Optimising Emotions, Incubating Falsehoods

How to Protect the Global Civic Body from Disinformation and Misinformation

Vian Bakir · Andrew McStay

OPEN ACCESS

palgrave
macmillan

Optimising Emotions, Incubating Falsehoods

Vian Bakir • Andrew McStay

Optimising Emotions, Incubating Falsehoods

How to Protect the Global Civic Body from Disinformation and Misinformation

palgrave
macmillan

Vian Bakir
Bangor University
Bangor, UK

Andrew McStay
Bangor University
Bangor, UK

ISBN 978-3-031-13550-7 ISBN 978-3-031-13551-4 (eBook)
https://doi.org/10.1007/978-3-031-13551-4

Cover illustration: Maram_shutterstock.com

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Few could have missed that in recent years we all seem to be feeling more. Maybe things have always been this way, but, frankly, things have just felt more intense, more 'feely'. Of course, one or two things have happened in recent years that caused everyone to ask fundamental questions about human life and its significance. As researchers, writers and teachers interested in politics, communications and the social implications and impact of new technologies, but also living through these 'feely' times, we wanted to put these together. This raised basic questions of what is going on in politics and in media to amplify and even tweak human emotion for given political and behaviour change goals.

As academics working in the UK, we wondered, too, from our locked-down homes: how is the rest of the world responding to political and technological interest in profiling, optimising and interacting with 'civic bodies', which seem to be feeling more than they have done for some time?

Also, as academics that study organisations and technologies built to profile human emotion in relation to biometrics and human body, our self-imposed task was to think: if this is now, what next? What of the emerging media environment where bodies, voices and faces are routinely profiled by devices, environments and services? Who will use these technologies and to what end? Much of this is yet to happen, but our hunch is that the current algorithmically mediated, feely world and body-based technologies will meet.

In addition to intuitions and personal motivations, the book came about because of our engagement with various parliamentary, regulatory and research council bodies across the world on what became the pressing

global topic of fake news and disinformation online. It seems a long time ago now, but there were two key lightening rod moments that captured decision-maker attention: the events of 2016 with the unexpected election of Donald Trump to US President and the unexpected referendum result in the UK to leave the European Union. Both political campaigns embraced false and emotive information targeted at the civic body. Decision-makers wanted to know what fake news and disinformation consisted of, how they worked, if they were harmful to society and what should be done about them. Academics proffered their advice from every conceivable discipline, and a range of new research programmes were funded by diverse research councils. Stakeholders from those parts of society affected by, and involved with, false information online weighed in with their views, some proving more influential than others.

Simultaneously, our Emotional AI Lab (EmotionalAI.org) was researching the increasingly important role of automated technologies that 'feel into' people's emotions, affects and moods, and these, we realised, were fundamental to understanding the contemporary phenomenon of false information online. Our book had many iterations as technologies, commercial practices and regulatory policies continued to evolve and as research started to accumulate to tentatively agree on the nature, scale, cause and impact of false information online. This version of our book, finalised across Spring 2022, is unlikely to be the last word on this still evolving phenomenon. We hope that it proves useful in spurring further discussion, and appropriate action, in protecting us all from the global rising tide of emotional profiling that we observe today.

Bangor, UK

Vian Bakir
Andrew McStay

Acknowledgements

We would like to thank our UK Research and Innovation funders, the Economic and Social Research Council (ESRC) and the Arts and Humanities Research Council (AHRC), as well as Japan's research council, the Japan Science and Technology fund (JST), for funding us across several grants that supported this book, especially its comparative focus. These grants comprise *Emotional AI in Cities: Cross Cultural Lessons from UK and Japan on Designing for an Ethical Life*, responding to an ESRC-JST Joint Call on Artificial Intelligence and Society (Grant ref. ES/T00696X/1), and ESRC-AHRC UK-Japan Social Sciences & Humanities Connections grant, *Emotional AI: Comparative Considerations for UK and Japan across Commercial, Political and Security Sectors* (Grant ref. ES/S013008/1). We'd also like to thank our wider team for invaluable cross-disciplinary conversations and very good company. From the UK end, these were Alexander Laffer, Diana Miranda, Lena Podoletz, and Lachlan Urquhart. From the Japan end, these were Nader Ghotbi, Tung Manh Ho, Peter Mantello, Hiroshi Miyashita, and Hiromi Tanaka. And as ever, gratitude goes to the ageing little cat that imparts daily life lessons on the importance of sleep, sunshine and, weirdly, salmon oil.

Contents

About the Authors

Vian Bakir is Professor of Journalism and Political Communication at Bangor University, UK. She is an expert in the impact of the digital age on strategic political communication, dataveillance and disinformation. Her most recent book is *Intelligence Elites and Public Accountability: Relationships of Influence with Civil Society* (2019). She has been awarded multiple research council grants and Arts Council grants on data governance and transparency. She has advised the UK All Party Parliamentary Group on Extraordinary Rendition (2022), the UK All Party Parliamentary Group on AI (2020), the UK All Party Parliamentary Group on Electoral Transparency (2019–2020), the UK Parliament's Inquiry into Fake News and Disinformation (2017–2019) and the Parliament of Victoria (Australia) Electoral Matters Committee on social media and political campaigning (2020–2021). Her academic outputs can be found at https://orcid.org/0000-0002-6828-8384.

Andrew McStay is Professor of Digital Life at Bangor University, UK. He is a world-leading scholar in how emotion-sensing technologies transform society, with his work regularly featured in international media and his major books translated into Chinese and Indonesian. His eight monographs include, most recently, *Emotional AI: The Rise of Empathic Media* (2018), which examines the impact of technologies that make use of data about affective and emotional life. Director of the Emotional AI Lab, current projects include cross-cultural social analysis of emotional AI in the UK and Japan. Non-academic work

includes membership of the Institute of Electrical and Electronics Engineers (P7000/7014) and ongoing advising roles for start-ups, non-governmental organisations and policy bodies. He has also appeared and made submissions to the United Nations Office of the High Commissioner on the right to privacy in the digital age, to the UK House of Lords AI Inquiry and to various inquiries for the UK Department for Digital, Culture, Media and Sport.

PART I

Conceptual Tools and Contexts

CHAPTER 1

Optimising Emotion: Introducing the Civic Body

Introduction

Emotion plays a vital role in modern societies, especially given circulation of knowingly and unwittingly spread false information. This book assesses how this has come to be, how we should understand it, why it matters, what comes next and what we should do about it. We start with three observations.

Firstly, *false information is prevalent online and causes real-world civic harms.* Multiple concepts associated with false information achieved linguistic prominence across the early twenty-first century, indicating the scale of the problem. In 2006, 'truthiness' was Word of the Year for Merriam-Webster: it refers to 'a truthful or seemingly truthful quality that is claimed for something not because of supporting facts or evidence but because of a feeling that it is true or a desire for it to be true' (Merriam-Webster, 2020). A decade later, 'post-truth' became Oxford Dictionaries Word of the Year, defined as, 'relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief' (Oxford Dictionaries, 2016). By 2017, a year after Donald Trump became US president, 'fake news' was word of the year for Collins' English dictionary, defined as 'false, often sensational, information disseminated under the guise of news reporting' (Collins English Dictionary, 2017). In 2018, 'misinformation' was Dictionary.com's word of the year, namely: 'false information that is

V. Bakir, A. McStay, *Optimising Emotions, Incubating Falsehoods*,
https://doi.org/10.1007/978-3-031-13551-4_1

spread, regardless of whether there is intent to mislead' (Dictionary.com, 2018). This linguistic infiltration indicates the prominence of false information in recent years, as well as widespread public concern.

