson_n_5e74da41c5b6eab77946c3b3?ri18n=true&guccounter=1

that day, American President Donald Trump announced in a press conference that chloroquine and hydroxychloroquine, which is derived from it, were showing very promising results. The path from a relatively unknown user to a nationwide presidential press conference took three tweets, a television program and just under eight days [273].

Clearly, influencers on Twitter have great potential to spread misinformation and disinformation. From the point of view of malicious agents, they are ideal targets for persuasion. Even without being the source, once they are convinced of a piece of false information they can greatly accelerate its spread. To illustrate, influencers have been found to be responsible for 69% of the engagement on COVID-19 misinformation, while rarely being the source themselves [274].

***Undirected networks***. Undirected connections, by contrast, require reciprocity for a connection to materialize, which naturally limits the number of connections an individual has (e.g. Facebook limits users to no more than 5,000 friends and most users have far fewer [275]). In consequence, there are fewer highly connected users and influencers. However, many of these connections are personal, between users who have met in real life or share interests or ideologies [276]. Stronger social ties and greater trust may lead to more influence on peer behaviour [277] and thus propagandists can spread disinformation through these more personal links [278]. Such networks are particularly vulnerable to rumours that require social conviction to spread, like social movements, extreme opinions or conspiracy theories [279].

Disinformation campaigns have been exploiting this attribute of Facebook. Russian government-sponsored activity on Facebook over the past five years has tended to disseminate information via Facebook "Pages", which users can choose to follow and which enable directed information transfer from a centralised source [280] (in that sense, Facebook Pages are structured more like Twitter than other Facebook products). The strategy is often to create affinity-based pages that attract members with certain interests or affiliations, such as political views or even love of pets and then push information out within the trusted network.

But undirected connections also support the formation of networks that are segregated into multiple modules of like-minded individuals [276, 281]. In those networks, shared content stays more localised and ideologically aligned [186]. Such networks, like Facebook or WhatsApp, limit the reach of individuals, but can nonetheless create ideological communities that can have other drawbacks. For example, segregated groups in modular networks can also be a breeding ground for radicalisation and polarisation [282, 283]. One counter-intuitive aspect of undirected networks such as Facebook is that despite the limited number of connections from any given user, it exhibits the features of a "small world" network [284, 275, 285, 286]. This means that the network has high clustering — i.e. it contains many tightly connected sub-networks or "cliques" — and a short average path-length (number of links between people), which in 2016 was measured to be just 4.5.[55] Any user on Facebook, anywhere in the world, is thus only a small number of "degrees of separation" removed from any other user. In networks of this type, viral information can propagate very quickly within cliques and it can spread broadly within the larger network despite the relatively few connections between users. As a result, it is more difficult to intervene on such networks to promote fact-checking or content moderation, because there are no single users with grossly disproportionate influence, as on Twitter. This makes it even more important to understand why individuals share misinformation, thereby engaging in participatory propaganda.

[55] https://medium.com/@duncanjwatts/how-small-is-the-world-really-736fa21808ba

***Participatory propaganda***. A particularly striking demonstration of participatory propaganda was provided in a study that presented participants with two side-by-side photographs of the inaugurations of Barack Obama in 2009 and Donald Trump in 2017 [287]. In the condition of interest here, the photographs were unlabelled and participants were asked to choose the photo that had more people in it. There is no doubt that far more people attended Obama's inauguration than Trump's.

The study found that among non-voters and Clinton voters, only 3% and 2% of respondents, respectively, chose the incorrect picture (i.e., the picture from Trump's inauguration with fewer people). Among Trump voters, this proportion was 15%. When the data were broken down further by level of education of respondents, the error rate rose to 26% among highly-educated Trump voters, compared to 1% for highly-educated Clinton voters. For participants with low education, the gap between Trump (11%) and Clinton (2%) voters was considerably smaller [287]. Given that inauguration attendance had become a matter of controversy at the time the study was conducted, with Trump's press secretary claiming that it was "the largest audience ever to witness an inauguration — period — both in person and around the globe," the results identified an instance in which people's partisan identity was more important than clear and unambiguous perceptual evidence [287]. Highly educated participants chose to participate in propaganda on behalf of their leader.

This is no isolated occurrence within a laboratory experiment. An NBC poll conducted in April 2018 revealed that 76% of Republicans thought that President Trump tells the truth "all or most of the time."[56] By contrast, only 5% of Democrats held that view. Essentially the same pattern was obtained by a Quinnipiac University poll in November 2018.[57] Clearly, partisanship is a major determinant of people's views of truthfulness, what counts as facts and even people's own perceptions of photographs.

What, then, makes people susceptible to misinformation and shock-and-chaos propaganda? Why do people participate in spreading misinformation and why do they consider a politician honest who, by any fact-checker's account, is prone to making statements that are easily shown to be false?

**Prevalence of receptivity**. Although much research attention has focused on Donald Trump, neither populism nor misinformation are limited to the United States. In Europe, populism and the ontology of truth it entails, have also made inroads, albeit to different extents in different countries. For example, a recent GLOBSEC survey of public attitudes towards democracy in 10 Central and Eastern European countries found that in half of the countries, a majority of respondents would choose an autocratic leader over liberal democracy.[58]

A recent study probed the extent to which populist party supporters had or lacked political knowledge, and whether that knowledge was based on correct or incorrect information, in nine European democracies [288]. One uniform finding was that increased political information related to support for both populist and non-populist parties. In addition, higher levels of misinformation were associated with greater support for right-wing populist parties [288]. Being incorrectly informed, but believing oneself to be correctly informed, made it more likely for a voter to prefer right-wing populist parties over other alternatives [288]. This result meshes well with other findings that over-claiming of knowledge predicts anti-establishment voting [289].

[56] https://www.nbcnews.com/politics/politics-news/poll-republicans-who-think-trump-untruthful-still-approve-him-n870521

[57] https://bit.ly/2FoZhqd

[58] https://www.globsec.org/publications/voices-of-central-and-eastern-europe/

Another recent study explored seven presumed determinants of a country's susceptibility to misinformation online [290]. Table 1 shows the variables and their effects on the perceived prevalence of disinformation in 18 Western democracies. The effects in the last column were obtained by regressing *self-reported* encounters of disinformation, as reported by the Reuters Digital News Report 2018 [291], on indicators for each of the predictors.[59]

The analysis in Table 1 identified three significant predictor variables: Countries with a greater trust in mainstream media reported reduced encounters with disinformation; whereas countries with greater social-media use and larger online advertising markets were more susceptible to misinformation [290]. A cluster analysis using all predictors identified three distinct clusters of countries. The first cluster consisted of Northern and Western European countries (Austria, Belgium, Denmark, Finland, Germany, Ireland, the Netherlands, Norway, Sweden, Switzerland and the UK) plus Canada, all countries characterized by a relatively high resilience to misinformation. The second cluster comprised Greece, Italy, Portugal and Spain; all countries with a polarised media system and historically late democratisation. The final cluster only included the United States, an outlier on many of the variables in Table 1 and the country most susceptible to disinformation [290].

| **Predictor variable** | **Effect size [a]** | |
|---|---|---|
| Populism | 0.10 | |
| Polarisation | −0.30 | |
| Trust in mainstream media | −0.56 | *** |
| Shared mainstream media | 0.32 | |
| Public broadcast media | −0.16 | |
| Social media use | 0.62 | ** |
| Ad market size | 0.41 | * |

*Table 1* - Potential determinants of a country's susceptibility to misinformation online

[a] Standardized coefficient and significance (*$p<.05$; **$p<.01$; ***$p<.001$) for variables in the model reported by [290] predicting exposure to disinformation online.

Although the data in Table 1 must be treated with caution because they are based on only 18 countries and on self-reported susceptibility to disinformation rather than more objective indicators, they provide a useful pointer to future research. At the very least, they establish the feasibility and necessity for cross-cultural comparisons of the resilience to misinformation.

One variable that has only recently attracted attention is language itself. Whereas most research on misinformation on social media has been conducted in English (i.e. involving English-speaking content), most social media activity in Europe takes place in languages other than English. A recent report that audited Facebook's efforts regarding COVID-19 misinformation found within a sample of misinformation

[59] The original report of the results [290] used a different format because several of the indicators were reverse coded for theoretical reasons [290]. In the present context this is unnecessary.

content in six languages, that while Facebook had provided warning labels for 71% of disinformation content in English and 86% in French, this fraction was considerably lower in Spanish (30%) and Italian (32%).[60] Overall, over half (51%) of non-English misinformation content failed to be accompanied by a warning label. The striking differences between languages point to the need for a more consistent approach by social media platforms.

Susceptibility to misinformation also varies within countries. Three variables in particular have attracted research attention: political beliefs, age and epistemological beliefs.

**Politically asymmetric susceptibility**. There is a large body of research into the cognitive and psychological differences between liberals and conservatives. There is little doubt that some of those differences are quite striking; a recent review of those asymmetries was based on data from more than 450,000 participants from a multitude of countries [292]. The study conducted a meta-analysis of surveys from 12 different countries (United States, England, New Zealand, Australia, Poland, Sweden, Germany, Scotland, Israel, Italy, Canada and South Africa) and found a relatively uniform set of psychological predictors of conservatism that transcended countries [293].

Given the depth of polarisation and cross-party animosity in many western societies, which can be stronger than affective polarisation based on race [294, 295], it would be peculiar indeed if there were *no* psychological differences between opposing partisans. Moreover, given the earlier analysis of the populist ontology of truth, it is unsurprising that recent research has repeatedly shown susceptibility to misinformation to be asymmetrically greater on the populist right and among strong conservatives than the left [211, 210, 218, 219, 296].

Turning to controlled observation, one class of stimuli that has been used in a number of recent studies involved "bullshit;" that is, utterances designed to impress but generated without any concern for the truth [297]. In one study, stimuli were sentences that were randomly generated from a set of buzzwords (e.g. "consciousness is the growth of coherence, and of us"). When participants rated these statements for how profound they appeared, those ratings were found to be modestly but significantly associated with a proxy of libertarianism-conservatism, namely people's endorsement of free-market economics [298]. Another study found that endorsement of pseudo-profound bullshit was associated with general conservatism and support for the Republican candidates for US president at the time [299]. No such association existed for mundane statements (e.g. "a wet person does not fear the rain") [299].

This line of research was recently extended to statements about urban myths relating to potential hazards (e.g. "kale contains thallium, a toxic heavy metal, that the plant absorbs from soil") that participants had to rate for their truth value [300]. Unlike bullshit, these statements have a clear discernible meaning. It was found that participants who were more conservative exhibited greater credulity for false information about hazards [300]. That is, conservatives were more likely to believe that kale contains thallium than liberals (there is no good evidence that it does).

This correlation was absent for similar statements that underscored putative benefits (e.g."eating carrots results in significantly improved vision"), which is consonant with a large body of research that has associated a greater negativity bias — i.e. a greater physiological response and allocation of more psychological resources to negative stimuli — with conservatism across numerous countries [301].

[60] https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/

***The role of cognitive style and epistemic beliefs.*** It is becoming increasingly clear that not everyone is equally susceptible to misinformation. A number of individual-difference variables have been identified that either increase or decrease susceptibility to misinformation, fake news or conspiracy theorising. Table 2 summarises those variables.

The overall pattern in Table 2 is unsurprising. The fact that people who have a consistent desire to base their opinions on evidence are less susceptible to hold beliefs that are unsupported by evidence [305] is unsurprising. Of further interest is the nuanced role of political views. As shown in the previous section, strong conservatives and right-wing populists are overall more susceptible to being misinformed. Similarly, there is an overall correlation between the variables in the table and political views. Nonetheless, one study found only a limited number of interactions between epistemic variables (e.g. need for evidence) and political views, suggesting that cognitive style outweighed worldviews [305].

| **Increase susceptibility** | **Decrease susceptibility** |
|---|---|
| Endorse delusion-like ideas (e.g. telepathy) [302] | Actively open-minded thinking [302, 303] |
| Dogmatism [302, 303] | Analytic thinking [302, 304] |
| Religious fundamentalism [302] | Need for evidence [305] |
| Strong distrust of the social and political system [306] [a] | |
| Low trust in media [219] | |
| View reality as a political construct [305] | |
| Age [211, 210, 218] | |

*Table 2* - Individual-differences variables that increase and decrease susceptibility to misinformation and fake news

[a] [306] examined trust in fact checkers, not fake news consumption per se.

***The effects of age***. A consistent finding in the literature is that older Americans are more likely to consume and share fake news [211, 210, 218]. This finding is particularly troubling in light of the fact that older Americans are more likely to vote than any other age group [307]. In Europe, the effects of age are far less clear because little research exists that has focused on this issue. In one study that examined the effectiveness of an intervention to reduce reliance on misinformation, pre-intervention baseline scores did not differ between younger and older adults in three countries (Germany, Greece and Poland) [308].

Although the role of age in belief and sharing of misinformation in Europe is presently unclear, given the prominence of age effects in other societies, possible reasons deserve to be explored. One possibility is that older people are less skilled with modern technology generally, having acquired those skills later in life because the technology became available only recently. On this view, the age effect should gradually diminish as people who acquired their online skills earlier in life are aging. Another possibility is that increasing susceptibility to misinformation with increasing age represents just another manifestation of a general cognitive decline that is often observed later in life. In the context of misinformation, older people are known to forget corrections more than younger people [309]. An obvious implication of this view is that the age effect will persist over time because it is not tied to a specific cohort. A further possibility is that older people share fake news not because they are failing to recognise it as false, but they have other reasons for sharing (e.g. to amuse or provoke friends and family [229]). Older people are known to use technology mainly to connect with others rather than to acquire new information and this social use of technology is related to their psychological well-being [310]. A recent review of the effects of aging on processing of misinformation, concluded that “cognitive declines alone cannot explain older adults’ engagement with fake news. Interventions in a ‘post-truth world’ must also consider their shifting social goals and gaps in their digital literacy” [307, p. 14].

***Countering misinformation: Fact-checking***. Independent fact-checking organisations have become an integral part of democratic discourse in many countries. Quality media have embraced fact-checking and in 2008, the fact checker Politifact won the Pullitzer prize.[61] Although there is some debate about the epistemological justification for fact-checking [311, 312, 313], a recent meta-analysis has confirmed its general effectiveness [314]. Specifically, across 30 experimental studies, fact-checking shifted beliefs in the expected direction in 28 cases. However, not unexpectedly, the effectiveness of fact-checking interacted with some of the variables considered above, such as pre-existing beliefs and partisanship. A recent large study of public perceptions of fact checkers in 6 European countries found greater acceptance of fact-checking in Northern Europe (Sweden, Germany) than elsewhere (Italy, Spain, France, Poland) [315]. The study also found that those with negative feelings towards the EU were less likely to embrace fact-checkers, raising the possibility that those most vulnerable to disinformation are also the hardest to effectively reach by fact-checkers [315].

Moreover, however effective fact-checking may be in controlled experiments, in the real world it can only be effective to the extent that people are exposed to fact checks. There is evidence that fact-checked validated information does not travel as far or as wide as misinformation [253].

In a democracy, accountability of politicians is critical. Fact-checking is one necessary element of providing this accountability. However, fact-checking alone will not resolve the “post-truth” crisis.

[61] https://www.politifact.com/article/2009/apr/20/politifact-wins-pulitzer/

## Key scientific findings

- Two core attributes from the attention economy and human psychology create the perfect conditions for the spread of misinformation: algorithms that promote attractive, engaging content and people's strong predisposition to orient towards negative news.
- Fake news generally makes up a small fraction of the average person's "media diet", but some demographics are disproportionately susceptible. Strong conservatism, right-wing populism and advanced age are predictors of increased engagement with misleading content. The problem of misleading online content extends far beyond strict fake news and when misleading content is considered in its entirety, the problem is extensive and concerning.
- The interpretation and status of misleading content often turns on subtle issues of intent and context that are difficult for third parties—especially algorithms—to ascertain, making it difficult to distinguish legitimate political speech from illegitimate content.
- There are asymmetries in how false or misleading content and genuine content spread online, with misleading content arguably spreading faster and further than true information. Some of this asymmetry is driven by emotional content and differing levels of novelty.
- Susceptibility to misinformation varies between people, with age and some cognitive attributes leading to greater vulnerability.
- The spread of misinformation is shaped by the network structures of social media. Some network structures can give rise to significant distortions in perceived social signals that in turn can affect entrenchment of attitudes.

# Chapter 7

## Taking democracy online

Specific characteristics

How this affects our behaviour

# Chapter 7: Taking democracy online

Early hopes that the web would automatically deliver democracy did not survive contact with reality. However, this failure need not be permanent. There is hope that democracy will reclaim online spaces and successful precedents are beginning to emerge at many levels of public engagement.

Political engagement — online and offline — can be categorised depending on how directly it can influence political decision-makers and affect political decisions.

At one end of the spectrum, private or public communications among citizens have indirect effects on political decisions by shaping the political public sphere in which decisions are taken. Those activities include seeking information online, participating in political debates, discussing solutions to societal challenges and commenting on policy-initiatives. When those discussions take place online, they often involve self-governed community spaces (e.g. Reddit) that sit outside the sphere of influence of social media giants. A general attribute of those conversations is that they are not officially regulated, recognised or managed. When unregulated political conversations scale up in size, they can become social movements that make use of online platforms to shape and sometimes transform, public debate (e.g. Arab Spring, Occupy Wall Street, Black Lives Matter, Yellow Vests, Fridays for Future). These social movements typically also have a large offline presence.

At the other end of the spectrum, direct public participation is officially recognised, managed and built into the political architecture, for example through acts of direct democracy (e.g. referenda) or deliberative democracy (e.g. deliberative assemblies). Although deliberative democracy has mainly been exercised offline, there have been recent attempts to design platforms that permit large-scale online deliberation.

Between these extremes, we find other forms of citizen involvement in policymaking, such as signing online petitions, joining crowd-sourced lawsuits or contributing to consultations. Successful precedents at the intermediate level are also beginning to emerge. We explore recent developments along all these levels, from self-governed community spaces to online opportunities for deliberative democracy.

## Specific characteristics

***Self-governed community spaces***. The web provides space for self-governed community fora that are not curated or controlled by the social media giants. For example, Reddit and outlets such as 4Chan (an anonymous site sympathetic to extreme-right content), cater to the interests and needs of a broad range of communities, including marginalised sections of society and niche communities. They can be utilised by different actors and serve a multitude of purposes, spanning a wide range of different design features [316].

Self-governed community spaces play a double-edged role in society. On the one hand, these spaces must be recognised as an opportunity for marginalised sections of society to build capacity for positive change. On the other hand, their potential to radicalise users must be critically examined as a potential threat to society. Both of these aspects can be illustrated by analysing Reddit. Reddit is a collection of fora ("subreddits"), most of which are user-run and moderated. There currently are around 138,000 active subreddits.[62]

[62] https://en.wikipedia.org/wiki/Reddit

Reddit's traffic consistently ranks among the top 10 US sites [317]. In the EU, Reddit ranked fifth among social media providers in 2018.[63] Reddit has a historical commitment to freedom of speech and has allowed a large number of fringe communities to flourish. It also provides opportunities for engagement between celebrities (including politicians and scientists) and the general public, through a feature called AMA ("Ask Me Anything"), which is hosted by the subreddit *r/IAmA*.[64] An AMA resembles an online press conference that is open to anyone, with questions and answers being upvoted or downvoted by the public. Notable participants in AMAs include Barack Obama, Bernie Sanders, Bill Gates and Donald Trump. A recent analysis of the experiences of 70 scientists who hosted an AMA revealed that most hosts considered it to be an enjoyable and productive experience [318].

However, other parts of Reddit may be more problematic. According to one study, Reddit's "karma point system, aggregation of material across subreddits, ease of subreddit and user account creation, governance structure, and policies around offensive content serve to provide fertile ground for anti-feminist and misogynistic activism" [319]. Reddit also supports a large community dedicated to conspiracy theories (at the r/conspiracy forum [317]). A retrospective analysis of users of the r/conspiracy forum found that prior to posting in conspiracy forums, users consistently exhibited anger more often than a control group of users who did not end up posting in the conspiracy forum [320].

***Social movements online.*** Online social networks can be empowering platforms for individuals and minorities to reach unprecedented audiences for creating awareness and enable grassroots movements (e.g. "hashtag activism" #metoo, #BlackLivesMatter). For example, the #metoo hashtag, denoting an experience with sexual harassment, was used 19 million times on Twitter in the year leading up to October 2018.[65] For example, 52% of French Twitter users surveyed in 2019 mentioned #metoo as being the most striking hashtag symbolising a collective movement since 2015. It ranked second place only behind #JesuisCharlie.[66]

Online activism thus promises easy participation — or at least the illusion thereof — from a safe place, via a simple "like" or share, although its actual impact in the real world is still being debated. On the one hand, there is research showing that online activism can facilitate future action to achieve social change [321]. Social affirmation through online interactions has been identified as a causal variable of how online activism can lead to offline collective behaviour [321]. On the other hand, it has been argued that online activism *inhibits* offline political participation — this is known as the "slacktivism" hypothesis and it is also not without empirical support [322]. A recent reconciliation of those divergent results [323] identified pre-existing activism as a critical variable. That is, online activism increased future activism for other campaigns in individuals who were already active and had a sense of efficacy of their actions. For other individuals, no such benefit was observed [323]. This result identifies the crucial role of affirmation or positive feedback for people who engage in online activism, in order to enhance their sense of efficacy.

Communication and the flow of information are especially critical for social movements due to their informal structure and large number of diverse participants. In addition, communication is critical because the goal of social movements is to attract widespread attention and support [324]. Providing easy and cheap ways to act politically lowers the barriers for participation [325]. Some prominent instances of

[63] https://ec.europa.eu/info/sites/info/files/osm-final-report_en.pdf

[64] https://en.wikipedia.org/wiki/R/IAmA

[65] https://www.pewresearch.org/fact-tank/2018/10/11/how-social-media-users-have-discussed-sexual-harassment-since-metoo-went-viral/

[66] https://www.statista.com/statistics/996368/most-memorable-hashtags-collective-movement-twitter-france/

mobilisation, such as the Egyptian revolution of 2011 and the Occupy Wall Street movement beginning the same year, have very visibly formed via social media [326]. While their prominent reliance on Twitter has led to the creation of the term "Twitter revolutions" or "Hashtag activism", other social media have played an important role as well (e.g. Facebook for the Women's March on Washington[67]).

Platform design has a major impact on the characteristics of a social movement's user base, the available avenues for political behaviour, the necessary transaction costs of any activity and the formation of a collective identity [327].[68] Different platform design choices can affect not only the amount of political activity, but also its characteristics. Preliminary findings indicate a significant amount of political activity on *TikTok*, which — potentially due to its more open design and "duet" function (a function that allows users to create new content featuring an initial video, with both videos appearing side by side, thus effectively permitting users to reply to video content with their own video) — creates more interaction across partisan lines [329].

A detailed analysis of the use of Twitter by three protest movements — Occupy Wall Street in the US, Indignados in Spain and Aganaktismenoi in Greece — revealed that Twitter was predominantly used for ongoing discussion among activists, media and the public and to sustain the movement, whereas original movement mobilisation had already occurred through other online platforms or offline [330]. Overall, social movements appear to be highly adaptive users of multiple platforms, demonstrating the high level of engagement that is possible through social media.

However, mobilisation through social media does not necessarily follow a democratic, consensus-oriented or dialogue-enhancing path [331]. If these crucial ingredients for a constructive process are missing, large movements can then render themselves ineffective in real world politics [72]. Several obstacles to democratic processes arise from the cognitive consequences of platform architectures, in particular how they communicate the opinion of others to a user.

Social signals from others are one of the most powerful determinants of attitudes and behaviours. The perception of a consensus opinion among others has been identified as a causal agent in shaping and changing of attitudes, including for politically-charged issues relating to stereotypes and discrimination. Receiving information about the predominant attitudes among one's peer group—namely their views towards minority groups—tends to shift a person's attitudes in the direction of the consensus [332, 333, 334]. The effect is enhanced if the purported consensus involves members of one's in-group, it can be long-lasting and is detectable outside the context of the initial manipulation [333]. A misperception or distortion of the social signal from others' opinions can have far-reaching consequences. One form of distortion, known as "pluralistic ignorance", arises when a minority opinion is given disproportionate prominence (e.g. in public debate or by the media), in which case the actual majority of people may think that their opinion is in the minority [335, 336].

[67] https://www.facebook.com/womensmarchonwash/

[68] This of course only applies so far as movements rely on such an identity, instead of existing in a much more loose, unstable, and ad hoc manner, such as the group Anonymous [328].

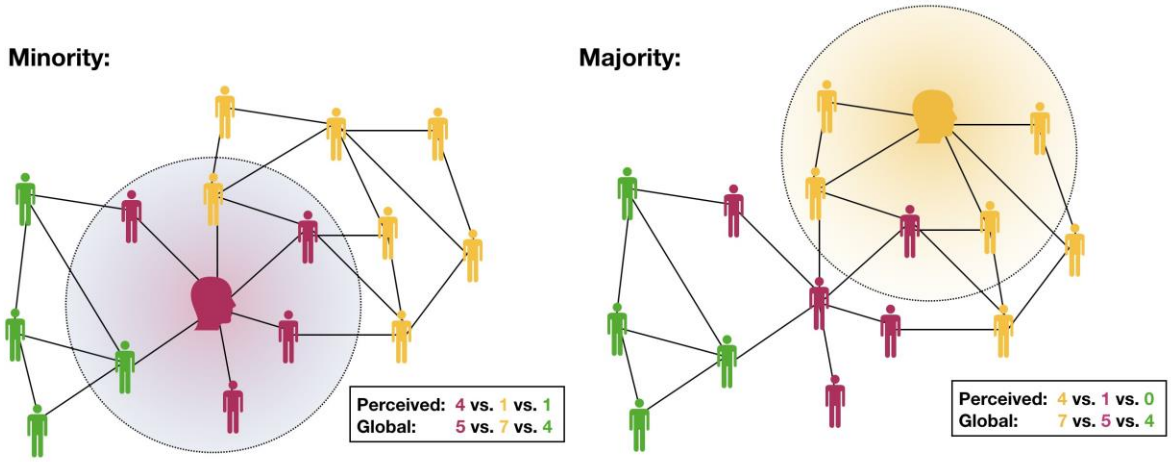


*Figure 5* - The perception bias in homophilic networks, where the colours correspond to some attribute - like political affiliation - the connections are more frequent between individuals sharing the same attribute and less if they are different. This causes individuals who belong to a minority (e.g. magenta) to overestimate their group's size and for members of the (yellow) majority to underestimate their size or even overlook them (green minority) completely.

In those circumstances, majority opinion may shift towards the minority position because its size is overestimated [337, 338]. The complementary distortion, known as "false consensus" applies when people overestimate the prevalence of their own opinion in their group or society at large (as they typically do, [339, 340]). The extent to which people over-estimate the prevalence of their opinion predicts their intention to engage in attitude-consonant behaviours: For example, the likelihood that someone might smoke marijuana increases with the extent to which the person over-estimates peer-support for the legalisation of drugs [341]. It follows that how platforms display the opinions of others to a user is a design decision with non-trivial implications.

Most platforms display metrics of social reactions and cues (e.g. the number of "likes" and emoticons [342]), which could in principle quantify the degree of consensus of a public discussion. However, these signals are asymmetrically positive — there typically is no "dislike" button—and are biased toward narrow groups of highly active users (as they do not show passive behaviours/engagement) [343, 344]. Recent results indicate how such biased metrics can even influence people to share misinformation and ignore fact checks [345]. The limitations can have further effects, such as dramatically changing a user's perception of group sizes (see Figure 5) and swaying collective decisions [346].

Another problem of existing social metrics is that they only represent a user's immediate online neighbourhood and there is low visibility of the global state of the network [347, 281]. This distortion can create the illusion of broad support [348]. Although large social media platforms routinely aggregate information that would foster a realistic assessment of societal attitudes, they currently do not provide a well-calibrated impression of the degree of public consensus [349]. This can cause social movements to exist in parallel, keeping them too small to have real-world impact, while at the same time withholding the stimulus to expand further, since one already thinks to be in the majority (e.g.[69]).

There is no compelling technological constraint that necessitates this distortion of social cues online. The interactive nature of social media could be harnessed to promote diverse democratic dialogue and foster collective intelligence. The positive examples in the past and the engaging character of social media, make them promising candidates for platforms of participation. In order to achieve this goal social media needs

[69] https://www.washingtonpost.com/news/the-intersect/wp/2014/05/08/bringbackourgirls-kony2012-and-the-complete-divisive-history-of-hashtag-activism/

to offer more meaningful, higher-dimensional cues that carry information about the broader state of the network rather than just the user's direct neighbourhood. For example, it has been shown that alternative polling algorithms that systematically sample perceptions of the global prevalence of an opinion among neighbours can effectively mitigate distorted perceptions caused by the network structure [350]. Social media platforms could provide a transparent crowd-sourced voting system [351] or display informative metrics about the behaviour and reactions of others (e.g. including passive behaviour, like the total number of people who scrolled over a post), which might put into perspective social engagement numbers and could counter pluralistic ignorance and prevent false-consensus effects.

***Public participation in actions.*** Social media does not only permit coordination of social movements, it can also direct users to "official" websites dedicated to political participation. A study on the use and design features of a UK government petitions platform illustrates the growing importance of social media, as 50% of traffic to the website came via Facebook (40%) and Twitter (10%) [352]. Once people reach a platform to express political action, platform design again plays a crucial role.

This can be illustrated with the results of a natural experiment in which the design of a UK government petition platform underwent a seemingly minor modification; namely, the introduction of a "trending" feature on the website [352]. The study found that displaying information about the behaviour of others on the petition website had a significant influence on the political choices of users, that is, which petitions they chose to sign. This altered the distribution of signatures, which shifted from the less popular petitions to the popular (trending) ones [352]. This design choice might allow for more effective mobilisation efforts (for successful petitions that have popular appeal), but could also have an adverse impact on others. The study also suggests that the placement of information (within the list of trending petitions) might have impacted the decisions of the users [352]. The fact that people tend to remember the first few and the last few items on a long list — including menu items — is well established [353, 8].

> *"When the world seems large and complex, we need to remember that great world ideals all begin in some home neighbourhood."*
>
> — Konrad Adenauer

Ease of access is a crucial determinant of participation also in areas that ordinarily have a very high threshold for political engagement. One example involves mass lawsuits, in which civil society groups recruit plaintiffs online. In recent years, the Court of Justice of the European Union had to decide a lawsuit that was brought by several thousand applicants (see, e.g. Sven A. von Storch and Others v. European Central Bank, Case T-492/12). The applicants were at least in part recruited online, through political platforms (in this case of a right-wing group) or social media. Applicants received information about the planned lawsuit, they could download, print and sign a form to appoint a mutual counsel and co-initiate the lawsuit. Legal costs were usually borne by the organisation behind the planned lawsuit to lower the threshold for participation.[70]

[70] After the more than 5,000 plaintiffs lost the case and appeal, the CJEU ordered them to pay costs; Case T-492/12 DEP, ECLI:EU:T:2016:668 and Case C-64/14 P-DEP, ECLI:EU:C:2016:846).

As part of its better regulation agenda, the European Commission provides citizens with the opportunity to comment on draft legislation through public consultation before it adopts a legislative act. Views are invited on the scope, priorities and added-value of EU action for new initiatives, as well as evaluations of existing policies and laws. The consultations take place via tailored questionnaires, which have to be submitted electronically. The public consultations are open for a response period of 12 weeks. Prerequisites and procedures are relatively easy but perhaps not that direct for an ordinary citizen [354]. The requirement is registration with an EU account (former ECAS account).

A tool devised by researchers at Cornell University, called RegulationRoom (/www.regulationroom.org/), sought to overcome those limitations by providing a moderated space for citizen engagement in rule making. The moderator provided information about a proposed rule and assisted commenters with substantiation of their views, encouraged them to consider opposing views and offer alternative solutions, while also maintaining civility [355]. RegulationRoom hosted several rulemaking discussions for the US Department of Transportation and the US Consumer Financial Protection Bureau. An analysis of one particular discussion, on a rule aimed at limiting the range of debt collection practices to protect consumers, revealed the crucial role of the human moderator [356]. The need for human moderation is a recurrent theme in the literature about online political deliberation.

To date, most of these initiatives create non-binding results and are mostly experimental. Empirical research exists with regard to the effective design of crowd-sourcing platforms serving private actors [357]. It remains unclear, however, whether the insights in a commercial context hold in cases of public crowd-sourcing, with its goals of generating legitimacy, trust and acceptance, and ultimately promoting the public good.

Soliciting comments on legislative or regulatory initiatives is only a partial step towards citizen involvement in governance. The most extensive form of involvement currently known involves “mini publics” [358] or “deliberative assemblies” constituted of ordinary citizens.

***Deliberative assemblies.*** Fears of global “democratic backsliding” are supported by the finding that in at least 45 democracies around the world, politicians and political parties have used computational propaganda tools by amassing fake followers or spreading manipulated media to garner voter support; and in 26 authoritarian states, government entities have used computational propaganda as a tool of information control to suppress public opinion and press freedom, discredit criticism and oppositional voices, and drown out political dissent[359].

This fear must be weighed against several striking counter-examples that provide a more positive outlook. These positive cases tend to involve deliberative forms of democracy. For example, The Republic of Ireland recently conducted referenda on two emotive issues — gay marriage and abortion — but the country has largely escaped demagoguery, populism and polarisation [360]. Integral to this success were two citizens’ assemblies, comprised of 99 randomly chosen voters who deliberated the issue one weekend every month for a year, which ultimately issued recommendations for the referenda.[71]

The recommendations did not take a side on the binary decision required by the referendum, but noted core principles established during deliberation and the distribution of opinions within the assembly about each principle. Deliberation was moderated throughout and was informed by numerous experts and public

[71] https://2016-2018.citizensassembly.ie/en/The-Eighth-Amendment-of-the-Constitution/The-Eighth-Amendment-of-the-Constitution.html

submissions [361]. The recommendations of the Irish Citizens' Assemblies were widely disseminated and had discernible impact on the public [362].

There is growing evidence that deliberative bodies, when properly moderated and facilitated, can ameliorate polarisation and "post-truth" discourse [363]. The involvement of citizens in deliberation has been nominated as one important countermeasure to the corrosive effects of disinformation campaigns by malicious actors [364]. Several factors contribute to the success of offline deliberative fora. The presence of moderation and input from experts have been identified as critical attributes of successful assemblies. Moreover, assemblies provide protection against internet trolls, populist demagoguery and tabloid headlines — none of which find currency within moderated deliberative assemblies.

There are, however, drawbacks to assemblies, the most obvious of which is cost in time and resources. Although the Republic of Ireland is currently conducting another assembly, on gender equality,[72] there are limits to how many issues can be deliberated at the current pace and cost. Research and practice has therefore increasingly focused on moving deliberative fora online in order to broaden citizen involvement without sacrificing the advantages offered by moderated deliberative assemblies. We briefly review existing successful precedents.

***Online deliberation platforms: successful precedents***. One of the most advanced and well-established platforms for deliberation and consultation exists in Taiwan, which uses a deliberation platform known as vTaiwan.[73] vTaiwan is an officially recognised service of the Taiwanese Government [365] that enables citizens, government officials, representatives and other stakeholders to discuss legislative proposals and generate non-binding policy solutions. As of 2018, 26 initiatives have been discussed [366]. There are two noteworthy design features that contribute to its functionality [367]. First, users are unable to reply to comments, thus preventing vitriolic exchanges. Second, the platform forms groups of users based on their opinions by statistical techniques [366], which makes argumentative dividing lines as well as space for consensus visible [367]. Taiwan has used this tool, for instance, to generate a proposal for the regulation of online alcohol sales and the introduction of gig economy services (such as Uber) [366].

In a similar way, the City Council of Barcelona created an online platform to discuss local issues, such as its strategic city plan on current policy objectives and initiatives [368]. The Council set up Decidim Barcelona, a platform on which different proposals can be discussed in threaded comments. A study of the platform revealed that comments opposing the respective proposals were particularly successful in sparking discussions [368]. The authors of the study attribute this result to the design of the platform that allowed "both conversation threading and comment alignment" [368]. Decidim (http://decidim.org) is a digital infrastructure for participatory democracy, available as open source software.

The software allows a wide range of configurations for use by different actors (from local city councils to universities, NGOs or national and supra national governments), including the blending of conventional in-person democratic events with online deliberation. A recent White Paper describes the system in full.[74]

Online deliberative procedures have also aided constitution-making. Most notable is the case of the post-financial crisis Icelandic constitution, which was crowd-sourced on the basis of online public participation

[72] https://www.irishtimes.com/news/social-affairs/gender-equality-citizen-s-assembly-moves-to-fulfil-1916-proclamation-aims-1.4151805

[73] https://info.vtaiwan.tw/

[74] Available at https://ictlogy.net/bibliography/reports/projects.php?idp=4017

[369]. Egypt also employed an online tool for public participation in the drafting process for the 2012 constitution with more than 68,000 participants and more than 650,000 votes and comments [370].

The recent COVID-19 pandemic has forced several deliberative processes to move online. For example, the UK climate assembly was initially intended to be a physical event, but was moved online because of the pandemic.[75]

Case studies on public deliberation platforms have also identified potential downsides. Among them are barriers to entry, such as accessibility issues, language and literacy requirements, cognitive demands and potentially necessary privacy limitations (particularly important vis-à-vis regimes that threaten to violate freedom-of-speech guarantees and police unwanted input) or the danger of capture. Therefore, deliberation platforms might provide unequal access and cement a certain power structure, all the while appearing to be open and inclusive.

In a recent report, the OECD analysed and summarised nearly 300 deliberative precedents around the world and catalogued good practices for offline deliberation [371].

The OECD identified the following principles for good deliberative practice:

- Clear objective and linked to a defined public problem.
- Deliberation must have influence on public decisions. Participants must be able to trust that their engagement informs subsequent action.
- The process must be fully transparent.
- Participation must be inclusive and encourage attendance by marginalised and underrepresented groups.
- The assembly must be a representative microcosm of society (e.g. use random or random-stratified selection).
- Participants must have access to relevant information (e.g. reading materials, expert testimony).
- Deliberation must involve skilled facilitation.
- Sufficient time must be available.
- The process must be run by people who are at arm's length from the commissioning public authority.
- Privacy of participants must be ensured to prevent lobbying or bribery.
- Participants must be given the opportunity to evaluate the process

These principles can be met by a number of different architectures, varying in size and purpose and duration. The OECD report recommended that deliberative fora be institutionalised to become an ongoing component of democratic governance. However, the report offers no guidance on how deliberation that complies with those principals can be taken online. Recent research has begun to identify the factors that make productive online deliberation possible.

[75] https://www.involve.org.uk/resources/blog/project-update/how-we-moved-climate-assembly-uk-online

Ensuring productive deliberation online. Moving constructive deliberation online has to address two classes of obstacles: one relating to recruitment of participants and the other to providing an environment for discussion that is inclusive, rational-critical, reciprocal and respectful [372].

> *“Only the people can change and enrich things in the institutions and transmit them to future generations.”*
>
> — Jean Monnet

Although recruitment may sound like a simple issue, in fact it is not if the deliberative forum is supposed to be representative of the population [371], which rules out self-nomination and requires targeted — but suitably random and stratified — recruitment. Experience with online experiments in Finland has revealed very low return rates, ranging between 2 and 4% [372]. Although it is unclear how these recruitment rates would compare to institutional deliberations outside an experimental context, they raise the possible need for incentives or remuneration [372].

Concerning the format of deliberation, there is good evidence that if suitable conditions for face-to-face deliberation are replicated online — that is, the discussion is moderated or facilitated by a human, is synchronous (i.e., a real-time back-and-forth) and strict rules for civil discourse are provided — then the outcome tends not to differ from offline deliberation [373, 374]. That is, polarisation can be avoided and participants’ knowledge measurably increases through deliberation.

Moreover, there may be at least two opportunities for improvement of deliberation online over its face-to-face counterpart. The first concerns temporality—that is, whether discussion is synchronous or asynchronous. Synchronous deliberation is inevitable offline as people are discussing an issue face-to-face in real time. If online deliberation is synchronous, it is therefore closest to an “ideal speech situation” resembling offline human communication. Synchronous deliberation is also more conducive to reciprocity. By contrast, asynchronous discussions, where people may take considerable time to respond to a point by another party, may allow more time for reflection. In an experimental comparison of the two modes, asynchronous online deliberation was found to enhance discussion quality compared to synchronous communication [375]. It is conceivable, therefore, that the opportunity for asynchronous communication afforded by online deliberation may give it an edge over its offline counterparts. However, the generality of this effect remains to be confirmed.

A second potential advantage of online over offline deliberation lies in the potential anonymity it affords. Online anonymity has caused considerable controversy. On the positive side, the recognised benefits of online anonymity include the elimination of hierarchical markers (e.g. gender and ethnicity) that may create a more balanced playing field for discussion [376]. On the negative side, anonymity has been frequently linked with online incivility in all its manifestations including trolling [42]. However, anonymity is typically confounded with a number of other variables, such as visibility and eye contact. When these variables are experimentally disentangled, the role of anonymity is found to be minimal [32], with eye-contact being the primary driver of online disinhibition and incivility. In an experimental examination of the effects of anonymity in the context of online deliberations, anonymity was found to have no effect [375]. Although this finding must await replication, it is reassuring that whether or not anonymity is permitted seems to be relatively unimportant.

## How this affects our behaviour

***The dark side of community fora: hate and radicalisation***. Radical and hateful content online is never far away and rarely out of reach. For example, in a German survey of more than 1,000 adolescents (age 14–19), 37% stated that they had encountered extremist content online, defined as content directed against the ideas and values of liberal democracy, for example by opposing the idea of the freedom and equality of all human beings [377].

As noted in Section 5.1, there has been much concern that recommender systems, in particular on YouTube, are guiding viewers towards increasingly radical content [68]. There is, however, another path towards radicalisation on mainstream social media platforms such as YouTube. A recent audit of radicalisation pathways on YouTube found that a large percentage of consumers of extreme "Alt-Right" content (e.g. defined as fringe views such as the creation of a white ethnostate) originally consumed less extreme content identified as belonging to the "Alt-lite" and "Intellectual DarkWeb" (IDW) [176]. The IDW is defined as an overtly respectable community that discusses controversial subjects such as race and IQ without necessarily endorsing extreme views, but which defines itself in opposition to mainstream intellectual discourse [176]. The Alt-lite community does not overtly embrace white supremacist ideology but is sympathetic to concepts associated with it, such as conspiracy theories about the "Great replacement."[76] The audit found that users in all three communities are more engaged with the content — as defined by the number of comments — than consumers of mainstream media [176]. Moreover, commenters systematically migrated from commenting on milder content to commenting on increasingly extreme content. Those comments are predominantly supportive of the content. The favourable stance of the community towards that content is also reflected in the high proportion of likes (median > 96%) as opposed to dislikes.

The radicalisation pathway identified by researchers [176] highlights the need to understand the role of fora that can serve both as gateways into extremism and as conduits that channel extremist content into the mainstream. An illustrative example of this path from fringe to mainstream involves the "pizzagate" event of 2016, which culminated in an armed individual entering a pizza parlour in Washington, D.C. and firing shots inside in search of a (non-existent) basement in which an alleged paedophile ring was thought to be operating [378]. The event originated with a tweet in October 2016 that linked presidential candidate Hillary Clinton to a paedophilia ring (without any evidence), a rumour that was taken up on Reddit and far-right sites such as 4Chan. Within days, the hashtag #pizzagate appeared on Twitter and was actively retweeted by accounts based mainly in the Czech Republic, Cyprus and Vietnam [378]. The rumour had now been linked with a specific pizza parlour and discussion on a dedicated subreddit (*r/pizzagate*) began to reveal private information about people in the pizza parlour and stores nearby (which ultimately led to it being shut down).

***The bright side of online communities: Civil conversation by design***. Online communities and movements are not only pathways to radicalisation, they also offer examples of productive civil conversation, knowledge building and advancement of progressive agendas. Importantly for the context of this report, a lot depends on the design aspect of such platforms. Wikipedia offers one example of how rules of collaborative editing can result in an impressive digital compendium of knowledge. Rules and design are equally important for successful civil conversation in online fora. One such example has originated on the Reddit forum ChangeMyView (*r/changemyview*).

[76] https://www.nytimes.com/2019/08/06/us/politics/grand-replacement-explainer.html

This subreddit later became a standalone website *ChangeAView* and is now known as *Ceasefire*; https://ceasefire.net/. The idea behind the forum is that participants present their opinions and reasoning on various topics, invite others to persuade them to change their views [379] and finally acknowledge if someone's arguments succeeded to persuade them. Topics of discussions range from climate change and gun control to religion and feminism (e.g. "Religion does more harm than good" or "Women already have equality"; https://www.reddit.com/r/changemyview/wiki/popular).

Crucially, *Ceasefire* provides several ground rules that ensure civility and productivity of the conversation and moderation is used to enforce the rules when necessary. Rules include "Don't be rude or hostile to another user", "Clearly express and explain your thinking" or "Contributions should inspire or add value to discussion", which are well defined with examples and counter-examples. Although these rules might appear self-evident, their presence and enforcement has non-trivial benefits. Recent modelling work using an agent-based approach has shown that *polarised* groups can identify optimal policy solutions with considerable skill, provided the agents are willing to talk and learn from each other [380]. By contrast, when even a small number of agents are impervious to evidence, these evidence-resistant minorities can prevent convergence on an optimal outcome [381]. Thus, disagreement or polarisation *per se* do not seem to present an insurmountable obstacle: what appears to be more important is *how* a conversation between political opponents is conducted.

The importance of a clear set of rules and moderation has also been emphasised in research on online deliberation [372]. If users are left to their own devices, online conversation readily deteriorates. Empirical evidence confirms that polarisation increases between opposing groups when like-minded people are left to their own devices [374]. By contrast, when the same issue was discussed under guidance by deliberative norms and an active facilitator, polarisation between groups was reduced [374].

Although online deliberation has shown considerable promise, at least two issues remain to be resolved: First, not all citizens have high-quality web access. The impact of the "digital divide" is becoming increasingly clear and many voices now consider web access to be a basic human right.[77] Second, there is no consensus on the distinction between online *participation* platforms and online *deliberation* platforms — the latter having been designed *per se* to be deliberative. In particular, there is no concrete agreement on what specific attributes a platform must have in order to be considered deliberative.

[77] https://webfoundation.org/2014/12/recognise-the-internet-as-a-human-right-says-sir- tim-berners-lee-as-he-launches-annual-web-index/

## Key scientific findings

- Some self-governed online fora have been identified as contributing to radicalisation and toxic extremism.
- Secluded online spaces can function as laboratories that develop extremist talking points that then find entry into the mainstream but they can also provide voices to marginalised and disadvantaged communities.
- Consumers of extremist content on self-governed fora often begin with less extreme content and shift over time.
- Current social media platform architectures provide social signals that can lead to misperceptions about relative group sizes. This has consequences for social movements who can believe that their ideas have broader or lesser penetration than they actually do.
- Government-supported platforms have successfully permitted large-scale public consultation.
- Existing research in online deliberative spaces suggests that online deliberation, when properly designed, may match the success of offline deliberative “mini publics” and citizens’ assemblies.

# Chapter 8

## What does this mean for policy?

Delineating policy parameters

Managing misinformation and tackling disinformation

Levelling the asymmetric landscape

Safeguarding the guardians

Safeguarding electoral processes

Safeguarding personalisation and customisation

Facilitating public deliberation

Enabling deeper policy reflections: Strategic foresight

# Chapter 8: What does this mean for policy?

There is a widespread sense that liberal democracy is in crisis but the reasons why are unclear. This is in part due to the complex nature of democracy that is not one single concept but comprises the three fundamental principles of equality, representation and participation.[78]

Consequently, the relationship between democracy and digital technologies is complex as their role and the importance attributed to them play out differently across these fundamental principles. There is no doubt that social media platforms spread polarising messages that can affect political behaviour offline but they also enable minority voices to be heard and can engage citizens in innovative ways in the political process. When it comes to policymaking, there is no one size fits all.

This report provides a state-of-the-science review of how online technologies influence political behaviour and decision-making. The fact that technological advances can have such profound effects on our democracies by shaping our human behaviour may be seen by some as symptomatic of the general health of democratic societies. Others may see it as an opportunity to harness our understanding of human cognition and use this as a force for good. Regardless, it should be clearly understood that the many evidence-based suggestions and implications outlined in this chapter will only serve policymakers well if they are undertaken in conjunction with broader efforts to meaningfully engage with citizens to understand their different values and perspectives and re-establish trust in political institutions.

Within the EU, a range of regulatory and non-regulatory instruments or combinations of instruments can be used to reach policy objectives. Firstly, action at the EU level is governed by the proportionality principle, which means that action should not go beyond what is necessary to achieve the objective. In short, the scope of the policy intervention needs to match the size and nature of the identified problem at the EU level. Secondly, the choice of policy instrument must take into account the experience obtained from the evaluation of the existing, relevant policies that are in place.

As the European Commission considers the types of regulatory steps to take, algorithms and design choices are in the meantime controlling the online environment. Made by corporations in pursuit of financial profits, these algorithms have little transparency or public oversight [72]. In parallel, however, a number of existing legislative measures support EU citizens in the online world, these include:

- Consumer protection rules which resulted in a large quantity of COVID-19 misinformation being removed from Facebook;[79]
- EU e-Commerce Directive that establishes content removal liability clocks;[80]
- The EU's Audiovisual Media Services Directive that regulates amongst others, content on YouTube;[81]
- The right to protect one's reputation, e.g. from slanderous fake news/disinformation is enshrined in the Charter of Fundamental Rights of the European Union and implemented through defamation law;[82] and
- It is criminal law that can convict fraudulent creators of disinformation.

[78] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12007L/TXT
[79] https://eur-lex.europa.eu/eli/reg/2017/2394/oj
[80] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX\%3A32000L0031
[81] https://eur-lex.europa.eu/eli/dir/2018/1808/oj
[82] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

While respecting the legislative measures that are already in place and the key scientific findings in this report, what more could and should policymakers be doing?

## Delineating policy  parameters

Due to the integrated nature of the pressure points identified in this report, it is not meaningful to recommend individual policy actions. Instead, the three fundamental democratic principles of equality, representation and participation are used as a framework to shape the proposals formulated in this chapter.

### Equality

- Managing misinformation and tackling disinformation;
- Levelling the asymmetric landscape;
- Safeguarding the guardians;

### Representation

- Safeguarding electoral processes;
- Safeguarding personalisation and customisation;

### Participation

- Facilitating public deliberation;
- The role of technology in the European Commission's forthcoming Conference on the Future of Europe initiative; and
- A final section "Enabling further policy reflections" is dedicated to the practical use of strategic foresight as a means of supporting complex policy reflections.

## Managing misinformation and tackling disinformation

In line with what they consider proportionate, policymakers have at least four classes of interventions at their disposal. They can regulate content directly, they can mandate that platforms regulate their content to limit misinformation, they can mandate the redesign of platforms to develop architectures that are more conducive to the spread of high-quality information than misinformation or they can request the development of tools that can ultimately empower people to become more resistant to misinformation. We explore the four classes in turn.

***Regulating content***. At the time of writing, the COVID-19 crisis has led to a vast expansion of government power in many countries, including some attempts to legislate against "disinformation" relating to the virus. Legislation is a powerful tool, however there is little evidence that on its own, it is as effective a tool against disinformation as the digital ecosystem requires. Conversely, there are legitimate concerns that overly broad legislative measures against disinformation will open the door to censorship.

Moreover, regulatory frameworks designed to combat disinformation (wilfully misleading content produced by malicious actors) must contend with the fact that other social media users may believe or share the same content without malicious intent. The fluidity between different forms of false and misleading content, as discussed in Section 6.1, often rests on intent and context, making it difficult for legislation and regulation to distinguish true disinformation from legitimate speech or unwitting sharing of false information.

There are also practical difficulties associated with regulating content. A recent example of these difficulties involves a video that claimed that the COVID-19 pandemic was the result of a conspiracy. The video was based on disinformation and flawed reasoning [382] and was therefore removed by YouTube and other platforms. This did not stop the video from appearing elsewhere on the web, often under the banner of preserving "free speech" [383], which was then re-shared on Twitter, Facebook and other social media sites after the original content was removed from these larger platforms.

Although regulation or legislation of content is fraught with risk, this does not mean that regulation cannot play a constructive role in other ways; for example, by mandating the reshaping of online architectures to ensure principles of fairness and transparency are adhered to.

***Mandating Content Regulation***. Regulators could insist that platforms suppress misinformation and disinformation by their own means. For instance, legislation could require platforms to maintain research groups aimed at finding and eliminating new forms and sources of misinformation as they develop. Such groups should be especially responsive to the worry that, due to its arms-race character, misinformation is constantly evolving and thus requires flexible, evolving techniques to combat it. Regulators could also insist on the involvement of independent fact-checkers, which are known to be at least partially effective in countering misinformation [314]. Moreover, the performance of platforms in removing misinformation must be subject to constant public audit on human rights compliance with enforceable consequences. At the time of this writing, Facebook had just failed an external civil rights audit, which found that the platform had "not done enough to protect users from discrimination, falsehoods and incitement to violence."[83]

Major platforms such as Twitter and Facebook already have internal groups designed to detect and limit misinformation. Government involvement might take the form of requiring that the amount of misinformation on platforms remain below some threshold or that robust efforts are made to prevent pollution of the information space by non-authentic or non-human actors such as "bots" or "sock-puppets".

There is some precedent for this approach. For instance, the German Network Enforcement Act (Act to Improve Enforcement of the Law in Social Networks of September 1, 2017) introduced reporting obligations for social networks with more than two million registered users in Germany about their management of complaints against certain kinds of unlawful content. Moreover, the law requires the platforms to set up effective and transparent procedures to address complaints against illegal content. These types of rules of course are based on other rules regulating content as the basis of these complaints (for instance, criminal laws prohibiting hate speech or slander). The effectiveness of these regulatory efforts depends on the design decisions platforms make when implementing regulatory obligations. A comparison of how Twitter and Facebook changed their management of user complaints in response to the legal requirements introduced by the NetzDG shows that

[83] https://www.nbcnews.com/tech/tech-news/weaponized-facebook-fails-protect-civil-rights-audit-says-n1233143

"Facebook makes the process of submitting a NetzDG complaint unnecessarily cumbersome while simultaneously trying to redirect user attention away from the NetzDG reporting process" [384].

Enforcing targets for successful self-regulation could also address problems like the one identified above, concerning content that is removed but then re-introduced on other sites. In such cases, the users who are reposting or rehosting previously-removed content originally found that content on the social media sites that removed it, but no mechanism is in place to prevent reposting elsewhere.

Technology exists to track photos and videos by creating a unique "hash" for an item, analogous to a unique DNA profile.[84]

Once a visual item has been thus identified and catalogued, any other site can detect the offending item and prevent uploading. One possibility is to create incentives for social media companies to participate in this technology to combat disinformation.

Additionally, proactive corrections could be automatically posted if a user has seen/engaged with mis/disinformation once it has been fact-checked; e.g. "You have shared XXX. It has been found that this was wrong, see here for more information."

***Redesigning platforms***. We already explored the power of choice architectures in Section 4. Design of online platforms is also important in the context of misinformation, sometimes in unexpected and unanticipated ways. A recent example relates to the misinformation-triggered "WhatsApp murders" in India (see earlier discussion in Chapter 1). By merely curtailing the number of times a message can be forwarded, a seemingly trivial change, WhatsApp may have contributed to the elimination of those lynch killings [13]. This approach, that blends platform design with knowledge of human cognition, has been labelled "technocognition" [246].

The idea behind technocognition is to redesign information architectures in a cognitively congenial way to assist in slowing the spread of disinformation. An example of a creative design intervention that embodies the technocognition spirit involved the Norwegian public broadcaster (NRK). Reader comments on news articles and blog posts are known to affect other readers' impressions and behavioural intentions [349, 385, 386]. The mere tone of blog comments — that is, whether they are civil or uncivil — has been shown to affect people's attitudes towards scientific issues they do not understand well [387]. In order to avoid those adverse consequences of commenting, the NRK trialled the requirement that readers must pass a brief comprehension quiz before posting comments.[85]

This slight increase in "friction" is intended to raise the standard of discussion by eliminating trolls or people who have not read the content. The friction also allows for a cooling-off period, thus contributing to tempering the tone of the discussion. This approach appears attractive because it can be automated and does not constitute censorship. It must be noted, however, that the Chinese government is using extreme forms of friction to censor online information [388]. By contrast, this report advocates friction in small doses and with the explicit intent to avoid censorship. Another approach that avoids censorship is to focus policymakers' efforts on empowering citizens to become more adept information consumers.

[84] https://www.iwf.org.uk/our-services/hash-list

[85] http://www.niemanlab.org/2017/03/this-site-is-taking-the-edge-off-rant-mode-by-making-readers-pass-a-quiz-before-commenting/

***Empowering citizens to reckon with disinformation***. People acquire and practice numerous and diverse competences throughout their lives, from riding a bike or reading to adopting digital technologies. The competences to navigate the digital information ecology — which is at times manipulative and hostile — can be developed by considering the cues that are already available in the online environment but have remained largely untapped to date. One can classify those cues into endogenous and exogenous cues [98]. Table 3 lists examples of endogenous or exogenous cues.

| **Endogenous cues** | **Exogenous cues** |
|---|---|
| Actors | Source/Publisher |
| Statements | Cited references |
| Plot | Language style |
| Reported events | Audience |
| Relations between actors | Social reactions |
| Circumstances of events | Sharing history |

*Table 3* - Examples of endogenous and exogenous cues of online content

Endogenous cues refer to the content itself, like the qualitative aspects of a story (e.g. who are the protagonists; what is their relationship; what is the story's plot). Modern search engines use natural language-processing tools that analyse content using endogenous cues.