Such concern built across the second and third decades of the twenty-first century. By 2013, massive digital misinformation was so pervasive in social media that it was listed by the World Economic Forum as a major societal threat. As the COVID-19 pandemic broke out in 2020, the World Health Organization (2020) expressed concerns about an 'infodemic', namely, 'too much information including false or misleading information in digital and physical environments during a disease outbreak' causing 'confusion and risk-taking behaviours that can harm health' and 'mistrust in health authorities'. COVID-19 conspiracy theories led to preventable deaths as many refused to be vaccinated against this highly infectious, novel disease. Yet, we should resist the idea that false information is something that *those* people believe. Although some viral claims were outlandish (such as the theory that co-founder of Microsoft, Bill Gates, masterminded the pandemic to implant microchips into humans alongside the vaccine), others were conceivable, while false (such as the vaccine's potential impact on fertility). Given limited understanding, persuasive and professional looking 'news', and the lingering 'what if' question, people reached their judgements. Philosophically, we take the view that although facts certainly exist, people mostly do not passively observe the world. Rather, we are coping and adapting, and while this involves rational decision-making, judgements are rightly informed by emotions.

Indeed, our second observation is that *emotion is fundamental to civic life*. Governments seek to influence their population's behaviour using insights from behavioural economics and cognitive psychology into the role of emotions in decision-making, for instance, to encourage compliance with COVID-19 biosecurity rules. Political parties and campaign groups use civic emotions to invigorate democratic life, drive voter mobilisation and nudge opinion formation. Direct action and protest against polluting corporations run on hope for change, channelling anger to shame decision-makers into addressing the climate emergency. Celebrations of hard-won civic rights invoke collective pride among marginalised communities while provoking others who feel unseen and left behind. Public efforts to help refugees are fuelled by pity and empathy, but equally, opposition to economic immigrants runs on suspicion and fear.

Our third observation is that *profiling and optimisation of emotions using automated systems (AI and machine learning algorithms) are*

escalating features of daily life. Such systems, present in social media and elsewhere, record and analyse people's psychological and behavioural characteristics. They do this to profile, label, classify and judge people, largely for the purposes of refining, targeting, boosting and otherwise optimising messaging that people are exposed to. This, today, is mostly automated and algorithmic, involving sets of instructions for computers to label, classify and inform how a system will 'decide' (Kitchin, 2017). Traditionally algorithms are rules predefined by human experts to perform a task and make sense of data, although today's digital platforms use machine learning techniques to look for patterns in big data. Supervised when the computer is being told what to look for and unsupervised when told to inductively create patterns and labels, both approaches used by digital platforms process vast amounts of information about human subjectivity to offer personalised services, content and advertisements (hereafter, ads) (van Dijck et al., 2018). Increasingly, the *datafied emotions* of consumers, users and citizens are also algorithmically gauged and profiled by private companies and governments worldwide for diverse purposes, including to persuade, influence and monetise.

At the established end of the spectrum, this involves profiling the behaviour and emotions of social media users to increase their engagement with the platform and their value to online advertisers. As we discuss in Chap. 2, the world's largest suite of social media platforms, Meta Platforms Inc. (formerly called Facebook until 2021), comprising Facebook, WhatsApp, Instagram and Messenger, continuously tweaks its machine learning algorithms to maximise user engagement. These technologies have real-world consequence, shown, for example, by internal Facebook documents leaked in 2021. These reveal how tweaks to its News Rank algorithm (that determines what posts users see in their News Feed) promoted politically and socially destructive, extremist, viral, false information. This, in turn impacts real-world politics. Another leaked Facebook report states that political parties across the world felt that Facebook's algorithmic change forced them into more extreme, negative policy positions and communications. Such algorithmic tweaks have global reach. Facebook had 2.9 billion monthly active users in 2021. Its parent company, Meta, claims 3.6 billion monthly active users in 2022 across Facebook, WhatsApp, Instagram and Messenger, reaching almost half of the world's population.

At the more emergent end of the spectrum, the profiling of behaviour and emotions involves a wide range of biometric profiling, variously

trialled and deployed worldwide to generate more persuasive ads and call centre workers, more reactive voice assistants and toys, cars that compensate for drivers' mental and emotional states, border guards that detect travellers' deception, police that anticipate dangerous situations in crowds and teaching that monitors children's concentration. In short, social life is becoming increasingly profiled in efforts to assess emotional and psychological disposition for the purposes of profiteering, persuasion, wellbeing, influence, resource allocation, social engineering, safety and security. Adequacy of methods is more than a little debatable, but we see this as an ongoing development of a process that began several decades ago with experimentation with biosensors in affective computing and the rise of social media platforms. It is now spreading to encompass diverse biometric data capture across existing media and emergent services.

In this book we focus on how current emotional profiling fuels the spread of false information online, consider the implications of emergent emotion profiling and suggest what can be done about these developments. Straddling our three opening observations, there is now evidence that emotion profiling incubates false information online, causing significant harms worldwide. This chapter frames these developments in terms of a *civic body* increasingly affected by processes of *optimised emotion*.

Optimising Emotion

Emotions are powerful drivers of decision-making and behaviour. As such, there is commercial, political, ideological and discursive power in understanding and influencing emotions and collective feeling. This has long been known by advertisers seeking to influence consumer behaviour, political campaigners seeking power, governments seeking behavioural management of populations, journalists seeking to maximise readership, trade unions seeking solidarity and social movements seeking social change. Unsurprisingly, then, the formation and manipulation of irrational (emotional) publics has long been of concern across multiple vectors of interest, most obviously, the Frankfurt School's mass society thesis (for instance, Marcuse (1991 [1964])) and public sphere theorisation and critiques (such as Habermas (1992 [1962]) and Calhoun (2010)).

Concomitantly, recent years have seen increasing datafication of emotion and the rise of emotional artificial intelligence ('emotional AI') and so-called empathic technologies. These technologies use machine training to read and react to human emotions and feeling through text, voice,

computer vision and biometric sensing, thereby simulating understanding of affect, emotion and intention (McStay, 2018, 2022). The roots of such processes that convert emotional life into data lie in 'affective computing' (Picard, 1997) which, in the 1990s, measured biometric signals such as heart fluctuations, skin conductance, muscle tension, pupil dilation and facial muscles to assess how changes in the body relate to emotions. Today, the emergent social picture is one where *datafied emotion is optimised to form a fundamental component of personalisation, communication and experience.*

Only a few years ago practical use cases of biometric emotional AI were rare, such as in outdoor ads that, enabled by emotional AI, changed themselves over time to elicit more smiles from passers-by (these captured by cameras embedded above the ad). Emotion optimisation is now becoming more mainstream, at least in certain consumer-facing sectors (McStay, 2018). For instance, major car manufacturers worldwide are deploying in-car cameras and affect and emotion tracking systems to profile drivers' emotional behaviour to personalise in-cabin experience and improve safety (McStay & Urquhart, 2022). Emotion-focused wearables are increasingly popular with Amazon's Halo (a fitness, mood and wellness tracker) and Garmin systems (tracking stress and the body's 'battery') to help users manage their mental health and overall day (Dignan, 2020, December 14; Garmin, 2022). Legacy companies such as Unilever and IBM in the USA, and SoftBank in Japan, use emotional analytics for recruitment purposes (Richardson, 2020). Although there is extensive academic scepticism about over-simplistic methodological approaches used by the emotional AI industry to translate biometric signals into emotional inferences (McStay, 2018, 2019), this has not prevented trialling and development of biometric forms of emotional AI in even more sensitive domains, often in countries with very different data protection and privacy regimes. This includes border security, policing (Wright, 2021), smart cities (McStay, 2018), education (Article 19, 2021; McStay, 2019) and children's toys (McStay & Rosner, 2021).

However, mass datafication and optimisation of emotion was pioneered and honed by social media platforms, which we observe as the dominant use case of emotional AI worldwide today. This follows two decades of continuous development of their emotional profiling and targeting tools, premised on surveillance and sharing of users' data for the purposes of modifying user behaviour and maximising user engagement with their platforms. We will return to this point in Chap. 2, but for now, we observe

that many accounts link dominant social media platforms' quest for greater user engagement to the viral spread of false information. Indeed, a survey in 2020 covering all five continents reveals that globally, people see social media as the biggest source of concern about misinformation (40%), well ahead of news sites (20%), messaging apps (14%) and search engines (10%). Overall, the greatest concern is with the world's biggest social media platform, Facebook (29% are most concerned about Facebook), followed by Google-owned YouTube (6%) and Twitter (5%) (Newman et al., 2020). As such, it is social media that predominate throughout this book when discussing false information online. In the final chapter, we move beyond social media to embrace more emergent emotional AI forms as we delve into near-horizon futures and assess implications for *civic bodies* that are being profiled and *optimised* in increasingly novel ways. This entails use of AI technologies and organisational claims to see, read, listen, judge, classify and learn about emotional life through biometric and human state data (McStay, 2018).