These cues have considerable promise. For example, it has recently been shown that a machine-learning classifier can be trained to detect organised influence operations on social media and differentiate those operations from organic social discourse solely based on content of messages [389]. Nonetheless, content analyses continue to have substantial shortcomings: they cannot (yet) reliably distinguish between facts and opinions; they cannot detect irony, humour or sarcasm [390]; they also have difficulty differentiating between extremist content and counter-extremist messages [68]. Importantly, current endogenous cues of epistemic quality either require background knowledge of the issue in question or sophisticated machine learning techniques — this, in turn, increases the risks from a lack of transparency, information asymmetries between platforms and users, and abuse for censorship purposes. It must also be noted that disinformation is created in an adversarial environment and that development of automated detection algorithms is likely to stimulate an arms race to develop more sophisticated forms of disinformation.

By contrast, exogenous cues do not tap into the content but the context of information. For a digital newspaper article, for instance, relevant, simple and intuitively easy to understand external cues include the sources being cited; when the article was first posted; how often it was posted and by whom; how often it was promoted (and, importantly, who paid for that); and how many people saw, shared and liked it. Slightly more complex cues could include, for instance, a sharing cascade for a social media post (see Figure 3a in [98]). Such a cascade reveals the informative history of the content before it reached the user, including metrics such as the depth and breadth of dissemination by others. Deep and narrow cascades (including repeated sharing) indicate extreme or niche topics and breadth indicates widely discussed issues. Other exogenous cues carry higher dimensional information beyond a user's direct neighbourhood. For instance, social media platforms could provide a transparent crowd-sourced voting system or display informative metrics about less active behaviour and reactions of others. This could, for example, include informative disengaging behaviours, such as information about how many people only quickly scrolled through a post rather than spending time reading it.

Once made available by the platforms, the cues outlined above could be used in behavioural interventions designed to make people more resilient to disinformation. Alternatively, social media operators could be mandated to offer the option to users to share their data with a secure research platform (this idea is further elaborated in Chapter 9) to establish a verifiable, common understanding of user behaviour. This is in line with the Eurobarometer findings on "Attitudes towards the impact of digitalisation on daily lives" that found that almost 60% of the representative EU population sample would be willing to share some of their personal information securely to improve public services. This in comparison to, e.g. the current 7-step Facebook process[86] that includes waiting for data to be combined into a file that the user then has to personally forward to a researcher.

Should the cue data be made more broadly available, the two main classes of behavioural interventions from which policymakers could draw conclusions are nudging and boosting (Table 4).

| Type | Examples | References |
|---|---|---|
| **Nudges:** The choice architectures that alter people's behaviour in a predictable way. | Privacy-protecting default settings | GDPR, Article 25 |
| **Educative nudges:** Reminders and subtle prompts to behaviour. | Fact-check labels | Facebook, 2020<br>Twitter, 2020 |
| | Accuracy nudges | [393, 394] |
| | Prominent epistemic cues | [98] |
| **Boosts:** targeting cognitive and motivational competences.<br>Boosts can target both cognition and/or the environment. | Inoculation against disinformation | [405, 404] |
| | Lateral reading | [400] |
| | Simple rules for digital literacy | [400, 401] |
| | Fast-and-frugal trees | [408, 24] |
| | Self-nudging | [24, 398]<br>Center for Humane Technology, 2019 |
| | Friction | [406]<br>WhatsApp, 2018<br>Twitter, 2020 |

*Table 4* - Interventions to empower digital competences and design better online environments (based on [24, 98].)

[86] https://www.facebook.com/help/1701730696756992?helpref=hc_global_nav

***Nudging***. Nudging interventions start from a "deficit model" of human cognition and behaviour, where both are seen as compromised by a range of hard-to-rectify cognitive biases and failures of self-control. Therefore, it is assumed to be more promising to co-opt cognitive biases and failure than to aim to overcome them [391]. This is often achieved by making small changes in the digital choice architecture (the environment) in which decisions are being embedded. This often means that institutional and public choice architects determine what is in the best interest of citizens. (See discussion in Chapter 4.)

A recent analysis of the current state of nudging online concluded that most websites included relatively invisible forms of nudges that were aimed at audience building [392], in addition to the more overt designs of choice architectures reviewed earlier in this report.

Some nudging interventions involve "educative nudging", such as the provision of disclosures, reminders and prompts. For instance, reminding people of the concept of accuracy made users more discerning in their subsequent sharing decisions of tweets (increase in the average quality of the news sources shared [393, 394]). Twitter recently introduced a facility that queries people whether they really want to retweet an article they had not read.[87]

With regard to anti-Black racist online harassment on Twitter, being called out by a white male with a high number of followers ("Hey man, just remember that there are real people who are hurt when you harass them with that kind of language") reduced the harassers' use of slurs [395]. The problem with these kinds of interventions is that warnings, prompts and moral appeals can wear off or can have unintended and potentially problematic side-effects. For instance, attaching a warning to fake news stories has been found to increase perceived accuracy of headlines that were not accompanied by warnings — thus creating an "implied truth effect" for anything not accompanied by warnings [396].

Clearly displaying epistemic qualities, such as the number of cited references and clearly distinguishing between content types (e.g. news, advertisement and posts from friends), without making a judgement about their truthfulness could be a more promising avenue because it would not run the risk of creating an implied truth effect.

***Boosting***. Boosting is a promising class of cognitive interventions from the psychological sciences [397]. Unlike nudging, boosting does not start with a deficit model of cognition and human behaviour but with a "growth" model, assuming that people's competences, skills and self-control strategies can be systematically and lastingly fostered. The objective is to empower people to make better decisions for themselves and in accordance with their own goals and preferences; ultimately, the individuals decide if they acquire a competence and once acquired, if they exercise it. In Finland, information discernment is being taught in schools and the country has the highest media literacy index of 35 European countries,[88] pointing to the possibility that boosting can be rolled out on a large scale.

Boosting can be bridged with nudging when people learn to design their proximate environment in a way that works best for them. This process is known as "self-nudging" [398]. While nudging redesigns choice architectures to prompt a behavioural change, self-nudging empowers people to act as their

[87] https://www.niemanlab.org/2020/06/twitter-wants-to-know-if-you-read-that-article-before-you-retweet-it/

[88] https://www.weforum.org/agenda/2019/05/how-finland-is-fighting-fake-news-in-the-classroom/

own choice architects. For example, one can choose to move tempting but undesirable foods (e.g. potato chips) to places that are harder to reach and not always in sight.

One competence worth boosting is people's ability to make inferences about the reliability of information based on the social context from which it originates. The structure and details of the entire cascade of individuals who have previously shared an article on social media have been shown to serve as proxies for epistemic quality [399]. A boosting intervention could provide this information (see Figure 3a in [98]), that is, display the full history of a post, including the original source, the friends and public users who disseminated it and the timing of the process. The intervention would also provide some practice opportunity to learn how to recognise informative patterns.[89]

Another competence that empowers people to evaluate the trustworthiness of information online is the ability to read laterally [400]. Lateral reading is a skill developed by professional fact-checkers that entails looking for information on sites other than the information source itself in order to evaluate its credibility (for example, "Who is behind this website?" and "What do others say about the quality of the evidence for its claims?") rather than evaluating a website's credibility solely by using the information provided there. This competence can be boosted with decision aids such as simple rules for digital literacy [401] or so-called "fast-and-frugal" decision trees [24]. Fast-and-frugal decision trees are simple protocols for the sequence of decisions that must be taken to reach a diagnostic conclusion.

A competence of particular relevance when it comes to staving off disinformation rests on insights into what makes disinformation so alluring (for example, novelty and the element of surprise) and the ability to resist its pull. This competence can be boosted by "inoculation" techniques. Inoculation targets people's ability to recognise misleading or manipulative strategies before they encounter them face-to-face or online. Metaphorically speaking, if disinformation is an infectious disease, spreading like a virus through a social network, then inoculation can immunise people against certain manipulative strategies and strains of false and misleading information (e.g. parasitic imitations of trustworthy sources and other sinister tactics [402, 403, 404]). Making people aware of such disinformation strategies or of their own personal vulnerabilities leaves them better able to identify and resist manipulation. For instance, having people take on the role of a malicious influencer in a computer game has been shown to improve their ability to spot and resist disinformation [405].

Boosting may also involve "friction", similar to what can be introduced by technocognition. In the boosting context, friction might involve asking people to "please explain how you know that the headline is true or false" before they rate their sharing intent of a story [406]. This brief contemplative pause has been found to reduce sharing intent for false headlines but not for true headlines (even though the decrease in sharing intention was relatively small [406]). Adding this friction, however, was not as effective for repeated headlines [406] — possibly because prior exposure increases perceived accuracy of fake news [407].

The preceding review illustrates that there is a wide range of possible behavioural interventions: some interventions nudge people without the explicit intent to foster competences; others aim to explicitly boost users' relevant competences; some are embedded in the digital choice architecture, others are external tools or mental routines. It is important to highlight that behavioural interventions can only

[89] See: https://tracemap.info/home

be part of an orchestrated response to disinformation, complementing but not replacing robust regulations, platforms' systematic curation of content or external fact-checking.

## Levelling the asymmetric landscape

The attention economy is characterised by a profound asymmetry between the power of platforms and the limited power of users. Any behavioural intervention therefore runs the risk of being instrumentalised by the platforms to shift the burden of responsibility for the detection and spread of disinformation and other externalities of the attention economy such as privacy violations, from the platforms to the users.

***Europe's 406 million Facebook users and their information diet***. To illustrate the extent of the issue, in June 2020, increasingly uneasy with how Facebook was handling misinformation and hate speech, high profile multinational companies (among them Unilever and Coca Cola) committed to suspending their advertising on the platform. The response by Facebook was notable: the platform first rolled out new measures to flag problematic political posts and expand its policies around hate speech.[90] A short time later, Facebook CEO Mark Zuckerberg announced that he was unperturbed by the advertising boycott, suggesting that Unilever and Coca-Cola would be back on the platform "soon enough". [91]

A healthy, collaborative relationship with the platforms across all segments of society should be encouraged; however, this must not detract from two relevant facts arising from this recent episode. First, decisions made by large corporations (e.g. Unilever) are having a direct impact upon the information diet of the 406 million active Facebook users across Europe. Second, although Facebook initially responded to the advertising boycott, it was unfazed by the actions of two major global corporations, Unilever and Coca-Cola, with a combined worth of US$123,000,000,000.[92] At the time of writing, Facebook's share price had grown by 24.6% in 2020.[93] This raises crucial issues about corporate power, governance and democratic accountability.

***Self-regulation, co-regulation or regulation?*** A related problem is that many empowering behavioural interventions require changes to the online environment (for example, transparent sorting algorithms or clear layouts). This requires the cooperation of industry, especially because some of these measures might reduce engagement and are in conflict with the platforms' commercial interests.

For these reasons, effective regulation of the attention economy could require more than behavioural insights, interventions and incentives for self-regulation: it could require the formation of specialised, dedicated bodies—with EU oversight but empowered at Member State level—analogous to those regulating financial markets and other aspects of the traditional economy. This is particularly important as the attention economy grows and becomes ever more intertwined both with more

[90] https://www.wsj.com/articles/unilever-to-halt-u-s-ads-on-facebook-and-twitter-for-rest-of-2020-11593187230

[91] https://www.bbc.co.uk/news/technology-53262860

[92] See https://www.statista.com/statistics/326065/coca-cola-brand-value/ and https://brandfinance.com/news/press-releases/value-of-unilever-brand-portfolio-more-than-double-kraftheinz/

[93] https://wallstreetexaminer.com/2020/08/top-tech-stocks-weather-the-storm/

traditional sectors of the economy and with the fora where political discourse occurs. Such a body would need to have strong oversight power and the ability to subpoena otherwise proprietary data and algorithms from online companies participating in the attention economy. This sort of enforcement power would be analogous to the power that market regulators have to implement disclosure rules for publicly traded companies, which similarly involves ensuring the release of otherwise proprietary content in the public interest.

Platforms are currently benefiting from the opacity of newsfeed algorithms because it permits unchallenged optimisation of advertising revenue, with no oversight, possibly at the expense of public understanding and democratic discourse. However, regulatory intervention could mandate other attributes to be included in newsfeeds, such as indicators of epistemic quality as well as information about the variables driving the algorithm and their weighting (for a review, see [98]). Users should be able to understand how an algorithm was used and to what effect. In some instances, this would include transparency about what data entered the decision-making system and how those data can be contested. However, without real explainability, users will remain powerless to digital technologies and the machine-based decisions that will have an immense impact on their lives [409]. Note that this is not mandating content and it does not constitute censorship: It simply mandates that the criteria for algorithms must include transparent epistemic attributes in addition to merely attracting users' attention.

***Experimentation with and without consent***. A related question is who develops, tests and then implements potential behavioural interventions? Behavioural innovations are likely to be developed both by independent academic research but also by the platforms' research departments. To illustrate, experiments that platforms conduct without user consent can have substantial ramifications, for example when different emotions are induced by altering newsfeed content on Facebook and that intervention is shown to spread through the social network [410]. There currently is no public control over their design and execution (e.g. through the approval of external ethics committees). This lack of oversight stands in striking contrast to the strict ethical review boards that oversee research by academics and public research institutes.

Without such oversight, internal experiments remain intransparent, are undertaken without explicit consent of subjects and, most importantly, target subsets of the population, exploiting information about demographics or personality. This can be detrimental to the democratic process in particular when these experiments influence election turnout, voting behaviour, the success of social mobilisation, the outrage against certain policies, the mood of constituents, vaccination behaviour and compliance with pandemic regulations, to name just a few potential effects. Policymakers should recognise that behavioural interventions are likely to exact only small effects — but even small effects can scale up over billions of users and have the potential to subvert democratic processes.

During particularly crucial periods for democratic processes (e.g. during election periods), moratoria on such experiments might be necessary in order to avoid the manipulation of outcomes (e.g. voting behaviour). This would counter the risk of a targeted intervention by the platforms themselves, which otherwise could masquerade as seemingly neutral changes of design features. Of course, sufficient oversight would be necessary to even spot this kind of activity.

To undo some of the asymmetry in the ability to conduct large-scale online experiments it seems desirable for independent agencies and the research community to have the opportunity to investigate promising interventions on the platforms. Without access to the platforms for independent research, industry enjoys privileged knowledge of the effects of behavioural interventions and modifications of choice architectures without public accountability.

***Data privacy vs. availability***. A more robust approach is needed to increase the transparency and access to data from the major social media and information sharing platforms. It is worth remembering that activities on social media contribute to our civilisation's cultural heritage. We know about ancient civilisations because, in their excavations, archaeologists unearth pottery and pieces of glass, cultural artefacts that are treasured not because of their utilitarian value but because of what we can learn from them about the everyday lives of people. Our data on Facebook, Twitter, Instagram or TikTok could be considered the contemporary equivalent of such cultural artefacts. Given the lack of access users have to data, this analogy could be stretched to the "cultural right of return." When writing about the importance of cultural restitution, French experts Savoy and Sarr stated: "The removal of cultural property not only affects the generation from whom it is taken, it becomes inscribed throughout the long duration of societies, conditioning the flourishing of certain societies while simultaneously continuing to weaken others" [411]. Our societies should resist being denied our right of cultural return.

Specifically, regulators may have to mandate the sharing of platform data with other entities. This would have to take into account the privacy implications of "big data" reviewed in this report; it would have to involve strict ethical procedures and supervision; and it would have to guard against abuse by political operatives. Significant efforts would be needed to use anonymous and anonymised data. Those data could be used by independent research organisations, NGOs and public bodies to audit algorithms and test for biases to uphold democracy in addition to safeguarding our cultural heritage.

Additionally, platforms could be mandated to provide every user with an annual summary of when their data point was sold. To avoid receiving a meaningless document, criteria could be established to ensure it is both user-friendly and insightful.

***Steps to level the asymmetry***. The Ranking Digital Rights (RDR) Corporate Accountability Index reflects over a decade of civil society and academic research into platform accountability. The recommendations below[94] have been inspired by this Index. It is intended to help policymakers understand the information that needs to be disclosed in order to reduce the asymmetry between corporations and citizens as the policy environment evolves from self-regulation to regulatory interventions.

### *Access to Key Policy Documents*

- Companies should publish the rules (otherwise known as terms of service or community guidelines) for what user-generated content and behaviour is/is not permitted.

- Companies should publish the content rules for advertising (e.g. what kinds of products and services can/cannot be advertised, formatting, types of language used…).

- Companies should publish the targeting rules for advertising (e.g. names, addresses, gender, ethnicity, personal interests…).

[94] Adapted by the NGO New America: https://www.newamerica.org/

- Companies should provide an inventory of sensitive inferences (e.g. personality) that are possible on the basis of user data for each individual.

- Companies should notify users when the rules for user-generated content, advertising content or for ad targeting change so that they can make an informed decision about whether to continue using the platform.

### *Rules and Processes for Enforcement*

- Companies should disclose the processes and technologies (including content moderation algorithms) used to identify content or accounts that violate the rules for user-generated content, advertising content and ad targeting.

- Companies should notify users in understandable language when they make significant changes to these processes and technologies.

### *Transparency Reporting*

- Companies should regularly publish transparency reports with data about the volume and nature of actions taken to restrict content that violates the rules for user-generated content, for advertising content and for ad targeting.

- Transparency reports should be published at least twice per year.

### *Content-shaping Algorithms*

- Companies should disclose whether they use algorithmic systems to curate, recommend and/or rank the content that users can access through their platforms.

- Companies should explain how such algorithmic systems work, including what they optimise for and the variables they take into account.

- Companies should enable users to decide whether to allow these algorithms to shape their online experience and to change the variables that influence them.

- Companies must provide opportunity for independent audits and reverse engineering of algorithms to ensure that they are free of biases and support informed deliberation rather than extremism.

Policymakers may want to consider extending these requirements to so-called data brokers or data merchants [412] before considering further measures such as mandating registration and licensing actors selling sensitive data. Key to the success of any policy intervention will be the ability for the relevant authorities at the EU and Member State levels to ensure  enforcement.

## Safeguarding the guardians

The growing movement of political fact-checking plays an important role in increasing democratic accountability and improving political discourse. Fact-checkers hold amongst others, governments and politicians to account through exposing false claims and exaggerated half-truths; this role is increasingly considered a public service, benefiting both journalists and citizens. However, the task is time-consuming, intellectually demanding and laborious, requiring more advanced writing skills than ordinary journalism. There are also motivational issues associated with this work, which finds dedicated journalists and civil society actors fearing that their human efforts are simply training algorithms during the profession's transition to automation.

Another pillar of management of the online environment is content moderation. Moderation seeks to detect and eliminate unlawful or otherwise inappropriate content, such as hate speech, pornography or graphic violence. Undertaken by tens of thousands of mostly subcontracted personnel, moderators aim to meet daily quotas by evaluating a maximum amount of questionable content according to corporate appropriateness criteria. The diversity of content reflects the global nature of social media's diverse user base. Consequently, there is a significant amount of illegal and objectionable content which moderation can prevent from appearing online.

This results in moderation negatively affecting the mental health and well-being of the moderators;[95] this is often compounded by strict non-disclosure agreements that prevent moderators from talking about their work.

In light of the important role that these occupations play in upholding democratic principles, policymakers may want to consider professional certification schemes that would guarantee minimal working conditions, clear explanations of inherent risks as well as the provision of psychological training and counselling. Through a certification scheme, independent audits could be used to monitor compliance and uphold standards.

## Safeguarding electoral processes

Elections are a pinnacle of democratic political expression and engagement. Everything in this report therefore also applies to the political processes and conversations leading up to elections. Regulations and policies that safeguard political discourse online also contribute to safeguarding elections. There are, however, several distinct aspects of elections that deserve being singled out for the attention of policymakers.

***Cybersecurity***. There is evidence of interference by foreign actors in several recent elections in the US and Europe [413, 414]. This interference raises issues of cybersecurity that are beyond the scope of this review and are the subject of a recent report by the JRC.[96]

***Political advertising***. Political advertising during electoral periods is heavily regulated in the EU, for both the broadcasting and the press media sectors however; social media are largely not covered by these measures.

Given the general absence of European or national-level rules, political advertising on online platforms is constrained only by the microtargeting options offered to political advertisers

[95] https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/;
https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation; https://www.npr.org/2019/07/01/737671615/from-nightmares-to-ptsd-the-toll-on-facebook-moderators?t=1600813927984; https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

[96] https://ec.europa.eu/jrc/en/news/put-cybersecurity-at-centre-of-society

established by the commercial platforms' respective advertising policies. Table 5 summarises policies for some of the most popular platforms in the EU.

Microtargeting of online political advertising has revolutionised political campaigns through the narrow segmentation of voters based on social media data coupled with sophisticated psychological profiling techniques using differentiated and customised political messages disseminated quickly and with precision.

Despite the use of personal data for microtargeting being subject to the GDPR, and the inclusion of data protection principles and compliance requirements applicable to microtargeting in the September 2019 electoral package, there are legitimate concerns that the imbalances between online and offline rules potentially pose a threat to democratic processes.

Policymakers can consider taking a number of steps including:

- Allowing only official candidate-bought and candidate-approved messages to reduce interference from third party actors;
- Ban microtargeting for political ads. Political actors will be representing and held accountable by broad constituencies, consequently their messages should be seen by all;
- To ensure accountability, all political ads should be made publicly available and centralised by an independent authority; and
- All political ads should be subject to fact-checking.

***Political Misinformation***. The spread of false information of political relevance in the period leading up to an election or other vote also deserves regulatory attention. Several jurisdictions have already taken on this special challenge in addition to general rules about the dissemination of false information (e.g. in laws against slander) and about the integrity of electoral processes. In particular, France adopted a law on the "fight against the manipulation of information" (Law 2018-1202) in December 2018. This law, inter alia, creates mechanisms to ward off attempts of foreign states to influence the outcome of votes through false information and to curb the spread of false mass communication during the three months before an election [415]. It also imposes reporting obligations on online platforms of a certain size concerning paid politically relevant content in the period before an election.

*Table 5* - Social media platform policies on political advertising and political content

| Platform | Overall policy | Definitions | Exemptions | Source |
|---|---|---|---|---|
| **Facebook,**<br>**Instagram** | "Advertisers can run ads about social issues, elections or politics, provided the advertiser complies with all applicable laws and the authorisation process required by Facebook. Where appropriate, Facebook may restrict issue, electoral or political ads." | Any advertiser running ads about social issues, elections or politics who is located in or targeting people in designated countries must complete the authorization process required by Facebook when the advertisement:<br>Is made by, on behalf of or about a candidate for public office, a political figure, a political party, a political action committee or advocates for the outcome of an election to public office; or<br>Is about any election, referendum or ballot initiative, including "get out the vote" or election information campaigns; or<br>Is about any social issue in any place where the ad is being run; or<br>Is regulated as political advertising.<br>Advertisers running these ads must comply with all applicable laws and regulations, including but not limited to requirements involving: Disclaimer, disclosure and ad labeling; Blackout periods; Foreign interference; or Spending limits and reporting requirements. | Advertisers who want to create or edit ads about social issues, elections or politics in a European Union country will need to go through the authorization process and place "Paid for by" disclaimers on ads. This includes any person creating, modifying, publishing or pausing ads that reference political figures, political parties or elections (including "get out the vote" campaigns). Advertisers will only be able to run ads in the country in which they are authorized. Then, ads will enter the Ad Library for seven years. | https://www.facebook.com/policies/ads/restricted_content/political# |

| Platform | Overall policy | Definitions | Exemptions | Source |
|---|---|---|---|---|
| **Google,**<br>**YouTube** | "We support responsible political advertising, and expect all political ads and destinations to comply with local legal requirements, including campaign and election laws and mandated election "silence periods," for any geographic areas they target." | Political content includes ads for political organizations, political parties, political issue advocacy or fundraising, and individual candidates and politicians. In the EU, election ads include ads that feature:<br>a political party, a current elected officeholder, or candidate for the EU Parliament.<br>a political party, a current officeholder, or candidate for an elected national office within an EU member state. Examples include members of a national parliament and presidents that are directly elected;<br>a referendum question up for vote, a referendum campaign group, or a call to vote related to a national referendum or a state or provincial referendum on sovereignty. | Note that election ads do not include ads for products or services, including promotional political merchandise like t-shirts, or ads run by news organizations to promote their coverage of referendums, political parties, candidates, or current elected officeholders. | https://support.google.com/adspolicy/answer/6014595?hl=en |

| Platform | Overall policy | Definitions | Exemptions | Source |
|---|---|---|---|---|
| **Microsoft** | “To offer a safe and positive online experience for users, we cannot accept ads that contain or relate to certain content.” | This includes, but is not limited to, the content covered in the policies listed below. We reserve the right to reject or remove any ad, at our sole discretion and at any time.<br>Advertising for the following content, products and services is either disallowed or subject to specific participation policies.<br>Areas of questionable legality<br>Dating<br>Defamatory, slanderous, libelous or threatening content<br>Hate speech<br>Peer-to-peer file sharing<br>Political and religious content<br>Sensitive advertising<br>Suffering and violence<br>Tax collection<br>Unregulated user-generated content | In France, ads containing content related to debate of general interest linked to an electoral campaign are not allowed. | https://about.ads.microsoft.com/en-us/resources/policies/disallowed-content-policies |

| Platform | Overall policy | Definitions | Exemptions | Source |
|---|---|---|---|---|
| **Twitter** | Twitter globally prohibits the promotion of political content. “We have made this decision based on our belief that political message reach should be earned, not bought.” | Political content is content that references a candidate, political party, elected or appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome. Ads that contain references to political content, including appeals for votes, solicitations of financial support, and advocacy for or against any of the above-listed types of political content are prohibited under this policy.<br>Twitter also does not allow ads of any type by candidates, political parties, or elected or appointed government officials. | News publishers who meet certain criteria may run ads that reference political content, candidates, political parties, or elected or appointed government officials, but may not include advocacy for or against those topics. | https://business.twitter.com/en/help/ads-policies/ads-content-policies/political-content.html |

## Safeguarding personalisation and customisation

***Renegotiating privacy and personalisation***. One potential solution to the privacy paradox is data-protection-by-design: the integration of proactive protections into the design, development and application of data systems and technologies [416]. Article 25 of the GDPR requires a reasonable level of protection as the default; thus, the burden of making rational decisions that are costly and require effort in the moment, but are beneficial in the long-run would be lifted from users. In addition, policymakers could consider regulation that directly addresses psychological targeting, for example, by restricting its use in specific contexts such as political campaigning and establishing clear standards for algorithmic fairness that prevent discrimination.

A potential difficulty with regulation and public standards in this arena is that sensitive attributes can be masked within online systems with clever renaming or by replacing an attribute with a proxy variable. For example, algorithms may infer "ethnic affinity" instead of "race" [417] or "upcoming important changes in the life of a woman" instead of "pregnant". Such euphemisms or use of proxy variables can enable discriminatory practices, such as the microtargeting of job or housing ads with respect to race and gender [105]. Such inferences, even when they may comply with the letter of the law, pose risks of harm [34], making it possible to manipulate user behaviour based on their most sensitive information.

One countermeasure involves personal data cooperatives. These cooperatives aim to democratise the decision process of the use of sensitive data [418]. Cooperatives aggregate personal data and provide access to other platforms and services on the terms decided by the members of the cooperative, serving as a way for individuals to build a common good without the need to depend on private companies or state agencies. An example of this is MIDATA in Switzerland, a cooperative that aggregates health data donated by its members to enable technological uses that can have a positive effect on their health without compromising privacy.

An alternative, technological approach would be to limit the power of the inferences made by online platforms and social networking services. Going beyond the individual's control over data, this approach would constrain the inferences that can legally be made.

***Redesigning privacy***. The challenges for the design of effective privacy protection revolve around several issues. Generally, data protection policies are based upon notice-and-choice rules mandating a certain level of disclosure. One of the main limitations of these rules (and perhaps a partial cause of the privacy paradox) is the users' limited capacity to read and process the details of the privacy policies [127].

Users also often cannot readily understand which kind of data they transmit, to whom and for how long. Consequently, notice-and-choice rules fail to incorporate the externalities of an individual's data disclosure, such as its relevance for predicting the behaviour of other users through algorithms using big data. Given its significance, the authors suggest that these externalities could be incorporated into future policies.

Another possible intervention could be to automate decisions about privacy settings, for instance by designing a tool that takes into account the individual user's privacy preferences (based on observation or one-time in-depth elicitation). Once those have been entered, the tool could automatically apply them to all new sites.

Other interventions aim at raising awareness by visualising the meta-data that is shared [131], for example through a smart phone app that shows a photo that the new app would have access to when asking for permissions during its installation process [419].

Policy interventions could consider mandating the use of such tools during the notice-and-choice process or when eliciting privacy-relevant information. In Europe, consumer protection law imposes similar responsibilities on firms, when it requires them to make general terms and conditions (the "fine print") salient in order to include it in the contract, which for example is the case in Germany.[97]

## Facilitating public deliberation

User interfaces have the potential to promote reasoned discussion. One successful model is the Reddit ChangeMyView community (see Section 7.2), which has been found to promote high quality discussion [379]. The platform provides an explicit expectation that users will state reasons for their beliefs. In addition, content is moderated to ensure posts are of high quality and engage properly with other users. Studies of this community have yielded useful insights [379]. For example, longer posts on ChangeMyView tended to be more persuasive as users can be more explicit about their reasoning [379].

By implication, sites like Twitter that put strong constraints on length may prevent users from engaging in good debate by design. In addition, emotional language was found to be less successful in online argumentation [379]. This stands in contrast to research showing that negative emotional content is likely to be retweeted [420, 421]. It follows that algorithms that promote calmer content over highly emotionally laden content may help users engage in improved discourse.

***Platform design for public consultation***. Small changes to platforms can make a big difference [352]. This creates a challenge for policymakers and designers who are responsible for public participation sites. It is very difficult to define the normative goal of platform design: what *should* a petition site look like? Civil servants should note that the most effective way they can implement optimal deliberative platforms is through a public procurement process.

As there is no neutral platform design, it is important to recognise the substantive influence of design choices and prevent undue influence, such as outright user manipulation. Also, it is not well understood yet, how this effect varies across platforms, with their different approaches to filtering and displaying social information [422]. Due to these insights, it might be advised, at least for government platforms, to conduct empirical studies before making final choices about platform design and to be aware of the effects on behaviour when changing them.

Regardless of a platform's design, learnings from behavioural decision science can help lower the threshold to public participation and increase the quality of interactions. When engaging with the public online, in addition to providing clear explanations of the topic in question, tools can be provided to empower participants as they engage in the deliberative process. Framing the process is a crucial part of any engagement and support applications can be designed to assist with e.g.: i) explaining why decision-making is difficult; ii) how to establish common understanding through the use of relevant information; and iii) how to reach a decision when there are diverse points of

[97] § 305.2 of the German Civil Code

view. Moderation is obviously key to the success of such a deliberative process, however this can be greatly supported by optional apps that provide background information and illustrative examples in multimedia formats for reading, listening and watching.

Since platform design is never neutral, public actors should make their underlying choices salient and subject to public debate. Recognising directed network platforms like Twitter and their resulting broadcasting character is important. For example, very popular accounts (i.e. with a high number of followers) could be identified as broadcasters and could be incentivised to apply journalistic standards to their content.

This will not curb their reach, but might make malicious actors easier to spot and good-faith influencers more careful before sharing content ("freedom of speech is not freedom of reach"[98]).

But in the spirit of empowerment, such network structures also allow individuals or minorities to get their voices heard, if they pick up enough social support. Limiting the number of possible connections might work against this positive achievement of social media. However, the platforms simply promoting already popular accounts to increase engagement on their site and by that amplifying rich-get-richer dynamics does not follow a democratic but a commercial goal and could be regulated. Similarly, self-affirming groups may play into the targeting business model of the platforms, as they allow groups with very specific interests to gather in a self-organised way, which can then be addressed with specific advertisement (e.g. a local toddler's group can be targeted with baby clothes). Recommendation algorithms that promote and amplify homophily further should be viewed in the light of these incentives and the potential dangers of radicalisation and perception biases they are posing.

> *"From the very beginning, I made clear that people need to be at the very centre of all our policies. My wish is therefore that all Europeans will actively contribute to the Conference on the Future of Europe and play a leading role in setting the European Union's priorities. It is only together that we can build our Union of tomorrow."*
>
> — Ursula von der Leyen, President of the European Commission

A transparent, but randomized diversification of content that goes beyond the direct neighbourhood in the network could potentially open up the discourse and prevent actors from posting extreme content, as the chance exists that it gets carried to the outside. However, such methods can backfire, when opposing content is repulsive to the other side of a polarized discussion and could even increase polarisation [187]. Alternative sampling/polling methods that consider not the directly

[98] https://www.theguardian.com/technology/2019/nov/22/sacha-baron-cohen-facebook-propaganda

opposing side of the network, but e.g. the extended neighbourhood beyond direct connections, can mitigate perception biases and potentially false consensus effects [423]. This could easily be achieved by the social cues that are transmitted along the network connections. If they were to carry higher dimensional information to draw a more realistic picture of the state of the network, they could, in principle, foster the democratic potential of large communication platforms to reach consensus or lead an informed debate. However, they need to be designed carefully and need to be hard to game by malicious actors.

Overall, such measures would neither curb freedom of speech, nor directly regulate the platforms' business model, but would have the potential to transfer responsibility for transparency to influential individuals and open up modular structures and, by that, set boundaries within which the current ecosystem and the emerging social networks can evolve further.

All of these implications should be considered by democratic institutions looking to strengthen their democracies when embarking upon meaningful citizen engagement. This includes the European institutions as they prepare for the forthcoming Conference on the Future of Europe.

## Enabling deeper policy reflections: Strategic foresight

Previous sections in Chapter 8 have identified policy options addressing specific issues based upon the status quo. However, policymakers face systemic challenges that require addressing multiple issues simultaneously. Moreover, they may have alternative visions of the web where, for example, global platforms no longer dominate the market but a multitude of alternatives exist that compete on quality and consumer protection features. This is where strategic foresight can help.

In light of the knowledge that has been garnered from this report, with a specific focus on the understanding gained about the stability of some basic behavioural principles, the authors invite readers to take the plunge and dive into the four possible futures co-created with key stakeholders which are as follows (see annex of this report for the in-depth scenarios):

*Scenario 1*: The "Struggle for information supremacy" scenario assumes that the future European information space will be marked by high degrees of conflict and economic concentration.

*Scenario 2*: In the "Resilient disorder" scenario, the EU has fostered a competitive, dynamic and decentralised information space with strong international interdependence, but is facing continuous threats from sophisticated disinformation campaigns and cyberattacks.

*Scenario 3*: The "Global cutting edge" scenario foresees a world in which societal and geopolitical conflict have been reduced significantly, while high degrees of competition and innovation have led to the emergence of a dynamic, global information space.

*Scenario 4*: The "Harmonic divergence" scenario assumes a world in which strong regulatory differences and economic protectionism between national and regional actors have resulted in a fractured global information space.

The scenarios have been prepared to give policymakers a tool to visualise the possible implications of their decisions. This can help them understand today the robustness of frameworks and how trade-offs can be made to improve policies for tomorrow.

These scenarios can be used by:

i) specialists in a single policy area to gain a broader systemic perspective;

ii) policymakers representing different policy areas whose collective efforts are needed to regulate technology and democracy; and

iii) by all policymakers as a way of engaging in meaningful stakeholder dialogue "So you consider Scenario 2 to be implausible, please tell me why?"

There are various ways to use scenarios. Four are described here. The first and most obvious is for an individual reader to read them carefully and to apply their own imagination and knowledge to assess the scenarios critically. The synoptic table (in annex) has been designed to facilitate this kind of work. While this exercise will challenge that person's thinking and will undoubtedly contribute to enrich it, the individual nature of the exercise makes it difficult to overcome personal biases and entrenched opinions. It also requires a willingness to spend enough time studying the scenarios to get the most out of them.

A second way to use the scenarios is to do the same as above, but in a group. Ideally, a small (5-6 people) and as diverse as possible group of interested people would have to read the scenarios critically before meeting to discuss the meaning and possible consequences of the scenarios. Again, the synoptic table (in annex) can be helpful to support the analytical discussion. The confrontation of diverse perspectives in this case allows for the development of richer and more robust reflections.

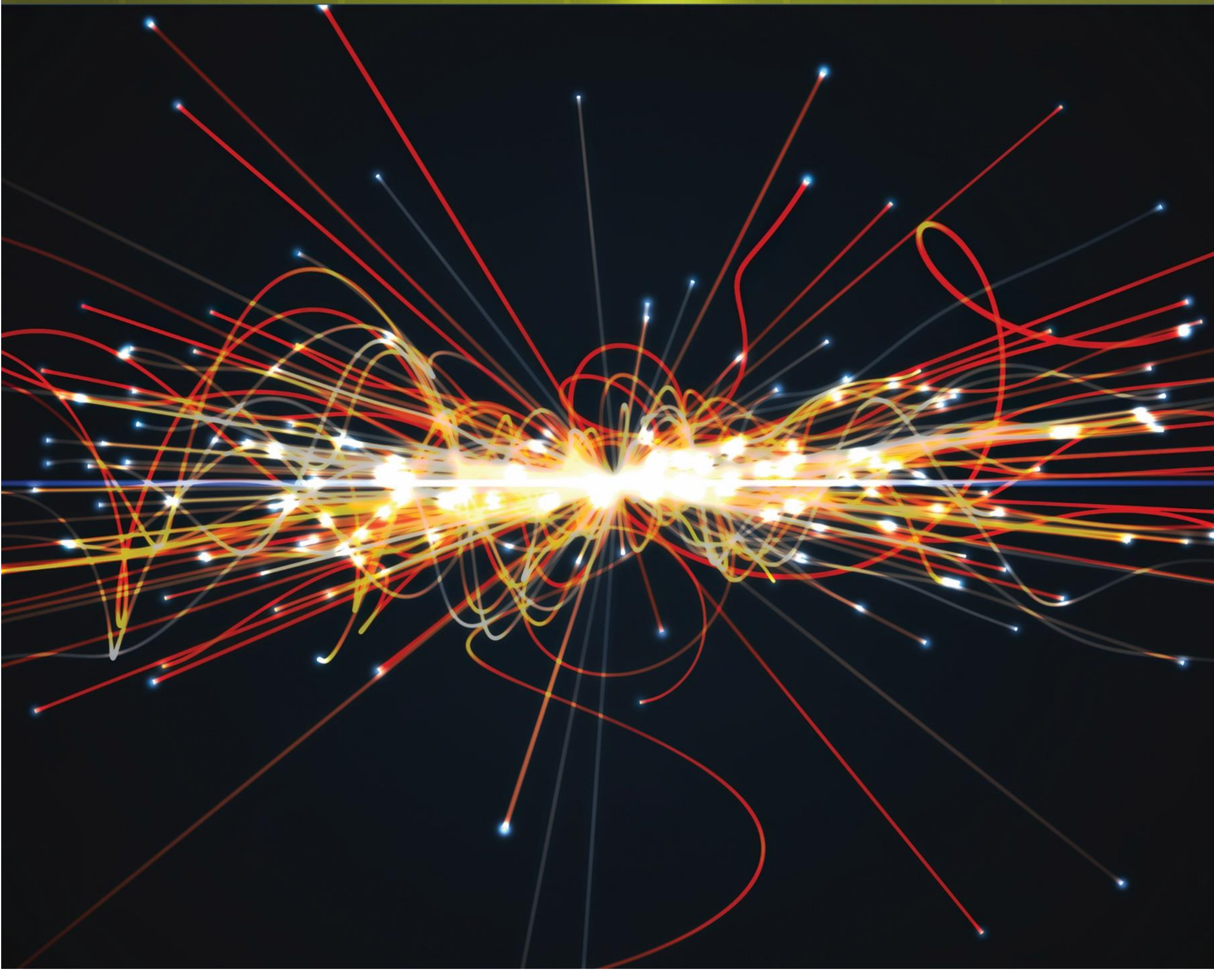
A third way, which is very relevant for policymakers, is to take a policy (or issue, e.g. a ban on microtargeting for political advertising), define clearly its objective and success criteria, and assess how the criteria would fare in each scenario. This is best done in a group. This is an applied approach and works well but tends to look at the situation from the perspective of just one stakeholder. This is a form of *ex ante* impact assessment.

A fourth way is to explore scenarios using role-playing techniques. This tends to be the richest approach in terms of understanding how the scenarios can influence the implementation of a policy but it requires a little extra preparation beyond scenario development (which in this case already exist; see annex). One tool to help do this efficiently is the Scenario Exploration System (SES),[99] a platform that engages participants in future-oriented systemic thinking developed by the JRC. The SES makes participants take action to reach their long-term objectives in contrasting scenario-related contexts while interacting with other stakeholders creating a realistic journey towards the future to simulate possible responses relevant for the issue of interest to the participants. This engagement platform helps people imagine what the scenarios of interest could mean for themselves and others and can be used for strategic development as well as anticipatory preparedness.

[99] https://ec.europa.eu/knowledge4policy/foresight/tool/scenario-exploration-system-ses_en
https://journals.sagepub.com/doi/abs/10.1177/1946756719890524

# Chapter 9

## Future research agenda



## Chapter 9: Future research agenda

To enable evidence-based policies that will shape Europe's future online information space, it is imperative that researchers have access to data from social media platforms. Current research relies upon data collected either painstakingly under uncontrolled conditions on the existing platforms,[100] or upon highly simplified experimental paradigms in the laboratory or online. None of these options are a sustainable solution to satisfy the urgent need to establish a full and transparent understanding of online behaviour, the importance of which has been outlined throughout this report.

Of all current and future human behaviours, political behaviours online are perhaps the most important ones for our collective future, but a full scientific understanding of these behaviours has been hampered by a mix of platform reticence and lack of regulatory clarity. To support the design and test of behavioural measures — they can be boosts, nudges, technocognition tools or any other measure — it is crucial to be able to carry out this research autonomously from the decision-making of existing platforms.[101] To illustrate, there is evidence that attempts to build measures by the research community can be thwarted by platforms terminating access to the required data.[102]

In response and serving multiple purposes, the EU could create a new online infrastructure for collaborative research and knowledge co-production between the public and scientists. The infrastructure would be a continuously evolving social-media environment that is designed by the public for the public in cooperation with researchers. This endeavour could generate new evidence in the service of creating an online space that fosters resilience towards the problems outlined in this report as well as anticipating future issues.

This is important as democracy is not static and should not be considered a system that can be perfected and then simply maintained and defended against threats. Democracy has always evolved and now needs to evolve further. The proposed infrastructure would enable the behavioural and scientific understanding of the "democratization of democracy" [424].

Participants would function as "citizen scientists" who — unlike current social media users — are fully informed about all aspects of the research, engaging with, employing and evaluating different tools on a consenting and consensual basis. Their behavioural data would then be used to identify the tool's efficiency, downsides, target groups and levers for improvement. Importantly, user privacy and transparency would be matched to the best legal and ethical standards of data protection, non-manipulative study design and fair compensation.

This infrastructure could also be a distributor of tools (like an "app-store" of digital assistants). Such a hub would also provide a real-time overview of the behavioural science and the success of interventions (akin to current approaches to collect behavioural science knowledge for COVID-19 responses: e.g. https://www.scibeh.org/), enabling fast responses in times of crises. A "knowledge for policy" interface would be developed, allowing big data results to become immediately available to policymakers, enabling them to make informed decisions despite the fast pace of change online.

[100] https://en.wikipedia.org/wiki/Social_Science_One
[101] https://citizensandtech.org/2020/01/industry-independent-research/
[102] e.g. https://tracemap.info

As a means of illustrating how the platform could be used, researchers and citizens could jointly pursue a research agenda dedicated to developing an understanding of the influence of technology on democracy.

An individual's ability to navigate potential traps in the online world is indispensable in the 21st century. Many of the core competences of what constitutes digital literacy have been and will increasingly be taught and practiced in schools, universities and institutions of life-long learning. Formal education, however, is slow and effortful. By contrast, the online world evolves at lightning speed. Consequently, institutionalised education in digital literacy needs to be complemented by research that designs, tests and implements behavioural measures (boosting tools) and smart online design (technocognition) that address ever-new emerging challenges and just-in-time interventions, delivered when users are most motivated to employ them or where platform redesign can be most effective.

To illustrate, "deepfakes" leverage powerful techniques from machine learning and artificial intelligence (e.g. [425]) to manipulate or produce fake visual and audio content — e.g. quite literally putting words into the mouths of politicians — with a high potential to deceive. Deepfakes are likely to dwarf the manipulative potential of fake news. The sudden emergence of such technological innovations requires a concerted research strategy with the objectives to anticipate such technological developments, analyse the challenges they pose and then to design, test, implement and evaluate boosting measures to empower citizens within a cognitively optimised design.

If the EU were to take leadership in this area, the future research agenda could be designed around three key principles:

- Building resilience through redundancy — users are highly heterogeneous (age, level of formal education, language ability, numeracy and risk literacy, level of motivation and self-control) — and no one-size-fits all behavioural measures exist. In response, a continuously growing *toolbox of boosting measures* would be developed to "fit" the specific cognitive and motivational needs of heterogeneous groups of users rather than those of a generic citizen. Redundancy by design means that such boosting interventions would be interlocked with other behavioural, techno-cognitive and regulatory measures, so that users' resilience is strengthened and their protection is retained if one of more measures fail.

- Safeguarding information autonomy — relative to the highly complex information ecology of the online world, users' bounded cognitive resources appear outmatched. Yet, a boundedly rational cognitive system [426] need neither be acting irrationally nor inefficiently and decision strategies can perform surprisingly well [427, 428]. Boosting measures are designed to respect the limits of realistic cognitive systems; in the fast-paced, complex digital environment, users will be empowered to take back some control by becoming citizen choice architects in the digital world [398].

- Competitive and efficiency testing — behavioural measures will be tested in the field with other behavioural measures and techno-cognitive interventions to gain a clear understanding of relative efficiency, including unanticipated side effects.

Implementation of these principles would be based upon the extensive theoretical and empirical body of research on bounded rationality in the behavioural sciences; analysis of expert decision-making in online environments; comparative analysis of how the cognitive system responds to offline and online technological design; simulations of structural changes to digital choice architectures to identify those that best support users' cognitive effort; and democratising boosting design by crowd-sourcing the creative process.

This may sound like a lofty vision yet, a powerful — and far more expensive — precedent for such a scientific infrastructure exists: CERN. At CERN, scientists work to uncover what the universe is made of and how it works. CERN's Member States are European (with the exception of Israel). Echoing CERN's goal, the mission of the proposed research structure for online behaviour would be to uncover the working of the digital universe and to safeguard the foundation of democratic and autonomous decision-making.

# Annex

The European information space in 2035

Scenario 1: Struggle for information supremacy

Scenario 2: Resilient disorder

Scenario 3: Global cutting edge

Scenario 4: Harmonic divergence

# A strategic foresight study of the European information space in 2035

## Four scenarios for the European information space in 2035

This report has presented the available science on the behavioural impact of online technologies on political decision-making. While many of the dynamics that have been discussed are generic and will therefore remain relevant even in light of future technological advances, the question of how our information environment will look in a few years time remains very much open. The extreme speed of development of the internet over the last two decades, as well as the acute challenges we currently encounter around issues such as disinformation and algorithmic content curation make it difficult to escape our overwhelming focus on the present. Given that the speed of innovation is not expected to subside for the foreseeable future, it is paramount for society and policymakers to anticipate the possible impacts of these developments on the European information space.

For this reason, the JRC has collaborated with a broad group of stakeholders to engage in a strategic foresight exercise to create possible futures for the "European information space". For the purpose of this exercise, the JRC defined this space as:

"All factors and entities that directly or indirectly contribute to the creation, dissemination and processing of information on the individual, group and societal level, including the translation of such information into political action."

Based on this definition, the stakeholders contributed to the construction of four possible scenarios describing what this environment could look like in the future. Balancing the need for imagining a sufficiently long time-horizon with the high degree of unpredictability of the digital environment, a fifteen-year horizon was chosen as an appropriate timeframe.

While it is impossible to predict the future, strategic foresight and explorative thinking can help us to uncover evolving trends and dynamics that may have a significant impact on the world we will eventually inhabit. For our digital information environment, such anticipation is all the more crucial, since we currently find ourselves at a number of significant crossroads. The way we will address issues such as digital privacy, hateful and discriminatory online behaviour, illegal content or online polarisation and radicalisation as a society will shape our future online experience and as proven in this report will result in offline consequences. With substantial legislation already applying to the online world and several regulatory initiatives currently taking shape at the European level, policymakers and legislators more than ever, need to look forward to identify how frameworks and policies can be made future-resilient.

***Methodology***. To best serve the needs of policymakers while respecting the time and resource constraints of this project, the JRC team applied the tried and tested foresight methodology of scenario building. Scenarios are stories illustrating possible futures or some aspects of possible futures. They are not predictions. In policymaking, scenarios help in policy design and analysis by providing realistic possible sets of future conditions. To be effective, scenarios must be:

- Plausible - they must remain within the realm of what might conceivably happen;
- Consistent - they should respect a coherent logic and not contain any inconsistency that would undermine their credibility; and
- Meaningfully insightful

Importantly, the coherence within scenarios does not mean that the developments described in them are mutually exclusive, nor are they weighted in terms of being more or less probable. In reality, individual trends from each of the scenarios might materialise in one or other form and the future will likely resemble a mixture of all four, rather than a "pure" representation of only one of four scenarios. Ultimately, scenarios should be useful.

A key strength of foresight scenarios is that they are qualitative. This gives them a potential richness of coverage that is not bound by the limitations of quantitative methods. They can easily explore relationships and trends for which either little or no numerical data are available or important dimensions that are impossible to quantify such as values, emotions, shocks and discontinuities, motivations or behaviours.

Scenario building also generates knowledge that can be applied in the formal decision-making process, helping policymakers to anticipate the context in which they have to act. They can also stimulate creativity and detach decision-makers from the priority mostly given to present and short- term problems. As such, scenario building can be a crucial method to foster longer-term leadership.

However, for scenarios to be used effectively, the participants in scenario building must be convinced of the soundness, relevance and value of the process. This is essential, as the foundations, on which scenarios are built, the structures that they use and the reasoning they employ, must stand up to critical examination. Only then is there a chance that they will contribute to decisions and actions. This is why extreme care must be taken in the setup of the group of participants in any good scenario-building workshop. The participants must cover a range of diverse relevant backgrounds, each able to provide useful insights for the topic of interest in the scenario-building process. They must also be able to bring both an inside and an outside view of the topic or system at stake. Direct participation of relevant decision-makers is also essential as it implies that they truly understand and co-own both process and the resulting scenarios, making it more likely that the insights generated by the scenarios will be used in their decision-making.

In the present case, despite the fact that the scenario building workshop took place in early March 2020, at the start of the disruptions caused by the COVID-19 pandemic, the participants associated with this foresight exercise represented the following stakeholders: digital technologies industry, non-profit organisations in the domain of digital rights, anti-disinformation organisations, think tanks, media and communications specialists, academics as well as EU policymakers (DG CNECT, DG COMM, DG HOME and DG JUST). We extend our grateful thanks to everyone who so actively contributed.

***What we actually did in practice?*** The scenario-building process for the European information space was carried out in an intense 1-day participatory workshop.

As most participants in the study did not know each other, a first session focussed on getting to know each other and creating a cohesive group. The participants were then presented with the JRC definition of the European information space and had some time to ask questions and become familiar with the concept. After that, participants were split into groups and asked to identify all the drivers of change that would affect the European information space over the next 15 years that they could think of.

The JRC team used the STEEP (Society, Technology, Environment, Economy, Policy) framework to help the participants in this exercise. All the identified drivers of change were then collected, discussed, clarified and posted in full view of everyone. The participants were then asked to look at the complete list and voted on which drivers of change would have the most influence on the future evolution of the European information space.

This yielded 11 drivers of change that received two votes or more. Participants were then asked to look at these drivers and to vote again, this time to identify those for which there was the most uncertainty regarding the way they would evolve in the future.

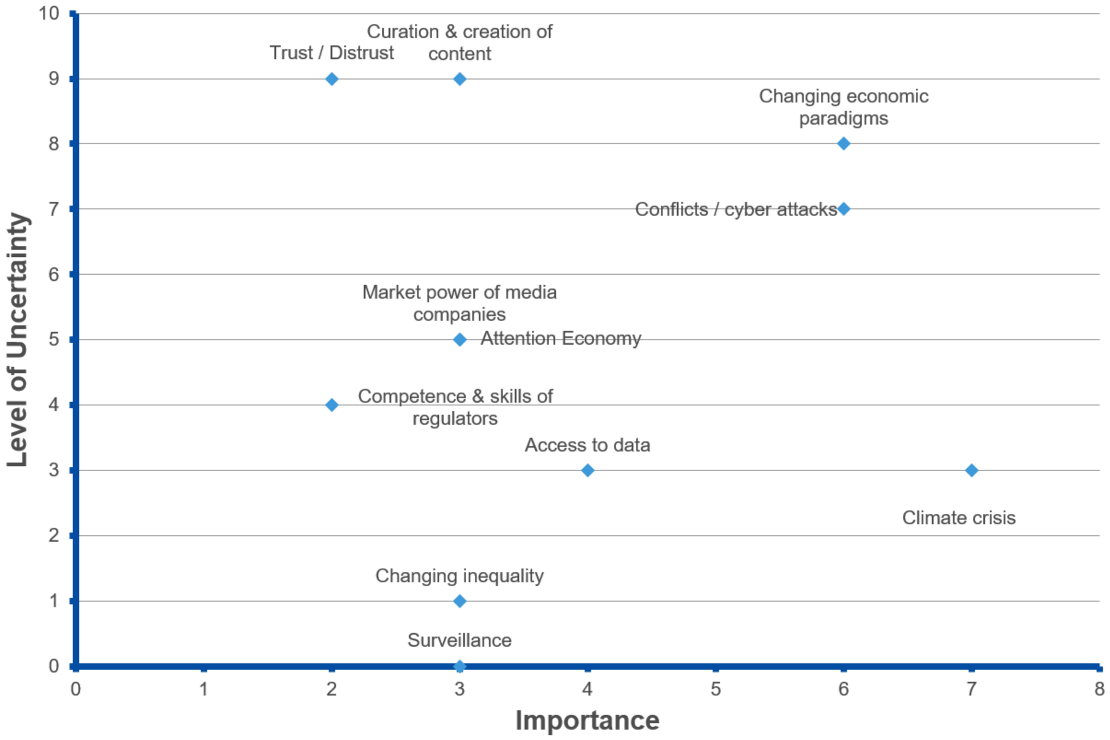


*Figure 6* - Identified drivers of change.

This process delivered a clear answer to which two drivers of change the group considered most important to drive the European information space towards the future while remaining the most uncertain regarding their future evolution: conflicts/cyber-attacks and changing economic paradigm. These two so-called "key uncertainties" were used as axes to build the logical space within which the scenarios were created.

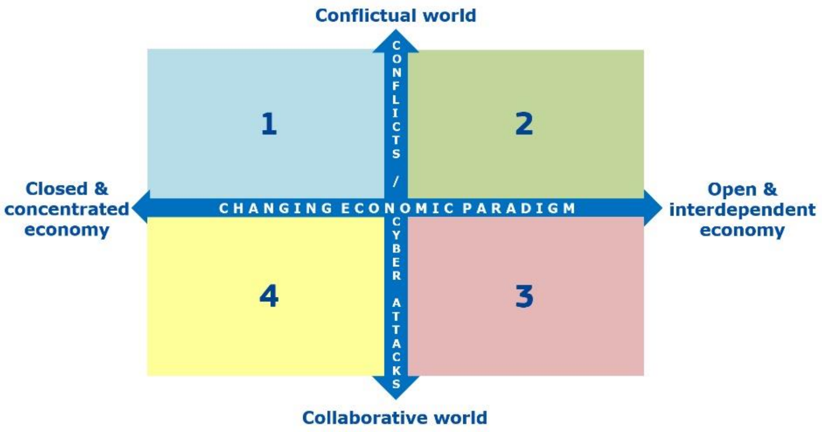

Conflictual world
Closed & concentrated economy
Open & interdependent economy
Collaborative world
CONFLICTS / CYBER ATTACKS
CHANGING ECONOMIC PARADIGM
1
2
3
4

*Figure 7* - Scenario logic.

In the afternoon, following clarifications about the axes and the scenario quadrants, the participants were split into four groups and engaged in a World Café, process to start developing the substance of the scenarios. The result of this work follows.

## Scenario 1: Struggle for information supremacy

The "Struggle for information supremacy" scenario assumes that the future European information space will be marked by high degrees of conflict and economic concentration.

**Regulatory environment.** Following intense economic protectionism and regulatory divergence, the internet has become geographically fragmented, with each region or state either having created their own digital-architecture champion or keeping a tight regulatory grip on a national sub-section of an international giant. Big companies with lots of technical capacity and oligopolistic market power are dominating all layers of the information architecture, ranging from the provision of broadband to the algorithmic curation and creation of content on social-media platforms. Information has become a deeply weaponised means of control, with national governments, political movements, businesses and foreign actors all struggling to steer the narrative on their own terms. A mostly disengaged public is circumventing the predatory nature of the European information space by focussing on apolitical, entertainment-oriented means of socialising online, losing touch with many social injustices in the process. The EU is struggling to keep at least some degree of coherence in the European information space, as Member States have introduced a patchwork of national legal frameworks.