Seen one way, 'optimisation' is the language of efficiency, making the best or most effective use of a situation or resource. Yet, when applied to datafied human emotion and civic functions (such as a public sphere of news, debate and information flow shaped by social media and search engines), 'optimisation' cannot fail to become something more political, more contentious. This is particularly so when set against critical understanding of Silicon Valley's neoliberal, free market worldview and 'techno-capitalism' where globalised, powerful corporations profit from intangibles such as new knowledge, intellectual property, research creativity and technological infrastructure (Suarez-Villa, 2012). Such critique raises classic questions of exploitation and choice: who decides, or has a say in, what is optimal, optimisable, or optimised in a public sphere shaped by datafied emotion? Furthermore, who benefits from these decisions, who is harmed and what is lost along the way? We hope, in this book, to provide some answers.

The 'Civic Body'

We advance the notion of the *civic body*, to capture the various ways by which datafied emotion is collected, processed and optimised, especially as it relates to information, between individuals and collectives. The *civic body* has an antecedent in the 'body politic', a principle originating with Plutarch, the Greek Middle Platonist philosopher who regarded the polity

as akin to a body having a life (Rigby, 2012). This has been through many iterations, focusing on different aspects of the body. Head-oriented accounts of the body politic focus on the head of the body, be this monarchs, rulers, decision-makers and hierarchical conceptions. Other bodily metaphors focus on limbs (such as the long arm of the law) or organs (typically the heart, belly and bowel) (Musolff, 2010). Still other bodily metaphors conceptualise the polity as based on equilibrium and interdependence. Just as a body requires balance to ensure homeostasis, social harmony is required among key stakeholders for the wellbeing of the entire social and political organism.

Although the 'body politic' is a classical way of understanding the relative importance and ecological interactions of government, monarchs, police, military and other heads, limbs and organs of the polity, the 'civic body' (as advanced in this book) is a more citizen-oriented concept in drawing attention to the datafied emotion of individuals and collectives. The *civic body* is also a less metaphorical concept than the body politic. Quite literally, we deploy the concept of *feeling-into* and optimising the *civic body* as a way to account for the variety of modalities by which interested parties seek to understand not only citizens' expressed and inferred preferences but increasingly their biometric correlates. This, for us, is a key change, in that the profiling of human behaviour increasingly involves information about the body, be this our faces, voices, or biometrics collected by body-worn technologies. This allies well with (but is not reliant on) biopolitical writing, crystallised in Rose's definition of biopolitics as the capacity 'to control, manage, engineer, reshape, and modulate the very vital capacities of human beings as living creatures' (Rose, 2006, p. 3) and the general interest in integrating bodies into systems (Rabinow & Rose, 2006). We also recognise biopolitical interest in organisational proclivity towards life sensitivity and what biopower scholars phrase as the 'molecular' (Deleuze & Guattari, 2000 [1972]; Foucault, 1977; Lazzarato, 2014; McStay, 2018) that represents a shift from macro- to micro-interests of a biological sort in populations, subjectivity and governance thereof.

There are a variety of modalities by which interested parties use technologies to *feel-into* the *civic body*. Today, this is still typically done through polling, interviews and focus groups. For instance, since 2017, Microsoft has compiled a Digital Civility Index through an online survey of adults and teens across more than 20 countries and most continents to explore their perceptions of online incivility. It covers topics such as being 'treated mean', trolling, hate speech, online harassment, hoaxes, frauds, scams,

discrimination and sexual solicitation. Its 2021 survey finds that the most civil country online is the Netherlands, followed by Germany, the UK, Canada and Singapore. The most uncivil country online is Colombia followed by Russia, Peru, Argentina, India and Brazil (Microsoft, 2021). Complementing such standard tools for feeling-into the *civic body*, emotional AI and wider empathic technologies are also deployed. Modalities may include (1) sentiment analysis, (2) psycho-physiological measures and (3) urban data, each discussed below.

The first modality, sentiment analysis, focuses on online language, emojis, images and video for evidence of moods, feelings and emotions regarding specific issues. Sentiment is often inferred through social media, and studies have been conducted on Twitter to measure life satisfaction in Turkey (Durahim & Coşkun, 2015); to create a 'hate map' to geolocate racism and intolerance across the USA (Stephens, 2013); and to measure and predict electoral outcomes (Ceron et al., 2017). Beyond social media sentiment analysis, music players such as Spotify sell data about specified user groups, providing insights on moods, psychology, preferences and triggers. ManTech, a US government contractor, has developed a model that uses open-source intelligence to predict and track foreign influence operations in any country, including covert actions by foreign governments to influence political sentiment or public discourse. Its main data source is Google's Global Database of Events, Language and Tone (GDELT) that monitors the world's broadcast, print and Web news in over 100 languages and identifies emotions, as well as people, locations, organisations, themes, sources, counts, quotes, images and events (Erwin, 2022, April 25).

The second modality, psycho-physiological measures, entails more focused assessment of bodies themselves, such as via laboratory-based tracking of facial expressions to gauge our expressions when shown political ads. Other psycho-physiological means include wearable devices that sense various responses (such as skin conductivity, moisture and temperature; heart rate and rhythms; respiration rate; and brain activity). Voice analytics try to parse not just what people say but how they say it. These include elements such as rate of speech, increases and decreases in pauses, and tone (McStay, 2018). Indeed, Amazon has long sought for its ubiquitous Alexa to profile users' voice behaviour to gauge emotion, despite technical and methodological challenges.

Finally, the third modality, urban data, moves beyond sentiment analysis and laboratory settings, to *feel-into* 'living labs' (Alavi et al., 2020).

Expanding beyond laboratory walls, this involves research, analysis and surveillance in more open settings, for example, a city or a larger polity. This harnesses insights gathered from traditional research techniques, online media and laboratory-based response analysis, but also *feels-into* the *civic body* through a vast array of public and private means. These include footfall, transport usage, data from mobile phones, spending patterns, urban cameras (that may register numbers, identities and expressions of emotion), health data and citizen complaints. This involves 'big data' logics to identify patterns in massive volumes of unstructured data from multiple sources and react quickly (Ceron et al., 2017). However, *feeling-into* the *civic body* goes further than this: quantity is used to help deal with political 'why' questions.

We posit that a healthy *civic body* requires a healthy media system. Unfortunately, so far, the datafication and optimisation of individual and civic emotion by digital platforms has helped incubate and amplify an ecology of false information throughout the *civic body*.

Incubating False Information in the Civic Body

There are many processes that fuel the ecology of false information among 'networked publics' (namely, publics restructured by networked technologies) (Boyd, 2010). These include epistemological processes (such as the rise of 'post-truth'); cultural processes (such as the decline of trust in political elites, experts and journalists); political processes (that fan emotion, such as nationalism and populism); economic processes (such as increasing competitive pressures on news outlets, generating tendencies to produce ever more engaging content); regulatory processes (such as uneven data privacy protections); and media and technological processes (such as the global rise of social media platforms and their interplay with legacy and alternative media systems). The view from Central and Latin America, for example, finds many of these processes underpinning widespread disinformation in elections held in 2018 in Brazil, Colombia and Mexico and in 2011 and 2015 in Guatemala. According to reports by news outlet, *The Intercept*, and by the Atlantic Council (a non-partisan think tank that seeks to galvanise American leadership to address global challenges), these processes include structural political corruption, mistrust in politicians and a desire for change; economic downturns and unemployment; an absence of data protection cultures; political challengers adept at using social media; and features within social media platforms that enabled virality (for

instance, in Brazil, each WhatsApp user could create up to 9999 groups, each with up to 256 people and could forward a message to 20 contacts simultaneously) (Bandeira & Braga, 2019; Bandeira et al., 2019; Currier & Mackey, 2018, April 7).