**Societal impact.** Society is marked by deep distrust of what is seen in the information space. Public broadcasters, publicly subsidised media and a small number of partisan outlets are seen as the only somewhat reliable sources of information, while citizens greatly distrust "outsider" information from unknown sources. This has led to the emergence of a number of strongly exclusionary digital services that provide sharing and discussion spaces for small numbers of users with cryptographically verified profiles. Since these are however relying mostly on paid subscription plans, most citizens still use bigger, less protected platforms. Due to the intentional lack of protection from corporate predatory practices, only individuals who are either highly tech-savvy or are closely connected to the corporate/public/security nexus have access to privacy and anonymity supporting technologies. The intensity of the political and commercial competition online leads to a disillusioning effect in large parts of the population where there is a clear trend of using technology primarily in an escapist way: in a society where reporting on the "real" world is not to be trusted, citizens have largely withdrawn from the political sphere and socialise in game-like virtual spaces. This trend is facilitated by a wide uptake of virtual and augmented reality tools, allowing for immersive experiences with a close circle of peers.

**Political impact.** Despite these tendencies towards apathy, the conflictual, polarised nature of the information space also results in a rise of ambient nationalism in the parts of the population who still partly trust state-sponsored information online. In Europe, this trend is particularly strong in countries where societies are traditionally more homogenic and closed or where populist governments have managed to cement their power through increased control of the media and judiciary. Governments respond to the danger of losing legitimacy vis-à-vis a disengaged public by introducing public deliberation platforms. However, these services often suffer from a lack of sophistication, as their primary development focus in a disinformation-rich environment is to detect and remove malicious meddling activities. Strict authentication conditions that frequently involve the processing of biometric data create a severe access barrier that only few citizens are willing to cross. Big social media platforms are filling this void by implementing their own polling and deliberation features that see higher participation rates due to their embeddedness, resulting in one of the most important means of accessing information from citizens. Although not representative and used only by some sections of society, the introduction of these polling features gives digital corporations large agenda-setting power that increases their capacity to influence the political process disproportionately. In opposition to such superficial and manipulative means of organising societal discussion, some groups still use information technology in an emancipatory way to challenge the status quo, using highly protected alternative deliberation platforms aiming at fair and transparent consensus building. Counter-culture movements also try, with varying degrees of success, to sneak political protest and debate into the virtual socialising and gaming worlds of their apolitical peers. However, activist and partisan groups ultimately fail to mobilise bigger groups of people for societal causes. Encouraged by floods of inflammatory content, some radical splinter groups of both political and religious nature have fostered isolationist online communities that give rise to a surge in "stochastic terrorism", meaning lone wolf violent attacks based on an ideology that largely functions without clear leaders and organisational structures.

With regard to political advertising, the large-scale harvesting of citizens' data from their connected devices has led to widespread hyperpolarised advertising based on psychological traits. Both the creation and dissemination of content have become increasingly automated: communication experts feed base parameters of a political narrative or commercial campaign into an automated system that translates the input into blog articles, entertainment videos or political campaign pieces. In the most extreme cases, the output content is additionally fine-tuned to the user that accesses the content. As only a limited number of proficient users has the desire to anonymise their activities on the web, filtering and targeting technologies determine the majority of content that is seen by most individuals in society. Technological and societal trends such as personalisation and inward-looking retreat are also reflected in the growing importance of virtual assistants. These automated, extremely personalised systems are an integral part of life of most citizens, organising and displaying most of their news content, entertainment and personal organisation alike. Since only the biggest digital corporations develop and deploy such systems, the market choice is limited and the companies therefore hold a great degree of power over one of the most important means of information access.

Although it is illegal in most jurisdictions to deploy "automated personas" that cloak themselves as humans, such chatbots are active in vast numbers on all digital platforms with discursive or economic relevance. Extensively trained on vast datasets of human behaviour and drawing on automated text- and audio-visual creation algorithms, the software can simultaneously manage thousands of accounts that each have a coherent, life-like persona. Despite pouring great amounts of resources into their detection and removal capacities, digital giants mostly fail to curb this activity; only some citizens still use internet platforms as forums of discussion, while the vast majority is retreating. The highly negative reputation of automated personas has also stalled any effort by legitimate political candidates and parties to deploy chatbot versions of themselves to give voters an opportunity to interact more deeply with their political ideas.

**Technological impact.** Because governments favour protectionism and autarchy, there is very little product and software compatibility between the services of different digital giants and there is no interoperability between different social media services. The internet has experienced a fragmentation of international conversation as users are clustered along with their geographical location across a variety of regional and national social media. Identity verification technologies have become crucial for maintaining at least some basic level of trustworthiness on deliberative and social media services, but they frequently come at the cost of encroaching on user privacy. Relying frequently on biometric data and facial recognition, both governmental and corporate digital identities are based on central data storage rather than decentralised authentication solutions.

**Economic impact.** Large, powerful companies heavily dominate the economy of the European information space. As the legal framework regulating issues such as illegal content has become increasingly difficult for international platforms to adhere to, some European alternatives have emerged and, in some cases, have become dominant in the European market. In the cases where large international providers remained dominant for a given service, the divergent legal frameworks across the globe result in starkly different user experiences depending on where the service is accessed. While the harnessing of vast amounts of citizens' data does result in a continuous development and improvement of the various national and regional digital services, the lack of direct competition results in a general lack of disruptive innovation. Where innovation does happen, it is mostly focussed on deterrence and defence against disinformation and cyberattacks. This is not least because of a mutually beneficial interconnectedness between on the one hand the digital champions and governments that builds on lax legal frameworks and on the other hand cooperation of the private sector for state surveillance and censorship purposes. Another consequence of this closed economy with a highly weaponised information space, is the development of a vibrant cryptocurrency market, developed to avoid tax regimes and create subliminal subsidy mechanisms.

Despite the heavy dominance of the bigger commercial platforms, there are still some isolated instances of alternative services: relying on alternative, highly securitised and cryptographically protected platforms, some dedicated deliberation services serve as important discussion venues for the politically engaged. As a result of economic concentration and the state's defence orientation, media plurality has decreased, with a number of big publishers owning most of the European media landscape. Small and medium sized news organisations, as well as private blogs and commentary are struggling in the face of strict online distribution regimes: it is still possible to host one's created (lawful) content on a website, but the possibility to disseminate it via open standards or to share its URL freely on social media has become subject to tight regulation. In accordance with the respective national or regional sets of rules, the access to recommender systems such as newsfeeds on digital platforms depends on a mixture of automated indicators and evaluations of nationally authorised fact-checkers that favours strongly the actors with greater resources. With regards to news content, media organisations struggle with the high levels of forgery and disinformation and therefore focus mainly on promoting as much as possible a shared understanding of reality. In many cases, this comes at the expense of deeper coverage and investigative journalism. Because geographical location is seen as one of the primary indicators of trustworthiness for a news organisation in a given territory, the rise of high-quality automated translation only had a marginal impact on the European and international media and even large publishers are drawing their audiences almost exclusively from the national level.

**The role of the European Union.** Policy, both at the European and national levels, is heavily focussed on security and economic protectionism. State entities further their security agenda through close ties with large companies and are thus unwilling and unable to limit corporate surveillance and enforce strong human rights compliance regulation for emerging technologies. National militaries and secret services have become an integral actor in the information space, working both towards neutralising foreign influence operations and conducting their own psychological warfare operations. The efforts towards increasing digital and media literacy are stalled in order to make citizens more susceptible to the authorities' own narratives. Data that is generated or collected by authorities is rarely made available in a non-discriminatory way and is instead passed only to a few select companies that align closely with the state. In turn, these commercial actors gain even more semi-autonomous power, resulting in lower quality software products.

The role of the European Union in shaping policies regarding the creation, dissemination and processing of information is severely limited as Member States have increasingly resorted to national measures to protect and control their respective information spaces. As a consequence, cyber-espionage exists across some Member State borders, undermining trust in further integration of the European project. Where common rules are agreed, they are only accepted if they include far-reaching implementation discretion. Such fragmentation is also visible in regulatory authorities: most Member States maintain one or more public entities to enforce obligations such as algorithm audits on digital services, but their orientation, legal bases and degree of investigative capacity vary widely. Coordination at EU level is thus reduced to the smallest common denominator. The EU is trying hard to counter external disinformation but due to fragmented national initiatives and low degrees of trust, these efforts are frequently undermined from within.

The balkanised, non-centralised structure of the information space leads to a multitude of localised data centres and server centres, which due to economies of scale harnessed by the big corporations have moderate degrees of energy efficiency. The environment is, due to visible effects of degradation, a topic that media and political communication frequently covers, but information battles usually substitute serious discourse and policy action.

## Scenario 2: Resilient disorder

In the "Resilient disorder" scenario, the EU has fostered a competitive, dynamic and decentralised information space with strong international interdependence, but is facing continuous threats from sophisticated disinformation campaigns and cyberattacks.

**Regulatory environment.** Private companies and civil society initiatives increasingly lead the response to uncertainties, as the degree of regulatory intervention has decreased. While hyper-partisan groups frequently undermine the consensus that polarisation and disinformation are problems that should be tackled by everyone, a general level of resilience towards the adversities of the online information environment has emerged.

**Societal impact.** With an increased focus on individualism, societies in Europe have become fragmented among various political and economic cleavages, which public policy generally does little to address. As a result of the high degrees of conflictual information, online tribalism has increased and is often visible in the clustering of communities along services whose content moderation and blocking policies members agree with the most.

Such grouping leads to somewhat heavier societal polarisation as echo-chamber effects between the different social media providers increase. As a result, these tribal means of accessing information also increase the general level of trust in the accuracy of online content that users receive through such channels as it conforms with their ontology of truth. Most services have developed robust moderation policies. This increases trust but does not generally reduce susceptibility to disinformation, although such tendencies are observable in some hyper-partisan communities. As tribal communities between or within different providers can still federate and access each other's content, politically engaged citizens and movements frequently compete to dominate the societal discussion. This can lead to emotional sensationalism and competition for the best argument.

**Political impact.** Despite many citizens actively participating in such online discussions, the building of bottom-up resilience capacity to disinformation has by far not rendered everyone politically active: many users opt instead to see as little political content as possible and focus on apolitical socialising and interactions. This move away from social media as a venue of political discourse is also facilitated by the introduction of both national and pan-European political deliberation platforms. Because these public services offer central means of societal discussions, large parts of society are beginning to expect their political peers to settle their differences there. However, the outcome of consensus-driven deliberative platforms is not unequivocally trusted: despite strict cybersecurity and strong user authentication, incidents of on-platform discussions that were significantly skewed by domestic and foreign intervention have occurred several times and both the open-source community and private contractors have yet to find a way to effectively shield their deliberation projects. Another means of maintaining a shared understanding of reality are collaborative projects to counter disinformation: along with the greater diversity of services and their interoperable APIs (Application Programming Interface), a number of professional as well as non-profit services are able to be active across a multitude of platforms to provide fact-checking or moderation services. While these do not fully offset the tribal nature of the served communities, their collective intelligence approach to mitigate disinformation has gained them trust and respect across most of society. This has a negative effect upon the creation of EU-wide collective narratives. As information becomes increasingly individualised, Member States offer different national narratives. Consequently, fact-checkers that correct an inclusive narrative, may be perceived as foreign intruders by some countries while others, whose own narratives are more closely aligned with the EU, would welcome such efforts. This would likely lead to increased political polarisation with inclusive EU rebuttals being met with anti-EU rhetoric campaigns.

**Technological impact.** Due to the enforcement of open standards and profile portability, digital services in the European information space have become increasingly decentralised. As competition on quality features such as privacy protection, communication features or content moderation standards is high, the pace of innovation is moderately fast and leads to continuous industry-wide development. Being confronted with a multitude of partisan, state-led and corporate disinformation operations, innovation is however commonly pivoting mostly towards improving security and trustworthiness. Common points of improvement include account verification or the accurate detection of forged content.

Providers of social media services however frequently diverge on their approach to curation and targeting technologies. While some open source platforms and social enterprises have built platforms that combine community maintenance with technological approaches to actively mitigate issues such as information overload or radicalisation, other niche providers have conversely pivoted towards the opposite: they want to keep their members perpetually outraged and politically engaged. In this process, the value orientation of the service chosen by a user gradually undermines cross-societal political ideologies. While this value orientation is in some instances closely aligned with universal human rights, other services— including most of the bigger ones—invoke and respect such norms only insofar as they serve their commercial interests. In addition, bigger platforms in the market can, despite official commitments to data protection, still leverage significant data harvesting from the activity and inferences on their user bases, marketing the use of behavioural targeting as a trade-off for increased security. There is also a high degree of topical divergence of platforms. While there is a multitude of services available for different purposes (e.g. for maintaining business contacts or exchanging ideas on niche hobbies), users generally lock the content they post on such dedicated platforms from being displayed on services serving another context.

Authenticated digital identities supported by decentralised technology (such as distributed ledgers) are a key feature for maintaining trust in the online environment. Anonymous and pseudonymous platforms and fora still exist but are, due to the insecurity induced by psychological warfare operations, usually only used by tight-knit communities that seek isolation. As most services therefore rely on granting access only to authenticated personas, the data attached to these profiles is extremely comprehensive and personal. In turn, the cybersecurity of these profiles is of paramount importance. There is a booming market for software that manages data access in accordance with citizens' preferences. Such access management is also common practice when it comes to systems facilitating and organising information for users, for instance virtual assistants. While in many of these services the data of authenticated profiles does not leave its storage and processing is happening on-device, some of the bigger commercial offers still seize the opportunity to gather troves of personal data from the use of their assistants.

Decentralised technology is also commonly used to provide transparency about the origin of a piece of content but given the flood of information available and the common re-mixture of cultural production, the impact on trustworthiness remains limited. Endeavours to solve the challenge of disinformation technically also run into an impasse because of an ongoing arms race with technology designed to create and disseminate disinformation. The proliferation of deepfake audio-visual content, automatically generated false news articles and sophisticated campaigns involving automated accounts is only slightly mitigated by initiatives such as industry-standards on timestamping and digital watermarking of original digital material.

Highly sophisticated automated personas are active across all instances of digital media. While governments had some initial success at curbing their spread by enforcing legislation requiring stricter authentication policies throughout the market, the growing sophistication of malicious chatbot systems ultimately rendered none of the federated social networks fully secure. Social media services of all sizes collaborate with volunteers in the open source community to develop common high standards of detection and removal systems. However, the arms race is overall still weighing in favour of foreign governments, domestic actors and, in some cases, private corporations who are deploying their inauthentic armies for political or economic gain. Debates in online spaces that rely on unauthenticated profiles thus become increasingly disfavoured by the public, who either move towards services that rely fully on governmental electronic authentication or cease to use social media for interactions with anyone other than their closest relatives and peers.

**Economic impact.** In principle, the introduction of profile portability and interoperability standards has made the economy of the European information space more competitive. Political advertising, as a consequence would take place on digital platforms via an interoperable, open standard (i.e. the same ad appearing on Facebook, Twitter and others), with options for microtargeting being limited by the amount of data that users allow to be accessed for advertising purposes (most platforms having quite granular controls). Microtargeting would still frequently happen on the bigger, "free" platforms. Nevertheless, such ads would be blocked by default by some providers who see it as a competitive advantage for their users not to receive political advertising. It follows that weakly enforced regulation based upon payment transparency of the ads is frequently overstepped by disinformation actors. Although the technical standard for political ads would be the same globally, national and regional laws would impose obligations on service providers to block or flag certain types of political advertising

However, the conflictual nature of the information environment has, despite the widespread availability of alternatives, led to a situation where most users rely on the services of a few select companies that are seen as the most protective and technologically advanced. Although far less monopolistic than in previous times, bigger digital platforms and services still scoop up the most significant share of revenue — not least because their advertising-based model grants free access to the service — whereas more secure alternative providers mostly function based on a subscription model. Competition is also skewed by an increasing reliance on public-private partnerships, initiated by public authorities in the hope that government sponsored authentication and filtering technologies will boost their societies' digital resilience.

When it comes to news media, the market features some remaining big, widely respected publishers and a growing multitude of small media, hyper-partisan outlets and blogs. Given the high level of disinformation and the diversification of news sources along granular political axes, maintaining a shared understanding of truth remains a fundamental challenge for the media landscape. Since the EU's open internet and economy model does not discriminate against external media, the media diet of European citizens has internationalised to a certain degree thanks to enhanced automated translation. This gives many users unprecedented access to content that was not created in their mother tongue. On the upside, this facilitates a better understanding of global events but on the downside also makes state-led disinformation campaigns of foreign actors much easier. Finally, the economy of the European information space frequently faces dangers to the integrity of its supply chains: geopolitical tensions around the attribution of influence campaigns sometimes escalate into foreign actors shutting off the supply of critical raw materials for hardware production or threatening to seize servers of companies that store the data of many Europeans. This has led to the creation of digital embassies at the national level, while the EU has put in place backup infrastructure for the entire European information space.

**The role of the European Union.** Despite facing significant degrees of internal and geopolitical manipulation attempts, the EU's commitment to a free and open internet is widely acknowledged but there are also frequent calls from citizens and Member States for tougher regulatory action on online content. EU policy has, however, focussed on providing higher degrees of transparency and data protection, for instance by requiring algorithm audits by competent regulatory authorities. While relying mainly on the private and civil society sector to decide how to address disinformation (e.g. by technological means or collective intelligence), the EU also concentrates on identifying and sanctioning malicious actors through enhanced attribution and imposition of sanctions regarding cyberattacks and influence operations. The Member States mostly follow this line but some countries have nevertheless introduced stricter liability and safety rules in their jurisdiction. In turn, market access in these territories is unviable for many smaller providers and the digital single market is therefore partly distorted. Such incoherence in turn undermines the capacity for coordination among the various national regulators at European level, making regulatory compliance with European standards depend largely on a company's legal residence.

The open economy of the European information space allows economies of scale for the storage of data, curbing energy costs. Despite climate change unfolding with tangible consequences, it remains a contentious issue that is still frequently subject to massive dis- and misinformation. However, many civil society led initiatives frequently collaborate on debunking and education campaigns and some of the more progressive providers of social media services have introduced strict filtering or rigorous community-led moderation to keep misinformation in check.

## Scenario 3: Global cutting edge

The “Global cutting edge” scenario foresees a world in which societal and geopolitical conflict have been reduced significantly, while high degrees of competition and innovation have led to the emergence of a dynamic, global information space.

**Regulatory environment.** Through strong anti-trust and interoperability frameworks, the EU has significantly contributed to the diversification of the online information economy and has cemented its role as an international innovator of technological and privacy standards. In an environment of intense economic, technological and collaborative dynamism, citizens and companies leverage the true potential of collective intelligence. However, issues such as tribalism, productivity-centred culture and a retreat from the ideal of collective goods can also make this environment mentally challenging for many European citizens.

**Societal impact.** A broadly peaceful world and a general spirit of collaboration and openness have fostered a climate of trust in networked societies. Information found online is mostly accurate and reliable and is provided by a vibrant mix of public broadcasters, small and large independent media and collective intelligence networks. Private individuals also frequently engage in the sharing and discussion of information in a generally collaborative manner. Enabled by auto-translation, in a low conflictual world, participation in political discussions is far-reaching, as governments have embraced deliberative online platforms both at the national and supranational levels that allow for nuanced, large-scale debate in a non-polarised manner. While political content is still widely shared on other social media, societal polarisation has decreased because of these participatory policy dialogues. Interoperable services with diverging curation policies have contributed to a situation where users have a wide degree of control over what they see online, which on the one hand reduces polarisation, but on the other hand also limits wider discussions on societal values as communities largely resort to their own moral standards. Groupthink is also a widespread phenomenon in some user clusters which, in cases where it remains unaddressed, results in less valuable output.

**Political impact.** Supported by decentralised technology, authenticated profiles that uniquely identify citizens are widely available, either for online participation in political matters or trust-based civil or economic cooperation. Since popular contributions tied to one’s verified persona have become an important source of societal status, many citizens actively try to be as productive and popular as possible—sometimes with the adverse effect that less motivated community members are looked down upon. However, anonymous profiles or profiles that are untied to an authenticator remain the norm for most services and most citizens actively keep both systems separate thanks to high levels of digital literacy. Authenticated personas also support a thriving deliberative platforms scene, including both private or non-profit services that cater to the needs of specific communities and public websites that facilitate larger societal discussions. These (mostly open source) services can federate with other digital infrastructures and are thus frequently embedded into social media newsfeeds and digital newspapers.

Participation is broad and the bridging of otherwise less connected communities significantly increases the quality of the discourse. Benefitting from a dynamic information economy with low degrees of tax avoidance, inequalities in society are mitigated by redistributive social policies. There are escapist tendencies in some parts of society that engage with political content only as far as necessary and otherwise use the information space to socialise and share non-political experiences such as games. For many citizens, adhering to this trend is actually a method of coping: because of society’s emphasis on productivity, dynamism and collaboration, those unable or unwilling to submit to this paradigm often feel excluded and less valued, leading to a small but significant spike in mental health problems.

One way of organising the abundance of information and the coordination of large networks of collaborating workers and volunteers is through dedicated automated systems. Such software can include summary and text mining algorithms or highly developed artificial moderator personas. The latter are complex models trained to display human traits such as compassion, authority and empathy in order to fulfil their task of coordinating both small and vast numbers of people in their common work. Automated moderators are almost always clearly labelled due to legal requirements and are often open sourced. Despite these transparency measures and the significant efficiency gains, many citizens struggle to adjust to the reality of automated systems directing much of their work. Consequently, issues of emotional well-being and human autonomy are gaining increasing traction in societal debates. Additionally, "bot wars" are becoming more common, referring to a situation where users deploy adversarial automated systems to sway moderators into their direction. In some cases political entities and activist movements also use digital automated personas as official representations of themselves. At the disposal of citizens, these unregulated systems can increase depth of thought interaction with political material but are seen as manipulative by some observers and questions over their legitimacy slowly arise.

**Technological impact.** Technology has become increasingly focussed on sharing and collaboration: crowd funding, participatory networks and validation technologies contribute to the information space's trust and openness. Decentralisation has become a guiding technological principle: from protocol interoperability of digital services to mesh networks and secure multi-party computing, data storage and processing in a central manner has mostly become outdated. Digital services have developed highly granular choice interfaces that allow users to choose their desired level of personalisation, both with regard to advertising and to the curation of content in newsfeeds and similar products. To help citizens cope with information overload, curation remains key. Apart from the recommender systems of social and collaborative platforms, the packaging and presentation of information is frequently managed by personalised virtual assistants. Since these systems only access, not copy, personal data from a given source and process the information fully on-device, privacy is mostly preserved. Thanks to the availability of open source products, citizens also have transparent insight into such crucial software. Generally, the legal and technological shift towards privacy protection has made unwarranted data collection difficult. As a result, the dominance of online behavioural advertising has decreased and contextual advertising is flourishing on media websites, supported by sophisticated automated systems that match content with advertisement.

Nevertheless, there is also a market for secretive tracking technology that links users' (anonymous) profiles to their real life, verified identity e.g. anonymous activity in a political forum can be linked to verified identity. These tools are used by both economic and political actors to improve their campaign narratives and targeting but an ongoing arms race with mostly privacy-oriented social services renders this a difficult and expensive undertaking. New means of accessing the information space, such as virtual and augmented reality (VR and AR), have spread far and wide and offer new means of entertainment and collaboration. While social, non-political games have proliferated and are used by large parts of society, VR and AR have also contributed to emancipatory and participatory usage. These can for instance include attending a realistic, online campaign rally or "in-person" coordination of activist activities.

There are a number of indirect effects on the environment and climate crisis. Firstly, the collaborative and non-conflictual manner of the global information space leads to faster development and spread of promising adaptation technologies. While the energy demands of a digitised society are high and threaten to leave a large carbon footprint, economies of scale that are harnessed by globally operating computing and cloud services have a mitigating effect.

**Economic impact.** Economically, the global information space is a dynamic mix of many competing digital services. Strong interoperability requirements have led to a diversification of platforms, giving citizens the choice among a large range of providers whose products compete on issues such as privacy protection, quality of service, design, curation, etc. This competition results in fast-paced innovation, as best practices are frequently adopted by other market players and disruptive improvements are key to retain users in the long term. Business models of digital services vary but rely mostly on contextual advertising, subscriptions, micro-payments or donations. In contrast, behavioural microtargeting has largely disappeared as a result of the continuous development of encryption standards co-developed by the private and civil society sectors, as well as hyper-granular consent and privacy options applied by most digital services.

The emergent, collaborative platform economy provides for a wide range of jobs and roles, making the distinction between remunerated labour and voluntary work increasingly fluid. The news media industry has experienced an impressive revival but looks very different when compared to earlier days of the internet. While a few — mostly high quality — trans-regional, big newspapers are still going strong, the prevalence of online group spheres and increasing professionalism outside of traditional media has also fostered a proliferation of small news sites, professional blogs and citizen journalism that steadily increase their share of revenue and influence. Due to the effect of automated translation, the vast majority of web and news content has become available in a multitude of languages. Initially, this resulted in a direct competition of many previously national media on the global scene, which in some cases led to the disappearance of these outlets. In the meantime, however, this loss of bigger, multi-issue news organisations has been compensated by a vast increase in highly specialised, small scale media that cover almost every aspect of human existence, regardless of whether the global audience interested in such content consists of 500 or five million persons. Regardless of size and market power, all outlets for the journalistic profession have benefitted from the impact of more privacy protected digital services as the decrease of on-platform behavioural advertising led to a revaluation of advertising on quality media.

**The role of the European Union.** Policy is focussed on enforcement of data protection, competition and interoperability legislation, leaving most issues in the information space to be solved by private and civil society innovation. Independent regulators with strong monitoring powers play a key role in this regard, also auditing algorithmic systems for compliance with legal frameworks such as labour or anti-discrimination law. Such enforcement is tightly coordinated at the European level to prevent a fragmentation of the digital single market, however the participation in global fora is also gaining increasing importance due to growing recognition that international coordination is crucial in a borderless information space. Intellectual property legislation has developed advanced “fair use” clauses to facilitate sharing and open access and a proliferation of open licenses has contributed to a general drop in piracy.

## Scenario 4: Harmonic divergence

The “Harmonic divergence” scenario assumes a world in which strong regulatory differences and economic protectionism between national and regional actors have resulted in a fractured global information space.

**Regulatory environment.** The general low level of international and societal conflict has prevented a crowded and predatory information environment for citizens. The EU has steadily expanded its role in setting — sometimes lauded, sometimes criticised — regional standards for digital services, ranging from data protection to illegal content. Emphasising the need for the online environment to conform with European values, this trajectory is in line with a general focus on common policies. Following a slow but steady long-term pace of innovation and policymaking, the stability of the European information space relies on the favourable conditions of reduced polarisation and foreign interference. This breeds vulnerability should such conflicts re-emerge.

**Societal impact.** Society in the European information space has become more inward looking. Facilitated by an ongoing policy focus on shared values, redistributive policies and social equality, the “European model” actively participates in international cooperation while simultaneously promoting and protecting its own values and standards. There is a tendency towards more regionalised content, based on a reliance on big platforms that are either based directly in Europe or that have implemented heavily localised versions of their services. Trust in information found online is generally high but the awareness of filtering technologies and geographically dispersed services gives some citizens an increasing feeling of not getting the full picture. While many subscribe to digital services that are dominant in other parts of the world to get additional information from foreign sources — in some cases even via virtual private networks (VPNs) — most people rely primarily on the dominant European and national digital platforms. Due to the low level of societal conflict and egalitarian policies, politics does not dominate most citizens' online experiences. The main uses of both traditional hardware as well as new means of access such as virtual reality remain mostly apolitical. However, both national governments and the EU have had some success in engaging citizens via dedicated deliberative polling websites. Although often contracted out to external providers who only give low degrees of transparency into the workings of moderation and information processing, these services mostly manage to secure the trust of those citizens passionate enough to use them. Through negotiations with bigger social media platforms to make them embed polling links into their digital architectures, public authorities manage to gradually increase participation of citizens. This however comes at the cost of further cementing the position of these platforms in the market. In addition to governmental deliberative websites, some platforms administered by civil society and non-profit enterprises also exist but are rarely used by most internet users. When it comes to the sharing of information and collaboration in collective intelligence processes, citizens and businesses benefit and actively make use of the European information space's internal coherence. This for instance includes the consolidation of intellectual property regimes to public oversight over platform algorithms. However, the leveraging power of big digital players persists and the diverging trajectory of the EU's digital economy makes cooperation with the rest of the world more difficult. As a result, the dynamism of collective intelligence is not as high as it could be and the EU stands at the verge of losing out vis-à-vis economies that took a more open path to innovation.

**Political impact.** In the absence of geopolitical and domestic conflict, malicious bot-networks and automated personas do not hold a large presence in the European information space. Far-reaching EU legislation on labelling and operating apply to automated persona systems coordinating or otherwise affecting the work, social lives and political decisions of citizens. While these standards do increase the trust placed in software agents increasingly contributing to economic sectors, the lack of transparency in their inner workings (e.g. because it is proprietary software) frequently raises legitimate fears of dark patterns and threats to human autonomy. Politicians, political parties and international organisations have developed automated personas in order to give citizens the opportunity for a “personal” interaction. Although for many citizens this has sparked greater attachment to politics, the deployment of such placeholder chatbots — in particular those of the most publicly known actors — is also confronted with increasingly difficult challenges like cybersecurity and reputation management.

**Technological impact.** As larger platforms dominate the market and continue to benefit from significant network effects, innovation in digital services is steady but rarely makes big leaps. Technologies related to the creation and dissemination of content are constrained by strict sets of rules. For instance, the algorithmic curation of content needs to balance the blocking of illegal content with carrying obligations for political speech. Where automated systems for content creation, alteration and dissemination are deployed, transparent labelling obligations apply. While this approach has facilitated the emergence of a more protected European information space, the insurmountable imperfections in filtering technologies and their underlying legally mandated moderation policies are frequently subject of heated debates on the rights to freedom of expression and access to information. Trust in this model of internet governance has also declined in light of some documented abuse cases in which the implementation of the curation guidelines led in some Member States to discrimination against opposition candidates and social movements.

Tracking and corporate surveillance have intensified and product advertisements are heavily personalised. Political advertising is also making use of such targeting options. However, due to EU internal legislation on transparency requirements and data protection, such targeting is generally not very sophisticated. Authentication technologies such as cryptographically verified identities have proliferated and are mainly issued by national governments. While the tying of inferences from user behaviour on a given service to such an identity is in principle subject to strict consent obligations, big digital corporations are partially successful in circumventing legal constraints due to loopholes and a general lack of competition. The cybersecurity of most online services is however strong,and therefore authenticated identities are rarely used to gain service access.

**Economic impact.** There is far-reaching economic concentration in the digital single market, with a select few platforms reaping the main economic benefits from providing means of information creation and dissemination. Strict regulatory frameworks on data protection, local storage, content moderation and illegal content have contributed to a partial withdrawal of international platforms from the European information space, giving initially innovative European services an edge that resulted in a limited number of European digital giants. Without profile portability or interoperability requirements, big players can benefit from strong networks, making it hard for smaller competitors to enter and stay in the market. The lack of direct competition results in low innovation with most information services focussing their development on incremental improvements. As start-ups and smaller services get acquired quickly at international and European level, the level of dynamism in the European information space economy is limited. Regarding media organisations, there is a clear trend towards internationalisation due to the widespread uptake of automated multilingualism. Supported by measures such as mandatory revenue sharing from digital distribution channels and in some cases even via public funding, Europe's media landscape has consolidated and its quality outlets are able to compete at international level. As the funding and business models facilitated by public policy mainly benefit larger players, small and medium sized media are often at an impasse when it comes to long-term stability. They nevertheless play a vital role in the European information space often due to their thematic specialisation.

**The role of the European Union.** Politically, the EU has achieved further integration on digital matters as Member States have collectively pushed for stronger digital sovereignty. Following a utilitarian mode of policymaking, most interventions into the European information space are guided by an emphasis on collective goods and European values. Access to data held by public authorities is usually granted mostly for projects with a clear social or economic benefit in mind and digital services are required to share parts of their company data for purposes such as academic research and public projects. While heavy-handed legal frameworks on data protection, intellectual property, intermediary liability and content moderation have sparked some degree of international tension, they have also helped to create an integrated and well-functioning European digital single market. In turn, policy has shifted from reigning in large platforms' market dominance towards strong oversight and moderation rules. Having largely overcome initial problems of attracting skilled staff, a network of European digital service regulators with a single board at European level now tightly monitors the status of implementation of EU rules. Internal borders for intellectual property have mostly been abolished, facilitating sharing and collaboration in Europe. Far-reaching legal obligations to counter terrorist content, disinformation or social discrimination on digital intermediaries have resulted in the European information space becoming increasingly separate from the global information community. Some instances of Member State governments using European legal frameworks such as counter terrorism to force removal and de-prioritisation of political content that they define as extremist have also raised many concerns among civil society.

Regarding the environmental impact of digitalisation, the EU has imposed resolute standards for the energy consumption of ICT services. Bringing carbon border adjustments into the online sphere, these standards have in some cases resulted in heavy fines imposed against foreign digital services. However, the insistence on geolocation of servers and the reluctance to foster more decentralised methods of data storage does ultimately still result in comparably high energy costs of the European information space.

*Table 6* - The future of European information space 2035 at a glance

| | **Scenario 1**<br>**Struggle for information supremacy** | **Scenario 2**<br>**Resilient disorder** | **Scenario 3**<br>**Global cutting edge** | **Scenario 4**<br>**Harmonic divergence** |
|---|---|---|---|---|
| **Scenario axes** | **Conflictual World / Closed & concentrated economy** | **Conflictual World / Open & interdependent economy** | **Collaborative World / Open & interdependent economy** | **Collaborative World / Closed & concentrated economy** |
| **Trust in online information** | Information online, as well as news reporting on the "real" world, is trusted very little.<br>Public broadcasters, publicly subsidised media and a small number of partisan outlets are seen as the only somewhat reliable sources of information. "Outsider" information from unknown sources is rarely trusted. | Societal resilience facilitates some degrees of trust in the accuracy of digital content, but the struggle to maintain a shared understanding of reality persists in both social and news media. | Popular online contributions tied to one's verified persona are an important source of societal status.<br>Emphasis on productivity, dynamism and collaboration, those unable or unwilling to submit to this paradigm often feel excluded and less valued. | Trust in information found online is generally high but the awareness of filtering technologies and geographically dispersed services gives some citizens an increasing feeling of not getting the full picture. |

|  | **Scenario 1**<br>**Struggle for information supremacy** | **Scenario 2**<br>**Resilient disorder** | **Scenario 3**<br>**Global cutting edge** | **Scenario 4**<br>**Harmonic divergence** |
|---|---|---|---|---|
| **Political engagement and participation** | Generally low engagement, high degree of escapism with few exceptions<br>Rise of ambient populism and nationalism.<br>Counter-culture movements try, with varying degrees of success, to sneak political protest and debate into the virtual socialising and gaming worlds of apolitical peers.<br>Public political deliberation platforms have poor sophistication and are scarcely used by citizens, not least because of their biometric access requirements. | Moderate level of political engagement and political news sharing.<br>Move away from political discussions on social media towards dedicated platforms. Many users choose not to see political content at all.<br>Political deliberation platforms are used but have been subject to meddling and cyberattacks, resulting in lower degrees of trust in their outcomes. | Participation in political discussions is far-reaching as governments have embraced deliberative online platforms that allow for nuanced, large-scale debate in a non-polarised manner.<br>Political content is widely shared on social media, societal polarisation has decreased as a result of facilitated participatory policy dialogues.<br>Escapist tendencies in parts of society engaged with political content only as far as it is necessary and otherwise use the information space to socialise and share non-political experiences. | Politicians, political parties and international organisations have developed automated personas in order to give citizens the opportunity for a “personal” interaction. Although for many citizens this has sparked greater attachment to politics, the deployment of such placeholder chatbots – in particular those of the most publicly known actors – is also confronted with increasingly difficult challenges like cybersecurity and reputation management. |

| | **Scenario 1**<br>**Struggle for information supremacy** | **Scenario 2**<br>**Resilient disorder** | **Scenario 3**<br>**Global cutting edge** | **Scenario 4**<br>**Harmonic divergence** |
|---|---|---|---|---|
| **Collaboration and sharing** | High degrees of escapism lead to focus on common online social experiences.<br>Hostile information environment severely limits economic productivity. | Focus on individualism, fragmentation among various political and economic cleavages.<br>Online tribalism has increased and is often visible in the clustering of communities along services whose content moderation and blocking policies they agree with the most. | Private individuals frequently engage in the sharing and discussion of information in a generally collaborative manner.<br>Social, non-political games have proliferated and are used by large parts of society, VR and AR have also contributed to emancipatory, participatory use cases. | When it comes to the sharing of information and collaboration in collective intelligence processes, citizens and businesses benefit and actively make use of the European information space's internal coherence.<br>Internal borders for intellectual property have mostly been abolished, facilitating sharing and collaboration in Europe. |
| **Pace and focus of innovation** | Little product and software compatibility between the services of different digital giants, no interoperability and portability between different social media services.<br>Lack of direct competition results in a general lack of disruptive innovation. | As competition on quality features is high, the pace of innovation is moderately fast and leads to continuous industry-wide development.<br>Innovation is commonly pivoting mostly towards improving security, surveillance and trustworthiness. | Decentralisation is a guiding technological principle: From protocol interoperability of digital services to mesh networks and secure multi-party computing, data storage and processing in a central manner has mostly become outdated.<br>Market for secretive tracking technology that matches citizens' various profiles with each other, in particular with their verified identity. | Technologies related to the creation and dissemination of content are constrained by strict sets of rules. |

|  | **Scenario 1**<br>**Struggle for information supremacy** | **Scenario 2**<br>**Resilient disorder** | **Scenario 3**<br>**Global cutting edge** | **Scenario 4**<br>**Harmonic divergence** |
|---|---|---|---|---|
| **Content curation and targeting** | Creation and dissemination of content have become increasingly automated. Widespread use of automated systems which translate human input into blog articles, entertainment videos or political campaign pieces.<br>Little to no anonymity, high degrees of filtering and targeting thus determine the majority of content that is seen by most individuals in society. | On bigger platforms, behavioural targeting is a trade-off for increased security.<br>There is a booming market for software that manages data access in accordance with citizens' preferences.<br>Some open source platforms and social enterprises combine community maintenance with technological approaches to actively mitigate issues such as information overload or radicalisation. Other providers have pivoted towards the opposite in order to keep their members perpetually outraged and politically engaged. | Highly granular choice interfaces that allow users to choose their desired level of personalisation, both in regard to advertising and the curation of content in newsfeeds and similar products.<br>Interoperable services with diverging curation policies have contributed to a situation where users have a wide degree of control over what they see online.<br>New means of accessing the information space, such as virtual and augmented reality, have spread far and wide in society and offer new means of entertainment and collaboration. | Algorithmic curation of content needs to balance the blocking of illegal content with carrying obligations for political speech. Where automated systems for content creation, alteration and dissemination are deployed, transparent labelling obligations apply.<br>Tracking and corporate surveillance have intensified and product advertisements are heavily personalised.<br>Political advertising is also making use of such targeting options, but due to EU internal legislation on transparency requirements and data protection, is generally more limited regarding targeting sophistication. |

| | **Scenario 1**<br>**Struggle for information supremacy** | **Scenario 2**<br>**Resilient disorder** | **Scenario 3**<br>**Global cutting edge** | **Scenario 4**<br>**Harmonic divergence** |
|---|---|---|---|---|
| **Authentication** | Identity verification technologies are crucial for maintaining a basic level of trustworthiness on deliberative and social media services, relying frequently on biometric data and facial recognition. | Verified digital identities, supported by decentralised technology (e.g. distributed ledgers), are a key feature for maintaining trust in the online environment.<br>The data attached to these profiles is extremely comprehensive and personal.<br>Decentralised technology is commonly used to give transparency about the origin of a piece of content (with limited effect). Initiatives such as industry-standards on timestamping and digital watermarking of original digital material has limited effect. | Supported by decentralised technology, centralised, “verified” identities that uniquely identify citizens are widely available, either for online participation in political matters or trust-based civil or economic cooperation.<br>Anonymous profiles remain the norm for most services and are kept untied from verified identities. | Authentication technologies such as cryptographically verified identities have proliferated and are mainly issued by national governments. While the tying of inferences from user behaviour on a given service to such an identity is in principle subject to strict consent obligations, big digital corporations are partially successful in circumventing such legal constraints through remaining loopholes and a general lack of less invasive competition. |

| | **Scenario 1**<br>**Struggle for information supremacy** | **Scenario 2**<br>**Resilient disorder** | **Scenario 3**<br>**Global cutting edge** | **Scenario 4**<br>**Harmonic divergence** |
|---|---|---|---|---|
| **Disinformation and information operations** | Sophisticated automatic generation tools are being used for highly granular, targeted political advertising campaigns.<br>Foreign and domestic forces interfere with online deliberation processes on dedicated platforms.<br>Efforts towards increasing digital and media literacy are stalled in order to make citizens more susceptible to the public authorities' own narratives. | Online communities are competing - trying to control the narrative. Proliferation of deepfake audio-visual content, automatically generated false news articles and sophisticated campaigns involving automated account<br>Foreign and domestic forces are interfering with online deliberation processes on dedicated platforms. | Disinformation exists but this is not a broad societal issue. | Absence of information operations, but increasing tensions over centrally mandated information curation legislation. Some Member States are accused of using EU acquis to censor political opponents. |

| | **Scenario 1**<br>**Struggle for information supremacy** | **Scenario 2**<br>**Resilient disorder** | **Scenario 3**<br>**Global cutting edge** | **Scenario 4**<br>**Harmonic divergence** |
|---|---|---|---|---|
| **Digital platforms market** | Heavily dominated by large, powerful companies.<br>Some European alternatives have emerged and, in some cases, have become dominant in the European market.<br>Fragmentation of international conversation, users are clustered along with their geographical location across a variety of regional and national social media.<br>Limited number of secure sharing and discussion spaces for small numbers of users. Reliance on paid subscription plans. | Competitive, mostly decentralised information space with strong international interdependence.<br>Majority of users rely upon the services of a few select companies that are seen as the most protective and technologically advanced.<br>Some Member States have introduced national liability and safety rules in their jurisdiction, partially distorting the digital single market.<br>More secure alternative providers mostly function based on a subscription model.<br>High degree of topical divergence of platforms, users generally lock the content they post on such dedicated platforms from being displayed on services serving another context. | Dynamic mix of a multitude of competing digital services.<br>Strong interoperability requirements have led to a diversification of platforms, giving citizens the choice between a variety of providers whose products compete on privacy protection, quality of service, design, curation etc.<br>Fast-paced innovation. Disruptive improvements are key to retain users in the long term.<br>Business models of digital services vary but rely mostly on contextual advertising, subscriptions or donations. | Far reaching economic concentration in the digital single market. Select few platforms reap the main economic benefits from providing means of information creation and dissemination.<br>Strict regulatory frameworks on data protection, local storage, content moderation and illegal content have contributed to a partial withdrawal of international platforms.<br>Without profile portability or interoperability requirements, big players benefit from strong network benefits.<br>Start-ups and smaller services are acquired quickly on international and European levels but level of dynamism in the European information space economy is limited. |

| | **Scenario 1**<br>**Struggle for information supremacy** | **Scenario 2**<br>**Resilient disorder** | **Scenario 3**<br>**Global cutting edge** | **Scenario 4**<br>**Harmonic divergence** |
|---|---|---|---|---|
| **News media market** | Media has become less investigative, focussing mainly on promoting as much as possible a shared understanding of reality. | News-media market features some big, widely respected publishers and a growing multitude of small media, hyper-partisan outlets and blogs. | Revival of news media, some few – mostly high quality – transregional, big newspapers are still going strong.<br>Prevalence of online group spheres and increasing professionalism outside of traditional media foster a proliferation of small news sites, professional blogs and citizen journalism increase share of revenue and influence.<br>Decrease in on-platform behavioural advertising leads to a revaluation of advertising on quality media. | Europe's media landscape has consolidated and its quality outlets are able to compete on an international level.<br>Small and medium sized media are often at an impasse when it comes to long-term stability, but play a vital role in the European information space. |

| | **Scenario 1**<br>**Struggle for information supremacy** | **Scenario 2**<br>**Resilient disorder** | **Scenario 3**<br>**Global cutting edge** | **Scenario 4**<br>**Harmonic divergence** |
|---|---|---|---|---|
| **Regulation and control** | Focussed on security and economic protectionism.<br>EU is struggling to keep some degree of coherence in the European information space. Member States have introduced a patchwork of national legal frameworks.<br>The EU is trying to counter external disinformation. Due to fragmented national initiatives and low degrees of trust, these efforts are frequently undermined from within.<br>Data generated or collected by public authorities is rarely made available in a non-discriminatory way and is instead passed only to selected companies that align closely with the state. | Enforcement of open standards and profile portability have resulted in digital services in the European information space becoming increasingly decentralised. | Focussed on enforcement of data protection, competition and interoperability legislation, leaving most issues in the information space to be solved by private and civil society innovation - independent regulators play a key role.<br>Intellectual property legislation has developed advanced “fair use” clauses to facilitate sharing and open access, and a proliferation of open licenses has contributed to a general drop in piracy.<br>Strong regulatory enforcement is tightly coordinated at the European level to prevent a fragmentation of the digital single market.<br>Participation in global fora has gained increasing importance as recognition that international coordination is crucial in a borderless information space. | High EU integration on digital matters.<br>Shift from reigning in of large platforms’ market dominance towards strong oversight and moderation rules<br>Some instances of Member State governments using European legal frameworks such as counter terrorism to force removal and de-prioritisation of political content that they define as extremist raise concerns among civil society.<br>Legal obligations to counter issues such as terrorist content and disinformation on digital intermediaries have resulted in the European information space becoming increasingly separate from the global information community.<br>Internal borders for intellectual property have mostly been abolished, facilitating sharing and collaboration in Europe. |

|  | **Scenario 1**<br>**Struggle for information supremacy** | **Scenario 2**<br>**Resilient disorder** | **Scenario 3**<br>**Global cutting edge** | **Scenario 4**<br>**Harmonic divergence** |
|---|---|---|---|---|
| **Transparency and public oversight** | State entities further their security agenda through close ties with large companies and are thus unwilling and unable to limit corporate surveillance and enforce strong human rights compliance regulation for emerging technologies. | Mandatory algorithm audits by competent regulatory authorities. Big challenges to coordination on European level as Member States have passed a patchwork of national laws with different legal obligations for intermediaries. | Independent regulators with strong monitoring powers play a key role in enforcing legislation, auditing algorithmic systems for compliance with legal frameworks.<br>Increasing participation in international fora as the information space is truly global. | A network of European digital services regulators with a single board at European level now tightly monitors the status of implementation of EU rules. |

# References

[1] R. H. Coase. The market for goods and the market for ideas. *The American Economic Review*, 64:384–391, 1974.

[2] H. Farrell and B. Schneier. Common-knowledge attacks on democracy. Technical report, Berkman Klein Center for Internet & Society, 2018.

[3] A. Marwick. Silicon Valley and the social media industry. In J. Burgess, A. Marwick, and T. Poell, editors, *Sage Handbook of Social Media*, pages 314–329, Sage, 2018.

[4] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108:1378–1384, 2018.

[5] P. N. Howard, S. Woolley, and R. Calo. Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *Journal of Information Technology & Politics*, 15:81–93, 2018.

[6] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9:4787, 2018.

[7] J. M. Jachimowicz, S. Duncan, E. U. Weber, and E. J. Johnson. When and why defaults influence decisions: A meta-analysis of default effects. *Behavioural Public Policy*, 3:159–186, 2019.

[8] J. Murphy, C. Hofacker, and R. Mizerski. Primacy and recency effects on clicking behavior. *Journal of Computer-Mediated Communication*, 11:522–535, 2006.

[9] M. J. Hurlstone, S. Lewandowsky, B. R. Newell, and B. Sewell. The effect of framing and normative messages in building support for climate policies. *PLOS ONE*, 9:e114335, 2014.

[10] R. Hertwig. When to consider boosting: Some rules for policy-makers. *Behavioural Public Policy*, 1:143–161, 2017.

[11] T. Wu. *The Attention Merchants*. Atlantic Books, 2017.

[12] P. Dixit and R. Mac. How WhatsApp destroyed a village. *BuzzFeedNews*, September 9, 2018.

[13] P. de Freitas Melo, C. Coimbra Vieira, K. Garimella, P. O. S. Vaz de Melo, and F. Benevenuto. Can WhatsApp counter misinformation by limiting message forwarding? In H. Cherifi, S. Gaito, J. Fernendo Mendes, E. Moro, L. M. Rocha, editors, *Complex Networks and Their Applications VIII*, pages 372–384, Springer, 2019.

[14] J. A. Tucker, Y. Theocharis, M. E. Roberts, and P. Barberá. From liberation to turmoil: Social media and democracy. *Journal of Democracy*, 28:46–59, 2017.

[15] John T. Jost, P. Barberá, R. Bonneau, M. Langer, M. Metzger, J. Nagler, J. Sterling, and J. A. Tucker. How social media facilitates political protest: Information, motivation, and social networks. *Political Psychology*, 39:85–118, 2018.

[16] R. J. Deibert. Three painful truths about social media. *Journal of Democracy*, 30:25–39, 2019.

[17] E. Geelmuyden Rød and N. B Weidmann. Empowering activists or autocrats? The internet in authoritarian regimes. *Journal of Peace Research*, 52:338–351, 2015.

[18] H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow. The welfare effects of social media. *American Economic Review*, 110:629–676, 2020.

[19] L. Bursztyn, G. Egorov, R. Enikolopov, and M. Petrova. Social media and xenophobia: Evidence from Russia. Technical report, National Bureau of Economic Research, 2019.

[20] K. Müller and C. Schwarz. Fanning the flames of hate: Social media and hate crime. Available at SSRN: https://ssrn.com/abstract=3082972, 2020.

[21] M. Schaub and D. Morisi. Voter mobilisation in the echo chamber: Broadband internet and the rise of populism in Europe. *European Journal of Political Research*, 59: 752–773, 2020.

[22] Y. Lelkes, G. Sood, and S. Iyengar. The hostile audience: The effect of access to broadband internet on partisan affect. *American Journal of Political Science*, 61:5–20, 2017.

[23] M. Cantarella, N. Fraccaroli, and R. G. Volpe. Does fake news affect voting behaviour? Available at SSRN: https://ssrn.com/abstract=3402913, 2020.

[24] A. Kozyreva, S. Lewandowsky, and R. Hertwig. Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest*, in press, 2020.

[25] R. I. M. Dunbar, V. Arnaboldi, M. Conti, and A. Passarella. The structure of online social networks mirrors those in the offline world. *Social Networks*, 43:39–47, 2015.

[26] S. A Myers, A. Sharma, P. Gupta, and J. Lin. Information network or social network? The structure of the Twitter follow graph. *Proceedings of the 23rd International Conference on World Wide Web*, 493–498, 2014.

[27] J. Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3:205395171562251, 2016.

[28] S. Kiesler, J. Siegel, and T. W. McGuire. Social psychological aspects of computer-mediated communication. *American Psychologist*, 39:1123–1134, 1984.

[29] J. B. Walther, T. Loh, and L. Granka. Let me count the ways: The interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity. *Journal of Language and Social Psychology*, 24:36–65, 2005.

[30] L. Bareket-Bojmel and G. Shahar. Emotional and interpersonal consequences of self-disclosure in a lived, online interaction. *Journal of Social and Clinical Psychology*, 30:732–759, 2011.

[31] K. M. Christopherson. The positive and negative implications of anonymity in internet social interactions: "On the internet, nobody knows you're a dog". *Computers in Human Behavior*, 23:3038–3056, 2007.

[32] N. Lapidot-Lefler and A. Barak. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, 28:434–443, 2012.

[33] A. G. Zimmerman and G. J. Ybarra. Online aggression: The influences of anonymity and social modeling. *Psychology of Popular Media Culture*, 5:181–193, 2016.

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

[35] Z. Leviston, I. Walker, and S. Morwinski. Your opinion on climate change might not be as common as you think. *Nature Climate Change*, 3:334–337, 2013.

[36] B. Enke and F. Zimmermann. Correlation neglect in belief formation. *The Review of Economic Studies*, 86:313–332, 2019.

[37] A. N. Joinson. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, 31:177–192, 2001.

[38] A. N. Joinson. Disinhibition and the internet. In J. Gakenbach, editor, *Psychology and the Internet*, pages 75–92, Elsevier, 2007.

[39] R. E. Robertson, F. W. Tran, L. N. Lewark, and R. Epstein. Estimates of non-heterosexual prevalence: The roles of anonymity and privacy in survey methodology. *Archives of Sexual Behavior*, 47:1069–1084, 2018.

[40] A. Kozyreva, S. Herzog, P. Lorenz-Spreen, R. Hertwig, and S. Lewandowsky. Artificial intelligence in online environments: Representative survey of public attitudes in Germany. 2020.

[41] S. Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security*, 64:122–134, 2017.

[42] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus. Trolls just want to have fun. *Personality and Individual Differences*, 67:97–102, 2014.

[43] M. Kaakinen, A. Oksanen, and P. Räsänen. Did the risk of exposure to online hate increase after the November 2015 Paris attacks? A group relations approach. *Computers in Human Behavior*, 78:90–97, 2018.

[44] B. Gardiner. “It’s a terrible way to go to work:” What 70 million readers’ comments on the Guardian revealed about hostility to women and minorities online. *Feminist Media Studies*, 18:592–608, 2018.

[45] P. Rossini. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication* Research, 10.1177/0093650220921314, 2020.

[46] M. J. Crockett. Moral outrage in the digital age. *Nature Human Behaviour*, 1:769–771, 2017.

[47] E. L. Glaeser, G. A. M. Ponzetto, and J. M. Shapiro. Strategic extremism: Why Republicans and Democrats divide on religious values. *The Quarterly Journal of Economics*, 120:1283–1330, 2005.

[48] C. McVittie and A. McKinlay. ‘Alternative facts are not facts’: Gaffe-announcements, the Trump administration and the media. *Discourse & Society*, 30:172–187*,* 2019.

[49] S. Lewandowsky. Wilful construction of ignorance: A tale of two ontologies. In R. Hertwig and C. Engel, editors, *Deliberate Ignorance: Choosing Not to Know*, pages 101–117, MIT Press, 2020.

[50] H. A Simon. Designing organizations for an information-rich world. *Computers, Communications and the Public Interest*, 70:37–72, 1971.

[51] D. Meshi, D. I. Tamir, and H. R. Heekeren. The emerging neuroscience of social media. *Trends in Cognitive Sciences*, 19:771–782, 2015.

[52] T. T. Hills. The dark side of information proliferation. *Perspectives on Psychological Science*, 14:323–330, 2019.

[53] C. R. Payne, M. Niculescu, D. R. Just, and M. P. Kelly. Shopper marketing nutrition interventions. *Physiology & Behavior*, 136:111–120, 2014.

[54] J. Hartmann-Boyce, F. Bianchi, C. Piernas, S. Payne Riches, K. Frie, R. Nourse, and S. A Jebb. Grocery store interventions to change food purchasing behaviors: A systematic review of randomized controlled trials. *The American Journal of Clinical Nutrition*, 107:1004–1016, 2018.

[55] D. Susser, B. Roessler, and H. Nissenbaum. Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review*, 2019.

[56] B. J. Fogg. Persuasive technology. *Ubiquity*, 2002:89–120, 2002.

[57] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4:eaao5580, 2018.

[58] B. Christian and T. Griffiths. *Algorithms to live by: The computer science of human decisions*. Macmillan, 2016.

[59] P. Mozur. A genocide incited on Facebook, with posts from Myanmar's military. *The New York Times*, October 15, 2018.

[60] K. Schaeffer. Nearly three-in-ten Americans believe COVID-19 was made in a lab. *Pew Research Center*, April 8, 2020.

[61] S. Lewandowsky and J. Cook. Coronavirus conspiracy theories are dangerous—here's how to stop them spreading. *The Conversation,* April 20, 2020.

[62] D. Freeman, F. Waite, L. Rosebrock, A. Petit, C. Causier, A. East, L. Jenner, A.-L. Teale, L. Carr, S. Mulhall, E. Bold, and S. Lambe. Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. *Psychological* Medicine, 10.1017/S0033291720001890, 2020.

[63] D. Allington, B. Duffy, S. Wessely, N. Dhavan, and J. Rubin. Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. *Psychological Medicine*, 10.1017/S003329172000224X, 2020.

[64] S. Khazaeli and D. Stockemer. The internet: A new route to good governance. *International Political Science Review*, 34:463–482, 2013.

[65] P. Lorenz-Spreen, B. Mørch Mønsted, P. Hövel, and S. Lehmann. Accelerating dynamics of collective attention. *Nature Communications*, 10:1759, 2019.

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

[67] P. Covington, J. Adams, and E. Sargin. Deep neural networks for YouTube recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems—RecSys '16*, 2016.

[68] J. B Schmitt, D. Rieger, O. Rutkowski, and J. Ernst. Counter-messages as prevention or promotion of extremism?! The potential role of YouTube recommendation algorithms. *Journal of Communication*, 68:780–808, 2018.

[69] L. Spinelli and M. Crovella. How YouTube leads privacy-seeking users away from reliable information. *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization,* 244–251. ACM, 2020.

[70] W. Youyou, M. Kosinski, and D. Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112:1036–1040, 2015.

[71] J. M. Balkin. The constitution in the national surveillance state. *Minnesota Law Review*, 93:1–25, 2008.

[72] S. Zuboff. Surveillance capitalism and the challenge of collective action. *New Labor Forum*, 28:10–29, 2019.

[73] B. Lindström, M. Bellander, A. Chang, P. N Tobler, and D. M Amodio. A computational reinforcement learning account of social media engagement. *PsyArXiv.* https://psyarxiv.com/78mh5/, 2019.