Mindful of these broader processes, we focus on the digital media and technological element. This is centrally important as false information online is greatly facilitated by the affordances of digital networked environments, namely, what these technological systems enable to happen and how they are used in particular contexts (Rice et al., 2017) and the sifting, sorting and judging processes therein. With few resources, 100% fake news websites hosting totally made-up stories can be created and made to look like genuine news content. Digital manipulation tools can increasingly easily be bent towards changing images and video (via deepfakes and shallowfakes), thereby deepening the rupture between recorded image and reality (a problem long discussed by Media Studies). These deceptive messages can have a long shelf life. As well as faking content, identities can be easily disguised online (the phenomenon of 'sock puppets'). These deceptive accounts and messages can be made to appear popular through amplification via campaigners, bots and targeting influential humans to manipulate online conversation. Indeed, in 2017, Facebook estimated that 2–3% of its worldwide monthly active users were 'user-misclassified and undesirable accounts', with a far higher percentage in developing markets such as India, Indonesia and the Philippines (Facebook, 2017, September 20).

People are concerned about false information online, as repeatedly shown in global surveys. Annual surveys by the University of Oxford's Reuters Institute (funded by the Thomson Reuters Foundation, Google, Facebook and other donors) conducted across scores of countries, and all five continents from 2018 to 2022 find over half (around 54% to 58%) of respondents are concerned about what is real and fake online (Newman et al., 2018, 2021, 2022). There are large variations between regions. In 2021, there was most concern in Africa (74%), followed by Latin America (65%), North America (63%) and Asia (59%), with the lowest concern in Europe (54%) (Newman et al., 2021). A survey of 27 countries in 2020 finds only 29% agree that they have 'good information hygiene' in engaging with news; avoiding echo chambers; verifying information; and not amplifying unvetted information (Edelman, 2021).

Alongside national differences in concern over false information, there are demographic differences. Surveys representative of the digital

population from Argentina, Chile and Spain across 2018 and 2019 show that concern increases with age, women, self-identified left-leaning users and those with high interest in political news (Rodríguez-Virgili et al., 2021). Age is also a factor in a 2018 survey of 28 European Union Member States, with older respondents less confident in their ability to identify fake news than other age groups (Eurobarometer, 2018).

Regions with the highest levels of concern (Africa, Latin America) over false information online correspond closely with high levels of use of social media for news. Different platforms also engender different levels of concern worldwide. Globally, the greatest concern is with Meta-owned Facebook (identified by 29% in 2020): this is unsurprising given that it is the most used social network worldwide. In parts of the Global South, such as Brazil, Chile, Mexico, Malaysia and Singapore, people are more concerned about closed messaging apps like (Meta-owned) WhatsApp where false information is less visible and harder to counter: for instance, 35% are concerned in Brazil. Twitter is seen as the biggest problem in Japan and YouTube in South Korea (Newman et al., 2020, 2021, 2022). Even in China, where the communication environment is tightly controlled and American digital platforms are absent, people are concerned about fake news. This is especially so on social networking site Sina Weibo (Twitter's equivalent in China, with over 307 million monthly active users in 2021) and on Tencent's popular messaging app WeChat (with over 1 billion monthly active users in 2021). A 2018 survey finds that about seven in ten respondents believe that fake news poses 'a great deal' or 'a fair amount' of threat to Chinese society and 12% think that 'most' of the news on social media is made up (Tang et al., 2021).

Not all types of false information are of equal concern. Surveys from Argentina, Chile and Spain across 2018 and 2019 show that participants are most concerned by stories where facts are twisted to push a particular agenda, followed by those completely made up for political or commercial reasons. They were least concerned by poor journalism (factual mistakes, dumbed-down stories, misleading headlines and clickbait) (Rodríguez-Virgili et al., 2021). While there remains widespread media coverage of attempts by outside powers to undermine elections abroad, Newman et al.'s (2020) survey across 40 countries finds that it is domestic politicians that are seen as by far the most responsible for false and misleading information online (40%), followed by political activists (14%), journalists (13%) and ordinary people (13%), with only 10% concerned about foreign governments. In some countries the figure holding domestic politicians as

most responsible for false information is even higher (for instance, in Brazil, the Philippines, South Africa and the USA) (Newman et al., 2020, p. 18).

Recognition of the harms that false information can inflict on the *civic body* has generated a search for solutions at national and supranational levels. A global response is necessary given the economic power, political value and transnational nature of dominant digital platforms. Especially where platforms operate in jurisdictions without data protection legislations, corporations may decide what data is collected, who can access and use the data, and why. This is not simply a privacy issue, but one of fairness and justice. For instance, in 2020, Facebook blocked an international investigation into use of hate speech on its platform to incite genocide against Rohingya Muslims in Myanmar in 2018 (Smith, 2020, August 18). This obstructs 'data justice', a term advanced by Taylor (2017) to advocate fairness in how people are made visible, represented and treated arising from their production of digital data.

Optimising Society and Subjectivity

Despite such concerns, optimising emotional data could be a force for civic good. Journalists have long appreciated the need to emotionally engage audiences: worthy stories that go unread have little value. Consider, too, the increase in engagement, mobilisation and togetherness across civic practices, where citizens care enough to go out and vote, or where they reach out to each other in solidarity and empathy. In a situation where data about emotion is increasingly ubiquitous, powerholders may seek to *feel-into* localised emotions and civic moods to form their policies and to help govern and better care for their wards. Already, since 2011, the UK's Office for National Statistics has tracked national and local authority-level average ratings of life satisfaction, happiness, anxiety and whether things feel worthwhile. For instance, it finds that in the build-up to the first national COVID-19 lockdown in March 2020, average anxiety jumped to its highest level since measurements began, and average happiness levels declined steeply (Office for National Statistics, 2020). More finally grained emotional data would prove useful to governments seeking to model and manage population behaviour at aggregate and localised levels, especially during upheavals like pandemics.

Yet, there is clearly scope for harms. These include the rise of 'empathically-optimised automated fake news' that exploits users'

outrage, tribalism and preconceived ideas (Bakir & McStay, 2018, 2020); computational propaganda that attempts manipulation of public opinion through an assemblage of social media platforms, autonomous agents and big data (Woolley & Howard, 2018); information warfare that provokes intense anxiety among targeted populations (Bolton, 2021); and political campaigning that profiles how we secretly feel in order to push anti-social emotional buttons (such as resentment towards specific groups) (Bakir, 2020).

For better or worse, the psychological and emotional behaviour of individuals and groups is increasingly quantified and datafied for the purposes of monetisation and influence (McStay, 2018). These processes are opaque but draw upon influential 'behavioural sciences' (Thaler & Sunstein, 2008) that downplay rationality in favour of a neo-behaviourist outlook. Furthermore, users are kept from seeing much of how their behaviour is monetised and shaped, whether by hidden trackers and behavioural advertising pixels that follow them around the Web, or by the secret processes that determine content moderation on platforms (Gorwa & Ash, 2020). While all these developments are observable, their impacts on human subjectivity and autonomy are more debatable.

Following a long tradition in Media and Cultural Studies that laments the loss of human agency, creativity and ability to think for oneself in the face of commercial or propagandistic mass communications (as in mass society studies), critical and biopolitically inclined scholars similarly object that neo-behaviourism and seeing people in psycho-physiological terms disregards (or denies) agency and civic autonomy. For instance, Andrejevic (2020, p. 2) highlights the peril of 'automated media' creating an 'automated subject' whose wants and needs have been anticipated and whose anti-social desires pre-empted, thereby diminishing the subject, politics and citizenship. After all, we are not the sum of our past preferences but engage in dialogue and community to reach collective decisions and to cultivate 'a willingness to adopt the perspective of others' (Andrejevic, 2020, p. 19). On a similar theme, Zuboff (2015, 2019) argues that 'surveillance capitalism' (exemplified by Google's AdWords) uses personal data to target consumers more precisely, thereby exploiting and controlling human nature and damaging the social fabric. For Zuboff (2015, p. 86), this replaces the 'rule of law and the necessity of social trust as the basis for human communities with a new life-world of rewards and punishments, stimulus and response'. Also decrying loss of human autonomy, Couldry and Mejias (2019, p. 346) advance the notion of 'data

colonialism' as a commercially motivated form of data extraction that advances particular economic and governance interests. They argue that we must protect 'the integrity of the self as the entity that can make and reflect on choices in a complex world'; that this is 'essential to all Western liberal notions of freedom' (p. 345); and that it 'cannot be traded away without endangering the basic conditions of human autonomy' (p. 345). Bösel (2020) argues that the blackboxing of media-assisted, automatic affect regulation of individuals and populations might lead to serious disempowerment of moral and political subjects.