[74] E. Reynolds. Has Tinder lost its spark?, *The Guardian*, August 11, 2019.

[75] J. Heawood. Pseudo-public political speech: Democratic implications of the Cambridge Analytica scandal. *Information Polity*, 23:429–434, 2018.

[76] S. Machkovech. Report: Facebook helped advertisers target teens who feel “worthless”. *arsTechnica*, May 1, 2017.

[77] 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*, 110:5802–5805, 2013.

[78] J. Hinds and A. N. Joinson. What demographic attributes do our digital footprints reveal? A systematic review. *PLOS ONE*, 13:e0207112, 2018.

[79] J. Hinds and A. N. Joinson. Human and computer personality prediction from digital footprints. *Current Directions in Psychological Science*, 28:204–211, 2019.

[80] Á. Cuevas, J. González Cabañas, A. Arrate, and R. Cuevas. Does Facebook use sensitive data for advertising purposes? Worldwide analysis and GDPR impact. *arXiv:1907.10672*, 2019.

[81] K. Kalimeri, M. G. Beiró, M. Delfino, R. Raleigh, and C. Cattuto. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior*, 92:428–445, 2019.

[82] R. L. Boyd, P. Pasca, and K. Lanning. The personality panorama: Conceptualizing personality through big behavioural data. *European Journal of Personality*, 10.1002/per.2254, 2020.

[83] J. Kokott and C. Sobotta. The distinction between privacy and data protection in the jurisprudence of the CJEU and the ECtHR. *International Data Privacy Law*, 3:222–228, 2013.

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

[85] J. A. T. Fairfield and C. Engel. Privacy as a public good. *Duke Law Journal*, 65:385–457, 2015.

[86] P. M. Regan. Response to privacy as a public good. *Duke Law Journal*, 65:51–65, 2016.

[87] Z. Tufekci. The latest data privacy debacle. *The New York Times*, January 30, 2018.

[88] S. Das and A. Kramer. Self-censorship on Facebook. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 120–127, 2013.

[89] C. Duhigg. How companies learn your secrets *The New York Times*, February 16, 2012.

[90] A. E Marwick and D. Boyd. Networked privacy: How teenagers negotiate context in social media. *New Media & Society*, 16:1051–1067, 2014.

[91] K. Knibbs. What's a Facebook shadow profile and why should you care. *Digital Trends*, July 5, 2013.

[92] D. Garcia. Leaking privacy and shadow profiles in online social networks. *Science Advances*, 3:e1701172, 2017.

[93] E.-Á. Horvát, M. Hanselmann, F. A. Hamprecht, and K. A. Zweig. One plus one makes three (for social networks). *PLOS ONE*, 7:e0034740, 2012.

[94] D. Garcia, M. Goel, A. K. Agrawal, and P. Kumaraguru. Collective aspects of privacy in the Twitter social network. *EPJ Data Science*, 7:3, 2018.

[95] D. Garcia. Privacy beyond the individual. *Nature Human Behaviour*, 3:112–113, 2019.

[96] D. Jurgens, Y. Tsvetkov, and D. Jurafsky. Writer profiling without the writer's text. In G. L. Ciampaglia, A. Mashhadi, and T. Yasseri, editors, *SocInfo: International Conference on Social Informatics*, pages 537–558. Springer, 2017.

[97] J. P. Bagrow, X. Liu, and L. Mitchell. Information flow reveals prediction limits in online social activity. *Nature Human Behaviour*, 3:122–128, 2019.

[98] P. Lorenz-Spreen, S. Lewandowsky, C. R. Sunstein, and R. Hertwig. How behavioural sciences can promote truth and, autonomy and democratic discourse online. *Nature Human Behaviour*, 10.1038/s41562-020-0889-7, 2020.

[99] M. Leiser. Regulating computational propaganda: Lessons from international law. *Cambridge International Law Journal*, 8:218–240, 2019.

[100] B. Bodó, N. Helberger, and C. H. de Vreese. Political micro-targeting: A Manchurian candidate or just a dark horse? *Internet Policy Review*, 6, 2017.

[101] F. J. Zuiderveen Borgesius, J. Moller, S. Kruikemeier, R. Ó. Fathaigh, K. Irion, T.Dobber, B. Bodo, and C. de Vreese. Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, 14:82–96, 2018.

[102] I. Faizullabhoy and A. Korolova. Facebook's advertising platform: New attack vectors and the need for interventions. *arXiv:*1803.10099, 2018.

[103] C. O'Connor and J. O. Weatherall. *The Misinformation Age: How False Beliefs Spread.* Yale University Press, 2019.

[104] M. Ali, P. Sapiezynski, A. Korolova, A. Mislove, and A. Rieke. Ad delivery algorithms: The hidden arbiters of political messaging. Technical report, *arXiv*:1912.04255v3, 2019.

[105] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke. Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 3:1–30, 2019.

[106] T. N. Ridout, E. Franklin Fowler, M. M. Franz, and K. Goldstein. The long-term and geographically constrained effects of campaign advertising on political polarization and sorting. *American Politics Research*, 46:3–25, 2018.

[107] K. H. Jamieson. *Cyberwar*. Oxford University Press, 2018.

[108] J. Horwitz and D. Seetharaman. Facebook executives shut down efforts to make the site less divisive. *The Wall Street Journal*,May 26, 2020.

[109] K. Dommett and S. Power. The political economy of Facebook advertising: Election spending, regulation and targeting online. *The Political Quarterly*, 90:257–265, 2019.

[110] J. B. Hirsh, S. K. Kang, and G. V. Bodenhausen. Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. *Psychological Science*, 23:578–581, 2012.

[111] 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*, 48:12714–12719, 2017.

[112] D. Eckles, B. R. Gordon, and G. A. Johnson. Field studies of psychologically targeted ads face threats to internal validity. *Proceedings of the National Academy of Sciences*, 115:E5254–E5255, 2018.

[113] B. Sharp, N. Danenberg, and S. Bellman. Psychological targeting. *Proceedings of the National Academy of Sciences*, 115:E7890–E7890, 2018.

[114] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. Reply to Sharp et al.: Psychological targeting produces robust effects. *Proceedings of the National Academy of Sciences*, 115:E7891–E7891, 2018.

[115] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. Reply to Eckles et al.: Facebook's optimization algorithms are highly unlikely to explain the effects of psychological targeting. *Proceedings of the National Academy of Sciences*, 115:E5256–E5257, 2018.

[116] C. A. Bail. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences*, 113:11823–11828, 2016.

[117] A. E. Latimer, N. A. Katulak, L. Mowad, and P. Salovey. Motivating cancer prevention and early detection behaviors using psychologically tailored messages. *Journal of Health Communication*, 10:137–155, 2005.

[118] W. Willett, J. Rockström, B. Loken, M. Springmann, T. Lang, S. Vermeulen, T. Garnett, D. Tilman, F. DeClerck, A. Wood, et al. Food in the Anthropocene: The EAT–Lancet Commission on healthy diets from sustainable food systems. *The Lancet*, 393:447–492, 2019.

[119] D. Garcia, V. Galaz, and S. Daume. EATLancet vs yes2meat: The digital backlash to the planetary health diet. *The Lancet*, 394:2153–2154, 2019.

[120] M. Eslami, A. Rickman, K. Vaccaro, A. Aleyasen, A. Vuong, K. Karahalios, K. Hamilton, and C. Sandvig. "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 153–162, 2015.

[121] T. Dienlin and M. J. Metzger. An extended privacy calculus model for SNSs: Analyzing self-disclosure and self-withdrawal in a representative U.S. sample. *Journal of Computer-Mediated Communication*, 21:368–383, 2016.

[122] E. J. Johnson, S. Bellman, and G. L. Lohse. Defaults, framing and privacy: Why opting in-opting out. *Marketing Letters*, 13:5–15, 2002.

[123] I. Lapowsky. Your old tweets give away more location data than you think. *Wired*, January 10, 2019.

[124] K. Kinder-Kurlanda, K. Weller, W. Zenk-Möltgen, J. Pfeffer, and F. Morstatter. Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society*, 4, 2017.

[125] A. Khalid. Twitter removes precise geo-tagging option from tweets. *The Verge*, June 19, 2019.

[126] Norwegian Consumer Council. *Deceived by Design*, 2018. Available from https://fil.forbrukerradet.no/wp-content/uploads/2018/06/2018-06-27-deceived-by-design-final.pdf

[127] F. Marotta-Wurgler. Does "notice and choice" disclosure regulation work? An empirical study of privacy policies. Technical report, University of Michigan, 2019.

[128] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan. Dark patterns at scale. *Proceedings of the ACM on Human-Computer Interaction*, 3:1–32, 2019.

[129] M. R. Leiser. 'Dark patterns': The case for regulatory pluralism. *Manuscript submitted for publication*, 2020.

[130] M. R. Leiser. Dark patterns: Light to be found in Europe's consumer protection regime. *Manuscript submitted for publication*, 2020.

[131] B. Henne, M. Koch, and M. Smith. On the awareness, control and privacy of shared photo metadata. *International Conference on Financial Cryptography and Data Security*, 77–88. Springer, 2014.

[132] B. Henne and M. Smith. Awareness about photos on the web and how privacy-privacy-tradeoffs could help. *International Conference on Financial Cryptography and Data Security*, 131–148. Springer, 2013.

[133] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020.

[134] F. Ricci, L. Rokach, and B. Shapira. *Recommender Systems: Introduction and Challenges*. Springer, 2015.

[135] F. Pasquale. *The Black Box Society*. Harvard University Press, 2015.

[136] P. B. de Laat. Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & Technology*, 31:525–541, 2018.

[137] N. Diakopoulos. Algorithmic accountability. *Digital Journalism*, 3:398–415, 2015.

[138] S. Garfinkel, J. Matthews, S. S. Shapiro, and J. M. Smith. Toward algorithmic transparency and accountability. *Communications of the ACM*, 60:5–5, 2017.

[139] A. Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6:175–183, 2004.

[140] M. Keller. The Apple ‘kill list’: What your iPhone doesn’t want you to type. *Daily Beast*, July 11, 2013.

[141] J. Valentino-DeVries, J. Singer-Vine, and A. Soltani. Websites vary prices, deals based on users’ information. *The Wall Street Journal*, December 24, 2012.

[142] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015:92–112, 2015.

[143] A. Lambrecht and C. Tucker. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65:2966–2981, 2019.

[144] L. Sweeney. Discrimination in online ad delivery. *Queue*, 11:1–19, 2013.

[145] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453, 2019.

[146] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

[147] S. Lawrence. Searching the World Wide Web. *Science*, 280:98–100, 1998.

[148] A. Mowshowitz and A. Kawaguchi. Assessing bias in search engines. *Information Processing & Management*, 38:141–156, 2002.

[149] A. Mowshowitz and A. Kawaguchi. Measuring search engine bias. *Information Processing & Management*, 41:1193–1205, 2005.

[150] E. Goldman. Search engine bias and the demise of search engine utopianism. In J. Mackenzie Owen, A. Spink, and M. Zimmer, editors, *Web Search*, volume 14, pages 121–133, Springer, 2008.

[151] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Topical interests and the mitigation of search engine bias. *Proceedings of the National Academy of Sciences*, 103:12684–12689, 2006.

[152] F. Tripodi. Searching for alternative facts: Analyzing scriptural inference in conservative news practices. Technical report, Data & Society, 2018.

[153] M. Golebiewski and D Boyd. Data voids: Where missing data can easily be exploited. Technical report, Data & Society, 2019.

[154] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry. Preconference at the 64th Annual Meeting of the International Communication Association*, 2014.

[155] R. E. Robertson, S. Jiang, D. Lazer, and C. Wilson. Auditing autocomplete: Suggestion networks and recursive algorithm interrogation. *Proceedings of the 10th ACM Conference on Web Science—WebSci '19*, 235–244. ACM Press, 2019.

[156] A. Hannak, P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson. Measuring personalization of web search. *Proceedings of the 22nd International Conference on World Wide Web—WWW '13*, 527–538. ACM Press, 2013.

[157] C. Kliman-Silver, A. Hannak, D. Lazer, C. Wilson, and A. Mislove. Location, location, location: The impact of geolocation on web search personalization. *Proceedings of the 2015 ACM Conference on Internet Measurement Conference—IMC '15*, 121–127. ACM Press, 2015.

[158] R. E. Robertson, D. Lazer, and C. Wilson. Auditing the personalization and composition of politically-related search engine results pages. *Proceedings of the 2018 World Wide Web Conference on World Wide Web—WWW '18*, 955–965. ACM Press, 2018.

[159] D. Trielli and N. Diakopoulos. Search as news curator: The role of Google in shaping attention to news information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems—CHI '19*, 1–15, ACM Press, 2019.

[160] A. Kawakami, K. Umarova, and E. Mustafaraj. The media coverage of the 2020 US Presidential election candidates through the lens of Google's top stories. *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 868–877. AAAI Press, 2020.

[161] R. E. Robertson, S. Jiang, K. Joseph, L. Friedland, D. Lazer, and C. Wilson. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, 2:1–22, 2018.

[162] D. Hu, S. Jiang, R. E. Robertson, and C. Wilson. Auditing the partisanship of Google Search snippets. *The World Wide Web Conference on WWW '19*, 693–704. ACM Press, 2019.

[163] E. Mustafaraj, E. Lurie, and C. Devine. The case for voter-centered audits of search engines during political elections. *Proceedings of the 202 Conference on Fairness, Accountability, and Transparency*, 559–569, 2020.

[164] D. Trielli and N. Diakopoulos. Partisan search behavior and Google results in the 2018 U.S. midterm elections. *Information, Communication & Society*, 10.1080/1369118X.2020.1764, 2020.

[165] S. Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, 2018.

[166] P. Baker and A. Potts. 'Why do white people have thin lips?' Google and the perpetuation of stereotypes via auto-complete search forms. *Critical Discourse Studies*, 10:187–204, 2013.

[167] P. Takis Metaxas. Web spam, social propaganda and the evolution of search engine rankings. *International Conference on Web Information Systems and Technologies*, 170–182. Springer, 2009.

[168] P. Takis Metaxas and J. DeStefano. Web spam, propaganda and trust. Adversarial Information Retrieval (AIRWeb), WWW 2005 Conference, Chiba, Japan.

[169] P. Takis Metaxas and Y. Pruksachatkun. Manipulation of search engine results during the 2016 US congressional elections. Proceedings of the ICIW 2017, Venice, Italy, 2017.

[170] D. Metaxa, J. S. Park, J. A. Landay, and J. Hancock. Search media and elections. *Proceedings of the ACM on Human-Computer Interaction*, 3:1–17, 2019.

[171] Z. Tufekci. YouTube, the great radicalizer. *The New York Times*, March 10, 2018.

[172] M. Alfano, A. E. Fard, J. A. Carter, P. Clutton, and C. Klein. Technologically scaffolded atypical cognition: The case of YouTube's recommender system. *Synthese*, 2020.

[173] K. Papadamou, A. Papasavva, S. Zannettou, J. Blackburn, N. Kourtellis, I. Leontiadis, G. Stringhini, and M. Sirivianos. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 522–533. AAAI, 2020.

[174] J. Kaiser and A. Rauchfleisch. Unite the right? How YouTube's recommendation algorithm connects the U.S. far-right. *Medium*, April 11, 2018.

[175] A. Rauchfleisch and J. Kaiser. YouTubes Algorithmen sorgen dafür, dass AfD-Fans unter sich bleiben. *Vice*, September 22, 2017.

[176] M. Horta Ribeiro, R. Ottoni, R. West, V. A. F. Almeida, and W. Meira. Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. ACM, 2020.

[177] K. Munger and J. Phillips. A supply and demand framework for YouTube politics. Available at https://osf.io/73jys/, 2019.

[178] R. Lewis. Alternative influence: Broadcasting the reactionary right on YouTube. Technical report, Data & Society, 2018.

[179] J. C. Wong and S. Levin. Youtube vows to recommend fewer conspiracy theory videos. The Guardian January 25, 2019.

[180] M. H. Raab, N. Auer, S. A. Ortlieb, and C.-C. Carbon. The Sarrazin effect: The presence of absurd statements in conspiracy theories makes canonical information less plausible. *Frontiers in Psychology*, 4:453, 2013.

[181] Global Web Index. GWI Coronavirus research: Multi-market research wave 5. Technical report, Global Web Index, 2020.

[182] T. Harris. How technology is hijacking your mind—from a magician and Google design ethicist. *Medium*, May 18, 2016.

[183] P. Lorenz-Spreen, F. Wolf, J. Braun, G. Ghoshal, N. Djurdjevac Conrad, and P. Hövel. Tracking online topics over time: Understanding dynamic hashtag communities. *Computational Social Networks*, 5:9, 2018.

[184] C. R. Sunstein. *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press, 2017.

[185] E. Pariser. *The Filter Bubble: What the Internet is Hiding From You*. Penguin Press, 2011.

[186] E. Bakshy, S. Messing, and L. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348:1130–1132, 2015.

[187] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. B. Fallin Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115:9216–9221, 2018.

[188] J. Bright, N. Marchal, B. Ganesh, and S. Rudinac. Echo chambers exist! (But they're full of opposing views). *arXiv:2001.11461*, 2020.

[189] C. O'Connor and J. O. Weatherall. Scientific polarization. *European Journal for Philosophy of Science*, 8:855–875, 2018.

[190] J. O. Weatherall and C. O'Connor. Endogenous epistemic factionalization. *Synthese*, 10.1007/s11229-020-02675-3, 2020.

[191] B. Nyhan. Does the US media have a liberal bias? A discussion of Tim Groseclose's "Left turn: How liberal media bias distorts the American mind". *Perspectives on Politics*, 10:767–771, 2012.

[192] K. Chadha and R. Wells. Journalistic responses to technological innovation in newsrooms: An exploratory study of Twitter use. *Digital Journalism*, 4:1020–1035, 2013.

[193] L. Vaughan and M. Thelwall. Search engine coverage bias: Evidence and possible causes. *Information Processing & Management*, 40:693–707, 2004.

[194] N. Diakopoulos, D. Trielli, J. Stark, and S. Mussenden. I vote for—How search informs our choice of candidate. In M. Moore and D. Tambini, editors, *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, Oxford University Press, 2018.

[195] J. Kulshrestha, M. Eslami, J. Messias, M. Bilal Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios. Search bias quantification: Investigating political bias in social media and web search. *Information Retrieval Journal*, 22:188–227, 2019.

[196] E. Borra and I. Weber. Political insights: Exploring partisanship in Web search queries. *First Monday*, 17, July 2012.

[197] N. J. Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information and Library Sciences*, 5:133–143, 1980.

[198] K. Munger. The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media + Society*, 5, 2019.

[199] L. Boxell, M. Gentzkow, and J. M Shapiro. Cross-country trends in affective polarization. *NBER Working Papers 26669.* National Bureau of Economic Research, 2020.

[200] A. Guess, B. Nyhan, B. Lyons, and J. Reifler. Avoiding the echo chamber about echo chambers. Technical report, Knight Foundation, 2018.

[201] J. Tucker, A. Guess, P. Barbera, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan. Social media, political polarization, and political disinformation: A review of the scientific literature. Technical report, Hewlett Foundation, 2018.

[202] L. Boxell, M. Gentzkow, and J. M. Shapiro. Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences*, 114:10612–10617, 2017.

[203] S. Flaxman, S. Goel, and J. M. Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80:298–320, 2016.

[204] P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26:1531–1542, 2015.

[205] L. A. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. *Proceedings of the 3rd International Workshop on Link Discovery*, 36–43. ACM, 2005.

[206] Y. Lelkes. Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, 80:392–410, 2016.

[207] E. Suhay, E. Bello-Pardo, and B. Maurer. The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics*, 23:95–115, 2018.

[208] R. Kelly Garrett and N. Jomini Stroud. Partisan paths to exposure diversity: Differences in pro- and counter-attitudinal news consumption. *Journal of Communication*, 64:680–701, 2014.

[209] A. M. Guess. (Almost) everything in moderation: New evidence on Americans' online media diets. *American Journal of Political Science*, in press, 2020.

[210] A. M. Guess, B. Nyhan, and J. Reifler. Exposure to untrustworthy websites in the 2016 U.S. election. *Nature Human Behavior*, 4:472–480, 2020.

[211] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363:374–378, 2019.

[212] A. Allam, P. J. Schulz, and K. Nakamoto. The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: Two experiments manipulating google output. *Journal of Medical Internet Research*, 16:e100, 2014.

[213] A. Novin and E. Meyers. Making sense of conflicting science information: Exploring bias in the search engine result page. *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval*, 175–184. ACM, 2017.

[214] R. Epstein, R. E. Robertson, D. Lazer, and C. Wilson. Suppressing the search engine manipulation effect (SEME). *Proceedings ACM Human-Computer Interactions*, 1:42:1–42:22, 2017.

[215] Directorate-General for Communication. *Flash Eurobarometer 464: Fake news and disinformation online*. Available at https://data.europa.eu/euodp/en/data/dataset/S2183_464_ENG, 2018

[216] A. Mitchell, J. Gottfried, G. Stocking, M. Walker, and S. Fedeli. Many Americans say made-up news is a critical problem that needs to be fixed. *Pew Research Center,* June 5, 2019.

[217] J. Allen, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6:eaay3539, 2020.

[218] A. M. Guess, J. Nagler, and J. Tucker. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5:eaau4586, 2019.

[219] A. M. Guess, D. Lockett, B. Lyons, J. M. Montgomery, B. Nyhan, and J. Reifler. "Fake news" may have limited effects on political participation beyond increasing beliefs in false claims. *Harvard Kennedy School Misinformation Review*, January 14, 2020.

[220] C. J. Vargo, L. Guo, and M. A. Amazeen. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, 20:2028–2049, 2018.

[221] Y. Benkler, R. Faris, H. Roberts, and E. Zuckerman. Study: Breitbart-led right-wing media ecosystem altered broader media agenda. *Columbia Journalism Review*, March 3, 2017.

[222] E. Nechushtai. From liberal to polarized liberal? Contemporary U.S. news in Hallin and Mancini's typology of news systems. *The International Journal of Press/Politics*, 23:183–201, 2018.

[223] A. Downs. An economic theory of political action in a democracy. *Journal of Political Economy*, 65:135–150, 1957.

[224] C. Wardle and H. Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policymaking. Technical report, Council of Europe, 2017.

[225] *Global Disinformation Index*. Available at https://disinformationindex.org/, 2019.

[226] E. C. Tandoc, Z. W. Lim, and R. Ling. Defining "fake news". *Digital Journalism*, 6:137–153, 2018.

[227] V. S. Greene. "Deplorable" satire: Alt-right memes, white genocide tweets, and redpilling normies. *Studies in American Humor*, 5:31–69, 2019.

[228] A. Marwick and R. Lewis. Media manipulation and dis-information online. Technical report, Data & Society, 2017.

[229] S. Altay, E. de Araujo, and H. Mercier. "If this account is true, it is most enormously wonderful": Interestingness-if-true and the sharing of true and false news. *PsyArXiv.* https://psyarxiv.com/tdfh5/, 2020.

[230] B. Lyons, V. Merola, and J. Reifler. Not just asking questions: Effects of implicit and explicit conspiracy information about vaccines and genetic modification. *Health Communication*, 34:1741–1750, 2018.

[231] D. Michaels. *Doubt is Their Product: How Industry's Assault on Science Threatens Your Health*. Oxford University Press, 2008.

[232] N. Oreskes and E. M. Conway. *Merchants of Doubt*. Bloomsbury Publishing, 2010.

[233] J. O. Weatherall, C. O'Connor, and J. P. Bruner. How to beat science and influence people: Policy-makers and propaganda in epistemic networks. *The British Journal for the Philosophy of Science*, axy062, 2018.

[234] C. O'Connor and J. O. Weatherall. Why false claims about COVID-19 refuse to die. *Nautilus,* April 16, 2020.

[235] A. M. McCright and R. E. Dunlap. Combatting misinformation requires recognizing its types and the factors that facilitate its spread and resonance. *Journal of Applied Research in Memory and Cognition*, 6:353–512, 2017.

[236] E. Herring and P. Robinson. Report X marks the spot: The British Government's deceptive dossier on Iraq and WMD. *Political Science Quarterly*, 129:551–584, 2014.

[237] E. Herring and P. Robinson. Deception and Britain's road to war in Iraq. *International Journal of Contemporary Iraqi Studies*, 8:213–232, 2014.

[238] O. D. Thomas. Good faith and (dis)honest mistakes? Learning from Britain's Iraq War inquiry. *Politics*, 37:371–385, 2017.

[239] S. Lewandowsky and J. Lynam. Combating 'fake news': The 21st century civic duty. *The Irish Times,* December 29, 2018.

[240] H. A. Giroux. Trump and the legacy of a menacing past. *Cultural Studies*, 33:711–739, 2018.

[241] J. M. Miller. Do COVID-19 conspiracy theory beliefs form a monological belief system? *Canadian Journal of Political Science*, 53:319–326, 2020.

[242] C. Paul and M. Matthews. The Russian “firehose of falsehood” propaganda model. *RAND Corporation*, September 29, 2016.

[243] A. S. Ross and D. J. Rivers. Discursive deflection: Accusation of “fake news” and the spread of mis- and disinformation in the Tweets of President Trump. *Social Media + Society*, 4, 2018.

[244] S. Waisbord. Why populism is troubling for democratic communication. *Communication, Culture and Critique*, 11:21–34, 2018.

[245] C. M. Corbin. Trump’s lies: The unconstitutionality of government propaganda. *Ohio State Law Journal*, in press, 2020.

[246] S. Lewandowsky, U. K. H. Ecker, and J. Cook. Beyond misinformation: Understanding and coping with the post-truth era. *Journal of Applied Research in Memory and Cognition*, 6:353–369, 2017.

[247] M. Judge. Garry Kasparov on the press and propaganda in Trump’s America. *Columbia Journalism Review*, March 22, 2017.

[248] S. Soroka, P. Fournier, and L. Nir. Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, 116:18888–18892, 2019.

[249] J. Berger and K. L. Milkman. What makes online content viral? *Journal of Marketing Research*, 49:192–205, 2012.

[250] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114:7313–7318, 2017.

[251] J. Paschen. Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal of Product & Brand Management*, 29:223–233, 2019.

[252] N. Bunzeck and E. Düzel. Absolute coding of stimulus novelty in the human substantia nigra/VTA. *Neuron*, 51:369–379, 2006.

[253] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359:1146–1151, 2018.

[254] E. Fenn, N. Ramsa, J. Kantner, K. Pezdek, and E. Abed. Non-probative photos increase truth, like, and share judgments in a simulated social media environment. *Journal of Applied Research in Memory and Cognition*, 8:131–138, 2019.

[255] E. Fenn, E. J. Newman, K. Pezdek, and M. Garry. The effect of non-probative photographs on truthiness persists over time. *Acta Psychologica*, 144:207–211, 2013.

[256] L. Fazio. Out-of-context photos are a powerful, low tech form of misinformation. *The Conversation,* February 14, 2020.

[257] S. Lewandowsky and L. Whitmarsh. Climate communication for biologists: When a picture can tell a thousand words. *PLOS Biology*, 16:e2006004, 2018.

[258] M. Smith-Rodden and I. K. Ash. Investigating the psychological effects of news imagery: A case for evidence-based decision making and practices. *Visual Communication Quarterly*, 19:20–32, 2012.

[259] G. Asmolov. The disconnective power of disinformation campaigns. *Journal of International Affairs*, 71:69–76, 2018.

[260] F. Giglietto, L. Iannelli, A. Valeriani, and L. Rossi. 'Fake news' is the invention of a liar: How false information circulates within the hybrid news system. *Current Sociology*, 67:625–642, 2019.

[261] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[262] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68:036122, 2003.

[263] M. Bossetta. The digital architectures of social media: Comparing political campaigning on Facebook, Twitter, Instagram and Snapchat in the 2016 U.S. election. *Journalism & Mass Communication Quarterly,* 95:471–496, 2018.

[264] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh. WTF: The Who to Follow service at Twitter. *Proceedings of the 22nd International Conference on World Wide Web—WWW '13.* ACM, 2013.

[265] B. Gonçalves, N. Perra, and A. Vespignani. Modeling users' activity on Twitter networks: Validation of Dunbar's number. *PLOS ONE*, 6:e22656, 2011.

[266] S.-Y. Kwak, S.-H. Yoo, and S.-J. Kwak. Valuing energy-saving measures in residential buildings: A choice experiment study. *Energy Policy*, 38:673–677, 2010.

[267] W. Lance Bennett and J. B. Manheim. The one-step flow of communication. *The ANNALS of the American Academy of Political and Social Science*, 608:213–232, 2006.

[268] S. S. Myers, A. Zanobetti, I. Kloog, P. Huybers, A. D. B. Leakey, A. J. Bloom, E. Carlisle, L. H. Dietterich, T. Fitzgerald, G. Hasegawa, N. M. Holbrook, R. L. Nelson, M. J. Ottman, V. Raboy, H. Sakai, K. A. Sartor, J. Schwartz, S. Seneweera, M. Tandsz, and Y. Usui. Increasing $CO_2$ threatens human nutrition. *Nature*, 510:139–142, 2014.