Although multiple critics decry the attack on human subjectivity, agency and autonomy, a cautionary note is needed when discussing impacts of any media text, system or technology on our beliefs, thoughts and actions. Historically, when new media technologies emerge, so do dystopian worries about their harms, alongside a desire to understand how to harness the new medium for social engineering. This was the case with the emergence of printing, radio, film, television, video games, the Internet and social media. Subsequent empirical studies tend to find that audience impacts are less pronounced, more difficult to interpret and more varied, with active rather than passive audiences, some of whom resist and reappropriate content rather than succumb to manipulation (Livingstone, 1998). Certainly, the empirical reality concerning false information online is messy. For instance, a representative online survey of Spanish adults finds (a) active users who are concerned about false news, are more aware of difficulties detecting it, and so make more effort to check news veracity; and (b) confident, passive users who feel less concerned about false news, view it as less difficult to detect, and so verify content less (Almenar et al., 2021). We acknowledge the long tradition of media research that engages with uses and gratifications and the politics of pleasure, finding active, oppositional and interactive audiences, as well as modes of resistance often mobilised by personal experience, and cultural differences and competences (for instance, Ang (1996), Morley (1992)). Indeed, agency can be evident even in encounters with algorithmic systems (Savolainen & Ruckenstein, 2022; Velkova & Kaun 2021). Ultimately, however, people (or audiences or users) do not get to design or set the rules on how such systems work, including technological systems of emotional optimisation. Furthermore, these algorithmic systems are abstract, opaque, personalised and of recent provenance, making it harder for people to develop an awareness of how they operate and whether they can be resisted or gamed.

Rather than starting from highly critical perspectives (as, for instance, adopted by those from the biopolitical or mass society camps), this book seeks out empirical insights and patterns to diagnose and evaluate the harms to the *civic body* from false information online. Most of this book focuses on false information incubated by digital platforms and especially social media (the globally dominant use case of emotional AI today). We appreciate that this is just part of the wider media system and that our focus neglects other areas such as the role of monopolistic, commercial legacy media systems and state-captured media systems in disseminating false and distorted information, but this is well-trodden ground in the political economy of media studies (McChesney, 2008; Túñez-López et al., 2022).

Throughout this book, we document strong currents seeking to optimise human emotion on behalf of platforms and influencers. Mindful to be even-handed rather than alarmist, we have sought out user-based studies to understand to what degree the agency of people is undermined across three component areas: false information, emotional information and microtargeting. Unfortunately, the studies are showing that most people are bad at recognising deception, especially in novel digital media forms (see Chap. 4), that emotions are viral online and that (some) people prefer news that bolsters their own worldview (see Chap. 5). The area least settled is microtargeting (see Chap. 6), and while this practice is on the rise, more user-based studies in this area are warranted.

One may rightly ask, who benefits from such emotional optimisation? As we will develop (especially in Chap. 2), it is the globally dominant digital platforms who ultimately profit from algorithmic optimisation of emotions, as this is driven by their business model that maximises user engagement. Justifiable and unjustifiable anger are fuelled by the algorithms, but so are joy, sadness and other emotions. As emotions drive user engagement, platforms can profit from any of these emotions; hence they are clear beneficiaries of this socio-technological arrangement. By contrast, the benefits to individuals and societies are mixed, affected by multiple contexts and accompanied by harms (for instance, proliferation of extremism, hate speech and false information online).

In seeking empirical insights into the causes and social consequences of globally dominant forms of emotional AI, we form a robust empirical base, from which we divine what may arise from more emergent forms of emotional AI. With interest in mediated emotion and datafied behaviour on the increase in biometric and in-the-wild contexts, we are interested in

the horizon line of emotionalised *civic bodies*. Ultimately, what should be done to socially prepare ourselves for the impact on the *civic body* of automated profiling, emotional AI and applications that simulate properties of empathy? We conclude by aligning with AI expert Stuart Russell who observes the pressing need to protect our 'mental integrity' from the global rise of AI and its profiling and predictive capacities (McStay, 2022; Russell, 2021). We argue that human *mental integrity* is not something to be lightly tossed aside in a technological, commercial, political or bureaucratic quest for something better, more efficient and optimised.

Aims, Approach and Argument

This is not a pessimistic book but one written in exceptional circumstances. People's growing concerns about false information online across the past decade have been spearheaded by governments and transnational bodies, producing political inquiries into, and legislation concerning, online fake news and disinformation. The COVID-19 pandemic also underscores harmful impacts of widespread, false information. Indeed, as the book progresses, we will suggest routes and means to address the problems we diagnose.

Given the rising tide of optimised emotion fuelling networked false information, we have six core aims:

1. To understand the significance of societal level profiling of human emotion through digital means.
2. To understand how the media and technological environment is (and will be) constructed to incubate false information, affect, emotions and user profiling and targeting.
3. To understand the economic and political incentives that drive the emotional profiling of society.
4. To understand the implications of societal profiling of emotion in a range of different societies and political arrangements.
5. To evaluate multi-stakeholder solutions to false information online proffered by supranational bodies, various sectors of society and multiple academic disciplines.
6. To generate principles necessary to strengthen the *civic body* while also looking forward to near-horizon futures.

Our approach deploys a multidisciplinary literature on contemporary misinformation, disinformation, digital marketing, digital advertising and emotional analytics. This scholarship is rooted in Communication Studies, embracing the disciplines of Advertising, Economics, History, Information Science, International Relations, Journalism, Law, Marketing, Media, Philosophy, Politics, Psychology, Public Relations, Science and Technology Studies and Sociology. Throughout, we have focused on studies with implications for the flow of false information throughout the *civic body*. There is growing evidence on the extent of false information on specific platforms and wider media, the techniques and pathways for its creation and spread, and how it may be tackled. There are also studies on its impacts, most of which detail behavioural impacts on platforms, national levels of concern and smaller-scale experiments into communication processes around false information.

As well as these academic studies, we draw on reports from national and supranational governmental bodies, regulators, non-governmental organisations, digital platforms, technology companies, think tanks (variously claiming to be independent, non-partisan, security focused, policy solutions oriented or technology based), research institutes, cybersecurity organisations, fact-checkers, journalists and, occasionally, bloggers. As the topic of this book is false information, it is pertinent to flag that some of these sources are focused on revealing and solving specific types of disinformation (for instance, emanating from some countries rather than others or deemed problematic for 'important' topics or groups of people). Even in countries such as the USA, where there is a considerable research effort into understanding disinformation, evidence is lacking on whether disinformation is about, or targeted at, people based on categories such as race and gender and whether it is effective (Thakur & Hankerson, 2021). Consequently, there are fewer finer-grained demographic insights into the phenomenon of false information online in this book. This would be important to address in future studies, as disinformation campaigns often rely on exploiting existing or historical narratives of discrimination to build credibility for the falsehoods being shared. Also, while we have cast our net widely geographically, some countries are well represented in terms of empirical studies, while others are lacking. This would also be important to address in future studies, as social media are global phenomena; as 'digital divides' are rapidly being breached in many countries, but digital literacies have not kept pace; and as digital means of engaging in information warfare and electoral influence can encompass any country.

Notwithstanding these empirical blind spots, we have aimed at a global approach that includes, but also looks beyond, the (comparatively) well-trodden ground of the USA.

As this book is primarily empirically based, most studies are on established use cases of optimised emotion, namely, social media and search engines, but we consider more emergent forms of emotional AI where there are supporting empirical studies. This includes substantive insights and trends emerging from the Emotional AI lab, where we are tracking cross-cultural developments in the fast-moving area of false information, datafied emotion and technological change. We flag here the difficulty of researching this emerging sector, not least because algorithms and datasets of the emotional AI industry remain largely off-limits to independent researchers, echoing the stance of dominant social media platforms.

The book has two parts. Part I (this chapter and Chaps. 2, 3, 4, 5 and 6) provides conceptual tools and contextual knowledge to understand the nature of false information online worldwide. We have devoted this chapter to introducing the metaphor of the *civic body*, to highlight the interconnectedness of bodies (individual and societal) and data about emotions. We also introduced the notion that this is leading to efforts to optimise emotions, this raising classic questions of exploitation and choice. Namely, who decides, or has a say in, what is optimal, optimisable, or optimised in a public sphere shaped by datafied emotion? Who benefits from these decisions, and who is harmed? What is lost along the way, and is there scope for resistance and reappropriation? In Chap. 2, we identify the two core incubators of false information to be the *economics of emotion* and the *politics of emotion*—namely, the optimisation of content for economic or political gain. We discuss, in Chap. 3, how different affective contexts worldwide fuel false information. This highlights the need to understand specificities of affective contexts and civic engagement, as well as their intersections with wider international information flows such as information warfare, ideological struggles and platforms' resources for content moderation.

Three chapters then each separately discuss a core component of contemporary false information online, covering false information (Chap. 4), affect and emotion (Chap. 5) and profiling and targeting (Chap. 6). In Chap. 4, we clarify the nature and forms of false information online (focusing on fake news and deepfakes, as well as wider misinformation), and its occurrence online (noting its prevalence, who spreads it and why). Observing that we are bad at recognising deception, especially new forms, we draw out implications for citizen-political communications, including

that rulers should not be deceptive, because of its erosion of social trust and democratic foundations.

Chapter 5 investigates the role of affect, emotion and moods as an energising force in opinion formation and decision-making that drives false information online across social media and news to potentially create post-truth environments. The chapter examines the resulting harms to the *civic body* by highlighting the challenges it poses to governmental efforts to manage their population's feelings and behaviour during the COVID-19 pandemic where uncertainty, anxiety, false information and conspiracy theories proliferated. As both mental harms (hate speech) and physical harms (reduced vaccine uptake) were evident, we conclude that we live in an informational environment that is sub-optimal for a healthy *civic body*.

In Chap. 6, we delve into profiling and targeting as the core means of delivering emotively charged, false information throughout the *civic body*, exploring this dynamic in political campaigning in democracies with different data protection regimes and digital literacies: the USA, UK and India. We find that political parties know increasingly more about their profiled, target audiences and adapt their campaigning accordingly. Worryingly, politicians and political parties have utilised platforms and built apps to mobilise electorates via delivery of inflammatory and deceptive messages targeted at profiled users. Less worryingly, the few empirical studies on profiling and microtargeting of voters find modest impacts on specific types of audience, and mixed findings regarding accuracy and prevalence of microtargeting. We conclude that more studies are needed on the effects of continuously refined profiling and targeting techniques on voting behaviour, especially as mobilisation of just a small sliver of the population (the persuadables) may generate decisive results. We also find that digital literacy, and awareness of profiling and microtargeting technologies for political purposes is uneven across the world, but where people are aware, most do not want it.

Building on this knowledge, Part II explores how we can strengthen the *civic body* across dominant and emergent uses of emotional AI. Opening this discussion, Chap. 7 identifies the following six civic harms arising from false information online. (1) It creates wrongly informed citizens that (2) in certain circumstances, for certain communities, may stay wrongly informed in digital echo chambers and (3) more widely, be emotionally provoked, leading to (4) contagion, where false, emotive information incubated online influences wider social media and mainstream news. Meanwhile, (5) profiling and microtargeting raise core democratic harms

comprising fragmentation of important national conversations; targeted suppression of voters; and even (potentially) undue influence over susceptible citizens. Also related (6) is the impact of false information in seeding distrust in important civic processes and institutions.

Chapter 8 evaluates solutions so far proffered by diverse stakeholders and by the multiple academic disciplines that embrace Communications Studies. It assesses seven solution areas: namely: (1) government action, (2) cybersecurity, (3) digital intermediaries/platforms, (4) advertisers, (5) professional political persuaders and public relations, (6) media organisations and (7) education. Noting that these are intrinsically difficult areas to solve individually, let alone in concert, and in every country, we conclude that such solutions merely tinker at the edges as they do not address a fundamental incubator for false information online: namely, the business model for social media platforms built on the *economics of emotion*.

The final chapter (Chap. 9) looks forward to near-horizon futures—an important angle given the rapid onset, scale and nature of false information online, and the rising tide of deployment of emotional analytics across all life contexts. While noting that false information, emotion, profiling and targeting are hardly new phenomena in citizen-political communications, we observe that the scale of contemporary profiling is unprecedented. We argue that a prime site of concern is the automated industrial psycho-physiological profiling of the *civic body* to understand affect and infer emotion for the purposes of changing behaviour. Exploring this, we look to near-horizon futures. This allows us to distil our core protective principle of protecting *mental integrity*. This is necessary to strengthen the *civic body* to withstand false information in a future where optimised emotion has become commonplace. How to have less of the harms and more of the positive elements is a difficult conundrum for policymakers. We hope that this book contributes to this ongoing global debate.

References

Alavi, H. S., Lalanne, D., & Rogers, Y. (2020). The five strands of living lab: A literature study of the evolution of living lab concepts in HCI. *ACM Transactions on Computer-Human Interaction, 27*(2), Article 10. https://doi.org/10.1145/3380958

Almenar, E., Aran-Ramspott, S., Suau, J., & Masip, P. (2021). Gender differences in tackling fake news: Different degrees of concern, but same problems. *Media and Communication, 9*(1), 229–238. https://doi.org/10.17645/mac.v9i1.3523

Andrejevic, M. (2020). *Automated media*. Routledge.

Ang, I. (1996). *Living room wars*. Routledge.

Article 19. (2021). *Emotional entanglement: China's emotion recognition market and its implications for human rights*. Retrieved April 13, 2022, from https://www.article19.org/wp-content/uploads/2021/01/ER-Tech-China-Report.pdf

Bakir, V. (2020). Psychological operations in digital political campaigns: Assessing Cambridge Analytica's psychographic profiling and targeting. *Frontiers in Political Communication, 5*(67). https://doi.org/10.3389/fcomm.2020.00067

Bakir, V., & McStay, A. (2018). Fake News and the economy of emotions: Problems, causes, solutions. *Digital Journalism, 6*(2), 154–175. https://doi.org/10.1080/21670811.2017.1345645

Bakir, V., & McStay, A. (2020). Empathic media, emotional AI and optimization of disinformation. In M. Boler & E. Davis (Eds.), *Affective politics of digital media* (pp. 263–279). Routledge.

Bandeira, L., & Braga, R. (2019). Brazil. In L. Bandeira, D. Barojan, R. Braga, J. L. Peñarredonda, & M. F. Pérez Argüello (Eds.), *Disinformation in democracies: Strengthening digital resilience in Latin America* (pp. 6–19). Atlantic Council. Retrieved April 13, 2022, from https://www.atlanticcouncil.org/in-depth-research-reports/report/disinformation-democracies-strengthening-digital-resilience-latin-america/

Bandeira, L., Barojan, D., Braga, R., Peñarredonda, J. L., & Pérez Argüello, M. F. (2019). *Disinformation in democracies: Strengthening digital resilience in Latin America* (pp. 20–29). Atlantic Council. Retrieved April 13, 2022, from https://www.atlanticcouncil.org/in-depth-research-reports/report/disinformation-democracies-strengthening-digital-resilience-latin-america/

Bolton, D. (2021). Targeting ontological security: Information warfare in the modern age. *Political Psychology, 42*(1), 127–142. https://doi.org/10.1111/pops.12691

Bösel, B. (2020). Affective media regulation: Or, how to counter the blackboxing of emotional life. In B. Bösel & S. Wiemer (Eds.), *Affective transformations: Politics-algorithms-media* (pp. 51–70). Meson Press.