[269] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, 591–600. ACM, 2010.

[270] R.Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:3200–3203, 2001.

[271] B. Mønsted, P. Sapieżyński, E. Ferrara, and S. Lehmann. Evidence of complex contagion of information in social media: An experiment using twitter bots. *PLOS ONE*, 12:e0184148, 2017.

[272] D. J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99:5766–5771, 2002.

[273] J. Wise. How Donald Trump's chloroquine campaign started, *Vanity Fair*, March 24, 2020.

[274] J. Scott Brennen, F. M. Simon, P. N. Howard, and R. Kleis Nielsen. Types, sources, and claims of COVID-19 misinformation. Technical report, Reuters Institute for the Study of Journalism, 2020.

[275] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the Facebook social graph. *arXiv:1111.4503*, 2011.

[276] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web*, 6:9, 2012.

[277] X. Wang, C. Yu, and Y. Wei. Social media peer communication and impacts on purchase intentions: A consumer socialization framework. *Journal of Interactive Marketing*, 26(4):198–208, 2012.

[278] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on Twitter. *Proceedings of the Fifth International AAAI Conference on weblogs and social media*, 89–96. AAAI, 2011.

[279] D. Centola and M. Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113:702–734, 2007.

[280] A. Satariano. Facebook identifies Russia-linked misinformation campaign. *The New York Times*, January 17, 2019.

[281] E. Colleoni, A. Rozza, and A. Arvidsson. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, 64:317–332, 2014.

[282] F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 124:048301, 2020.

[283] G. Resende, P. Melo, H. Sousa, J. Messias, M. Vasconcelos, J. Almeida, and F. Benevenuto. (Mis)information dissemination in WhatsApp: Gathering, analyzing and countermeasures. *The World Wide Web Conference*, 818–828. ACM, 2019.

[284] D. J. Watts and S. H. Strogatz. Collective dynamics of “small-world” networks. *Nature*, 393:440–442, 1998.

[285] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. *Proceedings of the 4th Annual ACM Web Science Conference*, 33–42. ACM, 2012.

[286] J. Wohlgemuth and M. T. Matache. Small-world properties of Facebook group networks. *Complex Systems*, 23:197–226, 2014.

[287] B. F. Schaffner and S. Luks. Misinformation or expressive responding? What an inauguration crowd can tell us about the source of political misinformation in surveys. *Public Opinion Quarterly,* 82:135–147, 2018.

[288] S. van Kessel, J. Sajuria, and S. M. Van Hauwaert. Informed, uninformed or misinformed? A cross-national analysis of populist party supporters across European democracies. *West European Politics*, 10.1080/01402382.2019.1700448, 2020.

[289] J.-W. van Prooijen and A. P. M. Krouwel. Overclaiming knowledge predicts anti-establishment voting. *Social Psychological and Personality Science*, 11:356–363, 2019.

[290] E. Humprecht, F. Esser, and P. Van Aelst. Resilience to online disinformation: A framework for cross-national comparative research. *The International Journal of Press/Politics*, 25:493–516, 2020.

[291] N. Newman, R. Fletcher, A. Kalogeropoulos, D. A. L. Levy, and R. Kleis Nielsen. Reuters Institute digital news report 2018. Technical report, Reuters Institute for the Study of Journalism, 2018.

[292] J. T. Jost. Ideological asymmetries and the essence of political psychology. *Political Psychology*, 38:167–208, 2017.

[293] J. T. Jost, J. Glaser, A. W. Kruglanski, and F. J. Sulloway. Political conservatism as motivated social cognition. *Psychological Bulletin*, 129:339–375, 2003.

[294] A. I. Abramowitz and S. Webster. The rise of negative partisanship and the nationalization of U.S. elections in the 21st century. *Electoral Studies*, 41:12–22, 2016.

[295] S. Iyengar and S. J. Westwood. Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59:690–707, 2015.

[296] K. Ognyanova, D. Lazer, R. E. Robertson, and C. Wilson. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*, June 2, 2020.

[297] G. Pennycook, J. A. Cheyne, N. Barr, D. J. Koehler, and J. A. Fugelsang. On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, 10:549–563, 2015.

[298] J. Sterling, J. T. Jost, and G. Pennycook. Are neoliberals more susceptible to bullshit? *Judgment and Decision Making*, 11:352–360, 2016.

[299] S. Pfattheicher and S. Schindler. Misperceiving bullshit as profound is associated with favorable views of Cruz, Rubio, Trump and conservatism. *PLoS ONE*, 11:e0153419, 2016.

[300] D. M. T. Fessler, A. C. Pisor, and C. Holbrook. Political orientation predicts credulity regarding putative hazards. *Psychological Science*, 28:651–660, 2017.

[301] J. R. Hibbing, K. B. Smith, and J. R. Alford. Differences in negativity bias underlie variations in political ideology. *Behavioral and Brain Sciences*, 37:297–350, 2014.

[302] M. V. Bronstein, G. Pennycook, A. Bear, D. G. Rand, and T. D. Cannon. Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, 8:108–117, 2018.

[303] M. Meyer. Fake news, conspiracy, and intellectual vice. *Social Epistemology Review and Reply Collective*, 8:9–19, 2019.

[304] G. Pennycook and D. G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, 2019.

[305] R. Kelly Garrett and B. E. Weeks. Epistemic beliefs' role in promoting misperceptions and conspiracist ideation. *PLOS ONE*, 12:e0184733, 2017.

[306] P. Bae Brandtzaeg and A. Følstad. Trust and distrust in online fact-checking services. *Communications of the ACM*, 60:65–71, 2017.

[307] N. M. Brashier and D. L. Schacter. Aging in an era of fake news. *Current Directions in Psychological Science*, 29:316–323, 2020.

[308] J. Roozenbeek, S. van der Linden, and T. Nygren. Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, February 3, 2020.

[309] B. Swire, U. K. H. Ecker, and S. Lewandowsky. The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43:1948–1961, 2017.

[310] T. Sims, A. E. Reed, and D. C. Carr. Information and communication technology use is related to higher well-being among the oldest-old. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 72:761–770, 2017.

[311] M. A. Amazeen. Revisiting the epistemology of fact-checking. *Critical Review*, 27:1–22, 2015.

[312] M. Marietta, D. C. Barker, and T. Bowser. Fact-checking polarized politics: Does the fact-check industry provide consistent guidance on disputed realities? 13:577–596, 2015.

[313] J. E. Uscinski and R. W. Butler. The epistemology of fact checking. *Critical Review*, 25:162–180, 2013.

[314] N. Walter, J. Cohen, R. Lance Holbert, and Y. Morag. Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37, 350–375, 2019.

[315] B. Lyons, V. Mérola, J. Reifler, and F. Stoeckel. How politics shape views toward fact-checking: Evidence from six European countries. *The International Journal of Press/Politics*, 25:469–492, 2020.

[316] W. Ben Towne and J. D. Herbsleb. Design considerations for online deliberation systems. *Journal of Information Technology & Politics*, 9:97–115, 2012.

[317] C. Klein, P. Clutton, and V. Polito. Topic modeling reveals distinct interests within an online conspiracy forum. *Frontiers in Psychology*, 9:189, 2018.

[318] N. Hara, J. Abbazio, and K. Perkins. An emerging form of public engagement with science: Ask me anything (AMA) sessions on Reddit r/science. *PLOS ONE*, 14:e0216789, 2019.

[319] A. Massanari. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19:329–346, 2015.

[320] C. Klein, P. Clutton, and A. G. Dunn. Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit's conspiracy theory forum. *PLOS ONE*, 14:e0225098, 2018.

[321] A. Kende, M. van Zomeren, A. Ujhelyi, and N. A. Lantos. The social affirmation use of social media as a motivator of collective action. *Journal of Applied Social Psychology*, 46:453–469, 2016.

[322] S. Schumann and O. Klein. Substitute or stepping stone? Assessing the impact of low-threshold online collective actions on offline participation. *European Journal of Social Psychology*, 45:308–322, 2015.

[323] D. J. Wilkins, A. G. Livingstone, and M. Levine. All click, no action? Online action, efficacy perceptions, and prior experience combine to affect future collective action. *Computers in Human Behavior*, 91:97–105, 2019.

[324] D. Della Porta and A. Mattoni. Social movements. *The International Encyclopedia of Political Communication*, pages 1–8. Wiley-Blackwell, 2015.

[325] E. Volokh. Cheap speech and what it will do. *The Yale Law Journal*, 104:1805–1850, 1995.

[326] K. M DeLuca, S. Lawson, and Y. Sun. Occupy Wall Street on the public screens of social media: The many framings of the birth of a protest movement. *Communication, Culture & Critique*, 5:483–509, 2012.

[327] L. Coretti and D. Pica. The rise and fall of collective identity in networked movements: Communication protocols, Facebook, and the anti-Berlusconi protest. *Information, Communication & Society*, 18:951–967, 2015.

[328] K. McDonald. From Indymedia to Anonymous: Rethinking action and identity in digital cultures. *Information, Communication & Society*, 18:968–982, 2015.

[329] J. C. Medina Serrano, O. Papakyriakopoulos, and S. Hegelich. Dancing to the partisan beat: A first analysis of political communication on TikTok. *arxiv.org/abs/2004.05478*, 2020.

[330] Y. Theocharis, W. Lowe, J. W. Van Deth, and G. García-Albacete. Using Twitter to mobilize protest action: Online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information, Communication & Society*, 18:202–220, 2015.

[331] N. Persily. Can democracy survive the internet? *Journal of Democracy*, 28:63–76, 2017.

[332] R. M. Puhl, M. B. Schwartz, and K. D. Brownell. Impact of perceived consensus on stereotypes about obese people: A new approach for reducing bias. *Health Psychology*, 24:517–525, 2005.

[333] C. Stangor, G. B. Sechrist, and J. T. Jost. Changing racial beliefs by providing consensus information. *Personality and Social Psychology Bulletin*, 27:486–496, 2001.

[334] E. M. Zitek and M. R. Hebl. The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology*, 43:867–876, 2007.

[335] J. Shamir and M. Shamir. Pluralistic ignorance across issues and over time: Information cues and biases. *Public Opinion Quarterly*, 61:227–260, 1997.

[336] A. Todorov and A. N. Mandisodza. Public opinion on foreign policy: The multilateral public that perceives itself as unilateral. *Public Opinion Quarterly*, 68:323–348, 2004.

[337] G. J. Botvin, E. M. Botvin, E. Baker, L. Dusenbury, and C. J. Goldberg. The false consensus effect: Predicting adolescents' tobacco use from normative expectations. *Psychological Reports*, 70:171–178, 1992.

[338] D. A. Prentice and D. T. Miller. Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, 64:243–256, 1993.

[339] J. Krueger and J. S. Zeiger. Social categorization and the truly false consensus effect. *Journal of Personality and Social Psychology*, 65:670–680, 1993.

[340] J. Krueger and R. W. Clement. Estimates of social consensus by majorities and minorities: The case for social projection. *Personality and Social Psychology Review*, 1:299–313, 1997.

[341] K. P. Bauman and G. Geher. We think you agree: The detrimental impact of the false consensus effect on behavior. *Current Psychology: Developmental, Learning, Personality, Social*, 21:293–318, 2003.

[342] P. Porten-Cheé and C. Eilders. The effects of likes on public opinion perception and personal opinion. *Communications*, 45:223–239, 2018.

[343] L. Muchnik, S. Aral, and S. J. Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013.

[344] P. Dandekar, A. Goel, and D. T. Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110:5791–5796, 2013.

[345] M. Avram, N. Micallef, S. Patil, and F. Menczer. Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review*, July 28, 2020.

[346] A. J. Stewart, M. Mosleh, M. Diakonova, A. A. Arechar, D. G. Rand, and J. B. Plotkin. Information gerrymandering and undemocratic decisions. *Nature*, 573:117–121, 2019.

[347] L. Ross. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz, editor, *Advances in Experimental Social Psychology*, pages 174–221, Academic Press, 1977.

[348] Z. Leviston, I. Walker, and S. Malkin. Third annual survey of Australian attitudes to climate change: Interim report. Technical report, CSIRO, 2013.

[349] S. Lewandowsky, J. Cook, N. Fay, and G. E. Gignac. Science by social media: Attitudes towards climate change are mediated by perceived social consensus. *Memory & Cognition*, 47:1445–1456, 2019.

[350] E. Lee, F. Karimi, C. Wagner, H.-H. Jo, M. Strohmaier, and M. Galesic. Homophily and minority size explain perception biases in social networks. *Nature Human Behavior*, 3:1078–1087, 2019.

[351] G. Pennycook and D. G. Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116:2521–2526, 2019.

[352] S. A. Hale, P. John, H. Margetts, and T. Yasseri. How digital design shapes political participation: A natural experiment with social information. *PLOS ONE*, 13:e0196068, 2018.

[353] H. Ebbinghaus. Teachers College Columbia University, *A contribution to experimental psychology*. New York, 1885.

[354] A. Deligiaouri and J. Suiter. Evaluation of public consultations and citizens' participation in 2015 Better Regulation Agenda of the EU and the need for a deliberative e-rulemaking initiative in the EU. *European Politics and Society*, 10.1080/23745118.2020.1718285, 2020.

[355] C. Farina, H. Kong, C. Blake, M. Newhart, and N. Luka. Democratic deliberation in the wild: The McGill online design studio and the RegulationRoom project. *Fordham Urban Law Journal*, 41:1527, 2013.

[356] C. Nam. Behind the interface: Human moderation for deliberative engagement in an eRulemaking discussion. *Government Information Quarterly*, 37:101394, 2020.

[357] A. Ghezzi, D. Gabelloni, A. Martini, and A. Natalicchio. Crowdsourcing: A review and suggestions for future research. *International Journal of Management Reviews*, 20:343–363, 2018.

[358] R. A. Dahl. *Democracy and its Critics*. Yale University Press, 1989.

[359] S. Bradshaw and P. N. Howard. The global disinformation disorder: 2019 global inventory of organised social media manipulation. Technical report, Project on Computational Propaganda, University of Oxford, 2019.

[360] J. Suiter, E. Culloty, D. Greene, and E. Siapera. Hybrid media and populist currents in Ireland's 2016 general election. *European Journal of Communication*, 33:396–412, 2018.

[361] D. M. Farrell, J. Suiter, and C. Harris. 'Systematizing' constitutional deliberation: The 2016–18 Citizens' Assembly in Ireland. *Irish Political Studies*, 34:113–112, 2018.

[362] J. A. Elkink, D. M. Farrell, T. Reidy, and J. Suiter. Understanding the 2015 marriage referendum in Ireland: Context, campaign, and conservative Ireland. *Irish Political Studies*, 32:361–381, 2017.

[363] N. Curato, J. S. Dryzek, S. A. Ercan, C. M. Hendriks, and S. Niemeyer. Twelve key findings in deliberative democracy research. *Daedalus*, 146:28–38, 2017.

[364] S. McKay and C. Tenove. Disinformation as a threat to deliberative democracy. *Political Research Quarterly*, 10.1177/1065912920938143, 2020.

[365] S. Maffei, F. Leoni, and B. Villari. Data-driven anticipatory governance. Emerging scenarios in data for policy practices. *Policy Design and Practice*, 3:123–113, 2020.

[366] Y.-T. Hsiao, S.-Y. Lin, A. Tang, D. Narayanan, and C. Sarahe. vtaiwan: An empirical study of open consultation process in Taiwan. Available at https://osf.io/jnq8u/, 2018.

[367] C. Horton. The simple but ingenious system Taiwan uses to crowdsource its laws. *MIT Technology Review*, August 21, 2018.

[368] P. Aragón, A. Kaltenbrunner, A. Calleja-López, A. Pereira, A. Monterde, X. E. Barandiaran, and V. Gómez. Deliberative platform design: The case study of the online discussions in Decidim Barcelona. *International Conference on Social Informatics*, 277–287. Springer, 2017.

[369] A. Hudson. When does public participation make a difference? Evidence from Iceland's crowdsourced constitution. *Policy & Internet*, 10:185–217, 2018.

[370] T. Maboudi and G. P. Nadi. Crowdsourcing the Egyptian constitution: Social media, elites, and the populace. *Political Research Quarterly*, 69:716–731, 2016.

[371] OECD. Innovative citizen participation and new democratic institutions: Catching the deliberative wave. Technical report, OECD, 2020.

[372] K. Strandberg and K. Grönlund. Online deliberation. In A. Bächtiger, J. S. Dryzek, J. Mansbridge, and M. Warren, editors, *The Oxford Handbook of Deliberative Democracy*. Oxford University Press, 2018.

[373] K. Grönlund, K. Strandberg, and S. Himmelroos. The challenge of deliberative democracy online: A comparison of face-to-face and virtual experiments in citizen deliberation. *Information Polity*, 14:187–201, 2009.

[374] K. Strandberg, S. Himmelroos, and K. Grönlund. Do discussions in like-minded groups necessarily lead to more extreme opinions? Deliberative democracy and group polarization. *International Political Science Review*, 40:41–57, 2019.

[375] K. Strandberg and J. Berg. Impact of temporality and identifiability in online deliberations on discussion quality: An experimental study. *Javnost – The Public*, 22:164–180, 2015.

[376] I. M. Young. *Inclusion and Democracy*. Oxford University Press, 2002.

[377] A. Nienierza, C. Reinemann, N. Fawzi, C. Riesmeyer, and K. Neumann. Too dark to see? Explaining adolescents' contact with online extremism and their ability to recognize it. *Information, Communication & Society*, 10.1080/1369118X.2019.1697339, 2019.

[378] M. Fisher, J. Woodrow Cox, and P. Hermann. Pizzagate: From rumor, to hashtag, to gunfire in DC. *Washington Post*, December 6, 2016.

[379] C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. *Proceedings of the 25th International Conference on World Wide Web—WWW '16*, 613–624. International World Wide Web Conferences Steering Committee, 2016.

[380] S. Shugars. Good decisions or bad outcomes? A model for group deliberation on value-laden topics. *Communication Methods and Measures*, 10.1080/19312458.2020.1768521, 2020.

[381] S. Lewandowsky, T. D. Pilditch, J. K. Madsen, N. Oreskes, and J. S. Risbey. Influence and seepage: An evidence-resistant minority can affect public opinion and scientific belief formation. *Cognition*, 188:124–139, 2019.

[382] J. Cook, S. van der Linden, S. Lewandowsky, and U. K. H. Ecker. Coronavirus, 'Plandemic' and the seven traits of conspiratorial thinking. *The Conversation*, May 15, 2020.

[383] Z. Kharazian and T. Knight. Why the debunked COVID-19 conspiracy video "Plandemic" won't go away. *The Verge*, May 12, 2020.

[384] B. Wagner, K. Rozgonyi, M.-T. Sekwenz, J. Cobbe, and J. Singh. Regulating transparency? Facebook, Twitter and the German Network Enforcement Act. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 261–271. ACM, 2020.

[385] C. D. Stavrositu and J. Kim. All blogs are not created equal: The role of narrative formats and user-generated comments in health prevention. *Health Communication*, 30:485–495, 2015.

[386] S. Winter and N. C. Krämer. Who's right: The author or the audience? Effects of user comments and ratings on the perception of online science articles. *Communications: The European Journal of Communication Research*, 41:339–360, 2016.

[387] A. A. Anderson, D. Brossard, D. A. Scheufele, M. A. Xenos, and P. Ladwig. The "nasty effect:" Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, 19:373–387, 2013.

[388] M. E. Roberts. *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton University Press, 2018.

[389] M. Alizadeh, J. N. Shapiro, C. Buntain, and J. A. Tucker. Content-based features predict social media influence operations. *Science Advances*, 6:eabb5824, 2020.

[390] A. Schmidt and M. Wiegand. A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, W17-1101, 1–10. Association for Computational Linguistics, 2017.

[391] R. H. Thaler and C. R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008.

[392] A. X. Wu, H. Taneja, and J. G. Webster. Going with the flow: Nudging attention online. *New Media & Society*, 10.1177/1461444820941183, 2020.

[393] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand. Understanding and reducing the spread of misinformation online. PsyArXiv. https://psyarxiv.com/3n9u8, 2020.

[394] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31:770–780, 2020.

[395] K. Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39:629–649, 2017.

[396] 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*, 10.1287/mnsc.2019.3478, 2020.

[397] R. Hertwig and T. Grüne-Yanoff. Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, 12:973–986, 2017.

[398] S. Reijula and R. Hertwig. Self-nudging and the citizen choice architect. *Behavioral Public Policy*, 10.1017/bpp.2020.5, 2020.

[399] S. Vosoughi, M. 'Neo' Mohsenvand, and D. Roy. Rumor gauge. *ACM Transactions on Knowledge Discovery from Data*, 11:1–36, 2017.

[400] S. McGrew, M. Smith, J. Breakstone, T. Ortega, and S. Wineburg. Improving university students' web savvy: An intervention study. *British Journal of Educational Psychology,* 89:485–500, 2019.

[401] A. M. Guess and K. Munger. Digital literacy and online political behavior. Available at https://osf.io/3ncmk/, 2020.

[402] J. Cook, S. Lewandowsky, and U. K. H. Ecker. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE*, 12:e0175799, 2017.

[403] J. Roozenbeek and S. van der Linden. The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, 22:570–580, 2018.

[404] J. Roozenbeek and S. van der Linden. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5:65, 2019.

[405] M. Basol, J. Roozenbeek, and S. Van der Linden. Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, 3:1–9, 2020.

[406] L. Fazio. Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, February 10, 2020.

[407] G. Pennycook, T. D. Cannon, and D. G. Rand. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147:1865–1880, 2018.

[408] S. Banerjee, A. Y. K. Chua, and J.-J. Kim. Don't be deceived: Using linguistic analysis to learn how to discern online review authenticity. *Journal of the Association for Information Science and Technology*, 68:1525–1538, 2017.

[409] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

[410] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111:8788–8790, 2014.

[411] B. Savoy and S. Cassard. Bénédicte Savoy: Co-auteur avec Felwine Sarr d'un rapport sur la restitution du patrimoine culturel africain. *La lettre du Collège de France*, (44):32–33, 2019.

[412] G. Venkatadri, P. Sapiezynski, E. M. Redmiles, A. Mislove, O. Goga, M. Mazurek, and K. P. Gummadi. Auditing offline data brokers via Facebook's advertising platform. *The World Wide Web Conference*, 1920–1930. ACM, 2019.

[413] D. A. Martin, J. N. Shapiro, and M. Nedashkovskaya. Recent trends in online foreign influence efforts. *Journal of Information Warfare*, 18:15–48, 2019.

[414] M. Wigell. Hybrid interference as a wedge strategy: A theory of external interference in liberal democracy. *International Affairs*, 95:255–275, 2019.

[415] R. Craufurd Smith. Fake news, French law and democratic legitimacy: Lessons for the United Kingdom? *Journal of Media Law*, 11:52–81, 2019.

[416] A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti, and D. Pedreschi. Privacy-by-design in big data analytics and social mining. *EPJ Data Science*, 3:10, 2014.

[417] C. Martinez. Driving relevance and inclusion with multicultural marketing. *Facebook for Business,* October 28, 2016.

[418] E. Hafen. Personal data cooperatives: A new data governance framework for data donations and precision health. In J. Krutzinna and L. Floridi, editors, *The Ethics of Medical Data Donation*, pages 141–149, Springer, 2019.

[419] M. Harbach, M. Hettig, S. Weber, and M. Smith. Using personal examples to improve risk communication for security & privacy decisions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2647–2656. ACM, 2014.

[420] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad news travel fast: A content-based analysis of interestingness on Twitter. *Proceedings of the 3rd International Web Science Conference*, 1:1–7. ACM, 2011.

[421] A. Osho, C. Waters, and G. Amariucai. An information diffusion approach to rumor propagation and identification on Twitter. *arXiv:2002.11104*, 2020.

[422] M. M. Malik and J. Pfeffer. Identifying platform effects in social media data. *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, 241–249. AAAI, 2016.

[423] N. Alipourfard, B. Nettasinghe, A. Abeliuk, V. Krishnamurthy, and K. Lerman. Friendship paradox biases perceptions in directed networks. *Nature Communications*, 11:707, 2020.

[424] H. Kundnani. The future of democracy in Europe: Technology and the evolution of representation. Technical report, Chatham House, 2020.

[425] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2387–2395. IEEE, 2016.

[426] H. A. Simon. Complexity and the representation of patterned sequences of symbols. *Psychological Review*, 79:369–382, 1972.

[427] G. Gigerenzer, R. Hertwig, and T. Pachur, editors. *Heuristics: The Foundations of Adaptive Behavior.* Oxford University Press, 2011.

[428] R. Hertwig, T. J. Pleskac, and T. Pachur, editors. *Taming Uncertainty*. MIT Press, 2019.

# List of abbreviations and definitions

**Algorithm:** An unambiguous procedure to solve a problem or a class of problems. It is typically composed of a set of instructions or rules that take some input data and return outputs.

**Algorithmic content curation:** Auto- mated selection of what content should be displayed to users, what should be hidden, and how it should be presented.

**Algorithmic decision-making:** The pro- cessing of input data to produce a score or a choice that is used to support decisions such as prioritisation, classification, association, and filtering.

**Attention economy:** Human attention limits what we can perceive in stimulating environments and what we can do. Attention has become a commodity online with platforms vying for user engagement. (Also sometimes called the "dopamine economy" because of the presumed addictive properties of social media.)

**Avatars:** An icon, graphic or other image by which a person represents herself online.

**Bots:** A software program that can execute commands, reply to messages or perform routine tasks, thus mimicking human communication either automatically or with minimal human intervention. Social media bots may retweet certain posts to gather attention.

**Choice Architectures:** Refers to the practice of influencing choice by organising the context in which people make decisions.

**Cognition:** The mental action or process of acquiring knowledge and understanding through thought, experience, and the senses.

**Dark patterns:** Design choices that benefit an online service by coercing, steering or deceiving users into making unintended and potentially harmful decisions.

**Digital fingerprint:** A distinct, data-driven identifier comprising tiny bits of personal data.

**Disinformation:** False, fabricated or manipulated content shared with intent to mislead or cause harm.

**Echo chamber:** An environment in which a person encounters only beliefs or opinions that coincide with her own, so that existing views are reinforced and alternative ideas are not considered.

**Epistemic quality:** The quality of information as valid components of knowledge.

**European Information Space (used in the foresight process):** All factors and entities that directly or indirectly contribute to the creation, dissemination and processing of information on the individual, group and societal level, including the translation of such information into political action.

**Explainability:** The possibility to explain the function of artificial intelligence technology so that their solutions are being under- stood by humans.

**Fake news:** A form of manufactured news consisting of deliberate disinformation or hoaxes that seeks to mimic the format of real news.

**Filter bubble:** A situation in which an Internet user encounters only information and opinions that conform to and reinforce her own beliefs, caused by algorithms that personalise information diets.

**Fundamental rights:** The Charter of Fundamental Rights of the European Union guarantees personal, civic, political, economic, and social rights and freedoms to individuals in the EU.

**Instrumental-variable analysis:** A sta- tistical technique that permits identification of causality in the associations between ob- served variables.

**Internet:** A global computer network con- sisting of interconnected nodes using stan- dardised communication protocols.

**Machine learning:** When computers dis- cover how they can improve performance from provided data, without being explicitly programmed to do so.

**Mal-information:** Genuine information shared with intent to cause harm, such as hate speech and leaks of private information.

**Microtargeting:** Customised marketing messages delivered to a niche audience shar- ing relevant interests, based on personal data provided by users or inferred from their on- line behaviour.

**Misinformation:** false or misleading con- tent created and initially presented without malicious intent.

**Online moderation:** Methods used to sort contributions that are irrelevant, obscene, illegal or insulting with regards to useful or informative online contributions.

**Profiling:** Any form of automated personal data processing that uses personal data to evaluate certain personal aspects relating to a natural person.

**Search Engine Optimization (SEO):** The process of growing the quality and quantity of website traffic by increasing the visibility of a website or a web page to users of search engines.

**Sock-puppet:** A fake persona used to dis- cuss or place comments online. Usually created to manipulate or deceive for political ends.

**Social media:** Websites and applications that enable users to create and share content or to participate in social networking.

**Strategic foresight:** A way of exploiting our inherent storytelling abilities to create a manageable and memorable number of plausible stories of the future.

**User authentication:** A security process that encompasses the human-to-computer interactions that require the user to log in.

**Web:** The World Wide Web, commonly known as the Web, is an information system where documents and other resources are identified by web addresses (URLs), which may be interlinked by hypertext, and are accessible over the Internet.

**Worldview:** The collection of attitudes, values, stories and expectations about the world around us, which inform our thoughts and attitudes and provide a schema for how the world should be structured.

# List of figures

# List of tables

KJ-NA-30422-EN-N

## The European Commission's science and knowledge service

Joint Research Centre

## JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

**EU Science Hub**
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office
of the European Union

doi:10.2760/709177

ISBN 978-92-76-24088-4