Boyd, D. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (Ed.), *Networked self: Identity, community, and culture on social network sites* (pp. 39–58). Routledge.

Calhoun, C. (2010). The public sphere in the field of power. *Social Science History, 34*(3), 301–335. Retrieved June 17, 2022, from https://www.jstor.org/stable/40927615

Ceron, A., Curini, L., & Iacus, S. M. (2017). *Politics and big data: Nowcasting and forecasting elections with social media*. Routledge, Taylor and Francis.

Collins English Dictionary. (2017). *Fake news*. Retrieved April 13, 2022, from https://www.collinsdictionary.com/dictionary/english/fake-news

Couldry, N., & Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media, 20*(4), 336–349. https://doi.org/10.1177/1527476418796632

Currier, C., & Mackey, D. (2018, April 7). The rise of the Net Centre. *The Intercept*. https://theintercept.com/2018/04/07/guatemala-anti-corruption-trolls-smear-campaign/

Deleuze, G., & Guattari, F. (2000). *Anti-Oedipus: Capitalism and schizophrenia*. Athlone [Original work published 1972].

Dictionary.com. (2018). *Dictionary.com's 2018 word of the year is ….* Retrieved April 13, 2022, from https://www.dictionary.com/e/word-of-the-year/

Dignan, L., (2020, December 14). Amazon Halo review: A creepy yet useful fitness band. *ZDNet*. https://www.zdnet.com/article/amazon-halo-review-fitness-tracker-band/

Durahim, A. O., & Coşkun, M. (2015). #iamhappybecause: Gross national happiness through Twitter analysis and big data. *Technological Forecasting and Social Change, 99*, 92–105. https://doi.org/10.1016/j.techfore.2015.06.035

Edelman. (2021). *Edelman Trust Barometer*. Retrieved April 13, 2022, from https://www.edelman.com/sites/g/files/aatuss191/files/2021-03/2021%20Edelman%20Trust%20Barometer.pdf

Erwin, S. (2022, April 25). ManTech tracking foreign influence using open-source intelligence. *SpaceNews*. https://spacenews.com/mantech-tracking-foreign-influence-using-open-source-intelligence/

Eurobarometer. (2018). *Fake news and disinformation online*. Flash Eurobarometer 464. Retrieved April 13, 2022, from https://europa.eu/eurobarometer/surveys/detail/2183

Facebook. (2017, September 20). *Facebook, Inc. quarterly report* (10-Q No. 001–35551). Securities and Exchange Commission. Retrieved April 13, 2022, from https://www.sec.gov/Archives/edgar/data/1326801/000132680117000053/fb-09302017x10q.htm

Foucault, M. (1977). *Discipline and punish*. Penguin.

Garmin. (2022). *Body battery frequently asked questions*. Retrieved April 13, 2022, from https://support.garmin.com/en-GB/?faq=VOFJAsiXut9K19k1qEn5W5

Gorwa, R., & Ash, T. G. (2020). Democratic transparency in the platform society. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 286–312). Cambridge University Press.

Habermas, J. (1992). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Polity (Original work published 1962).

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society, 20*(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

Lazzarato, M. (2014). *Signs and machines: Capitalism and the promotion of subjectivity*. Semiotext(e).

Livingstone, S. (1998). Audience research at the crossroads: The 'implied audience' in media and cultural theory. *European Journal of Cultural Studies, 1*(2), 193–217. https://doi.org/10.1177/136754949800100203

Marcuse, H. (1991). *One-dimensional man: Studies in the ideology of advanced industrial society*. Routledge (Original work published 1964).

McChesney, R. W. (2008). *The political economy of media: Enduring issues, emerging dilemmas*. Monthly Review Press.

McStay, A. (2018). *Emotional AI: The rise of empathic media*. Sage.

McStay, A. (2019). Emotional AI and edtech: Serving the public good? *Learning, Media and Technology, 45*(3), 270–283. https://doi.org/10.1080/17439884.2020.1686016

McStay, A. (2022). *Automating empathy: When technologies claim to feel-into everyday life*. Oxford University Press.

McStay, A., & Rosner, G. (2021). Emotional artificial intelligence in children's toys and devices: Ethics, governance and practical remedies. *Big Data & Society*. https://doi.org/10.1177/2053951721994877

McStay, A., & Urquhart, L. (2022). In cars (are we really safest of all?): Interior sensing and emotional opacity. *International Review of Law, Computers & Technology*. Advance online publication. https://doi.org/10.1080/13600869.2021.2009181

Merriam-Webster. (2020). *Truthiness*. Retrieved April 13, 2022, from https://www.merriam-webster.com/dictionary/truthiness

Microsoft. (2021). *Digital Civility Index*. Retrieved April 13, 2022, from https://www.microsoft.com/en-us/online-safety/digital-civility

Morley, D. (1992). *Television, audiences and cultural studies*. Routledge.

Musolff, A. (2010). *Metaphor, nation and the holocaust: The concept of the body politic*. Routledge.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2018). *Reuters Institute digital news report 2018*. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/digital-news-report-2018.pdf

Newman, N., Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Reuters Institute digital news report 2020*. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Newman, N., Fletcher, R., Schulz, A., Andı, S., Robertson, C. T., & Nielsen, R. K. (2021). *Reuters Institute digital news report 2021*. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K. (2022). *Reuters Institute digital news report 2022*. Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Office for National Statistics. (2020). *Personal well-being in the UK: April 2019 to March 2020*. Retrieved April 13, 2022, from https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/april2019tomarch2020#:~:text=Average%20ratings%20of%20happiness%20in,the%20South%20East%20(1.6%25)

Oxford Dictionaries. (2016). *Word of the year: Post-truth*. Oxford Living Dictionaries. Retrieved April 13, 2022, from https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016

Picard, R. W. (1997). *Affective computing*. MIT.

Rabinow, P., & Rose, N. (2006). Biopower today. *BioSocieties, 1*(2), 195–217. https://doi.org/10.1017/S1745855206040014

Rice, R. E., Evans, S. K., Pearce, K. E., Sivunen, A., Vitak, J., & Treem, J. W. (2017). Organizational media affordances: Operationalization and associations with media use. *Journal of Communication, 67*(1), 106–130. https://doi.org/10.1111/jcom.12273

Richardson, S. (2020). Affective computing in the modern workplace. *Business Information Review, 37*(2), 78–85. https://doi.org/10.1177/0266382120930866

Rigby, S. H. (2012). The body politic in the social and political thought of Christine de Pizan (Abridged Version). Part I: Reciprocity, hierarchy and political authority. *Journal of Medieval and Humanistic Studies, 25*, 461–483. https://doi.org/10.4000/crm.13136

Rodríguez-Virgili, J., Serrano-Puche, J., & Beatriz Fernández, C. (2021). Digital disinformation and preventive actions: Perceptions of users from Argentina, Chile, and Spain. *Media and Communication, 9*(1), 323–337. https://doi.org/10.17645/mac.v9i1.3521

Rose, N. (2006). *The politics of life itself: Biomedicine, power, and subjectivity in the twenty-first century*. Princeton University Press.

Russell, S. (2021). The Reith Lectures - Stuart Russell - Living with Artificial Intelligence - AI: A future for humans. *BBC Sounds*. Retrieved April 13, 2022, from https://www.bbc.co.uk/sounds/play/m0012q21

Savolainen, L., & Ruckenstein, M. (2022). Dimensions of autonomy in human–algorithm relations. *New Media & Society*. https://doi.org/10.1177/14614448221100802

Smith, M. (2020, August 18). Facebook wanted to be a force for good in Myanmar. Now it is rejecting a request to help with a genocide investigation. *TIME*. Retrieved April 13, 2022, from https://time.com/5880118/myanmar-rohingya-genocide-facebook-gambia/

Stephens, M. (2013). *The geography of hate*. Retrieved April 13, 2022, from http://www.floatingsheep.org/2013/05/hatemap.html

Suarez-Villa, L. (2012). *Globalization and technocapitalism: The political economy of corporate power and technological domination*. Ashgate.

Tang, S., Willnat, L., & Zhang, H. (2021). Fake news, information overload, and the third-person effect in China. *Global Media and China, 6*(4), 492–507. https://doi.org/10.1177/20594364211047369

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data and Society, 4*(2). https://doi.org/10.1177/2053951717736335

Thakur, D., & Hankerson, D. L. (2021). *Facts and their discontents: A research agenda for online disinformation, race, and gender*. Center for Democracy & Technology. Retrieved June 13, 2022, from https://osf.io/3e8s5/

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

Túñez-López, M., Campos-Freire, F., & Rodríguez-Castro, M. (2022). *The values of public service media in the internet society*. Springer.

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

Velkova, J., & Kaun, A. (2021). Algorithmic resistance: Media practices and the politics of repair. *Information, Communication & Society, 24*(4), 523–540. https://doi.org/10.1080/1369118X.2019.1657162

Woolley, S. C., & Howard, P. N. (2018). *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press.

World Health Organisation. (2020). *Infodemic*. Retrieved April 13, 2022, from https://www.who.int/health-topics/infodemic#tab=tab_1

Wright, J. (2021). Suspect AI: Vibraimage, emotion recognition technology and algorithmic opacity. *Science, Technology & Society*, 1–20. Advance online publication. https://doi.org/10.1177/09717218211003411.

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology, 30*(1), 75–89. https://doi.org/10.1057/jit.2015.5

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

CHAPTER 2

Core Incubators of False Information Online

Introduction

False information is incubated across complex, interconnected communication and technological environments, imbricating individuals and society. Here, we introduce two key concepts. The first is the *economics of emotion*: namely, the optimisation of datafied emotional content for financial gain. Our second concept is the *politics of emotion*: namely, the optimisation of datafied emotional content for political gain. Optimising emotions, whether for financial or political gain, entails understanding people in terms of demography, interests and disposition; creation of content (by machines or by people) optimised to resonate with profiled individuals and groups; strategic ambition to elicit emotion to cause contagion; and recording of this datafied emotion expression, to feed into the next wave of info-contagion. We see the *economics of emotion* as the core incubator of false information online, as this stems from the business model of globally dominant digital platforms while also enabling the business model of digital influence mercenaries. However, the *politics of emotion* readily exploits the tools at its disposal. This chapter foregrounds these economic and political incubators of false information, leaving the messier discussion of impacts on audiences to later chapters.

V. Bakir, A. McStay, *Optimising Emotions, Incubating Falsehoods*,
https://doi.org/10.1007/978-3-031-13551-4_2

The Economics of Emotion

In this section, we explore the link between emotions, attention and revenue, and how these have been optimised to monetise deception. We illustrate this by focusing on the business models of digital influence mercenaries and of globally dominant social media and search engine companies. We end by reflecting on how economic decisions made by these digital platforms have destroyed the business model for news production, thereby further propelling false information online.

The Attention Economy and Optimised Emotion

Social media and search engine platforms make most of their revenue from selling online advertising. In 2020, advertising revenue made up 98% of Meta's total revenue and 86% of Twitter's, and more than 80% of Alphabet's revenue came from Google Ads (Alphabet Inc., 2020; Iqbal, 2022, January 11; Statista Research Department, 2022, February 18). To maximise how much advertising they can sell, these platforms' algorithms, interfaces and default settings are designed to maximally attract new users and keep users on their platform by holding their attention.

Alphabet-owned Google was the first company to create a standard market for online attention when, in the early 2000s, it launched Google Adwords (rebranded as Google Ads in 2018). Fully automated, Google Ads uses the PageRank algorithm to thematically match the offer and demand for advertising (namely, keywords searched for by users and targeted by marketers). It establishes advertising prices via automated asynchronous auctions for the keywords, pairing the bid amount with a 'Quality Score' assessment of marketers' ads, keywords and landing pages (higher quality ads potentially lead to lower prices and better ad positions). This allows Google to handle microtransactions unprofitable to traditional advertising agencies and to scale up its network, thereby helping to establish behavioural advertising as the dominant business model for websites (McStay, 2016).

Catchily termed 'surveillance capitalism' (Zuboff, 2019), this informational capitalism pioneered by Google (followed by Facebook) extracts as much data as possible about users. These digital platforms have built technical infrastructures and business models that link individual sites into a suite of services (like Google's many services) or ecosystem (as with

Facebook's 'Like' buttons scattered across the web), creating incentives for users to remain within the platform's ecosystem. They then turn that data into increasingly comprehensive 'profiles' or behavioural predictions (Gillespie, 2014). These profiles are monetised through internal use or sale to third parties in order to know, predict and modify behaviour. While Google uses keywords from search queries as its bidding criteria, Facebook uses demographic and behavioural information about its users based on their activity on Facebook and the wider web (Levy, 2020).

Attention is grabbed by emotions. Therefore, it was perhaps inevitable that informational capitalism, which transforms human subjectivity into usable quantitative representations (McStay, 2014), would embrace 'affective economies' where marketers seek to manage consumers by data-mining sentiment, as well as demographic and behavioural information (Andrejevic, 2011, p. 606). Arguably, across the entire neoliberal platform architecture, where all aspects of life are marketised, affect is a prime currency, as contemporary digital platforms are designed to commodify and manipulate formats for emotional expression (McStay, 2018; Stark & Crawford, 2015).

Facebook, for instance, maintains its attention economy by continually experimenting with algorithms, new data types and platform design, measuring users' actions to improve user interfaces (to increase users' time on, and engagement with, the site) and encouraging virality of posts. This includes design features to collect and manipulate emotional data about users' interests to fuel its advertising-based business model (Levy, 2020; McNamee, 2019; Stark, 2018). In 2009, Facebook introduced the 'Like' button (a 'thumbs up' emoji), forming one of its social plug-ins that can be placed on third-party websites, so allowing Facebook to track users across the web, providing it with a massive data source. Clicking 'Like' provides a crucial signal to help rank posts in a user's News Feed (renamed simply 'Feed' in 2022), also making the content appears in News Feeds of that user's friends. 'Getting 'Likes' incentivises users to habitually return to Facebook to see how many 'Likes' their posts received' (McNamee, 2019, p. 63). In 2010, the ability to 'Like' comments was added. In 2016, after several years of testing, Facebook rolled out its new Reaction Icons globally (users long-press the 'Like' button for an option to use one of five predefined emojis, namely, 'Love', 'Haha', 'Wow', 'Sad' or 'Angry'). Reactions were extended to comments in May 2017. In April 2020, responding to COVID-19, Facebook added a new Reaction for 'Care'.

Such data helps earn advertising revenue. It allows Facebook to understand users on a more emotional level, enabling personalisation of what content Facebook shows each user; and businesses can quickly tell which content resonated with target audiences.

While much as been written (and leaked) by, and about, Facebook, other social media platforms are similarly *emotional by design* in encouraging production of attention-grabbing content. All social media platforms use 'vanity metrics' that encourage users to return to, and engage with, the site (Rogers, 2018). Reaction buttons are used by platforms such as Meta-owned Instagram (which has eight quick Reactions), Twitter (its 'Like' button predates Facebook's) and Reddit (with 'upvotes' and 'downvotes'). Other platforms are structured so that only the most engaging material survives, such as threads on social media site 4Chan (Vaidhyanathan, 2018). Such affordances made 4Chan's environment an incubator for outlandish conspiracy theories that confirm users' preconceived biases through emotional appeals (Tuters & Hagen, 2020; Tuters et al., 2018). TikTok, which excels at engaging users (it was the most popular domain in 2021), injects a continuous fire hose of short videos into peoples' screens by guessing what users like based on their passive viewing habits, and signals such as likes, comments and who a user follows or blocks (Benton, 2022). An experiment by NewsGuard (a business that provides trust ratings for online content) in March 2022 on how TikTok funnels information about the Ukraine war finds that a new account that does nothing but scroll TikTok's algorithmically curated 'For You' page, watching in full videos about the war, results in analysts' feeds being almost exclusively populated with both accurate and false war-related content, with no distinction made between disinformation and reliable sources (Hern, 2022, March 21). Google-owned YouTube's recommendation algorithm promotes video clips that draw strong traffic: with news-related subjects, such results tend to be those with more extreme views (Larson, 2020). YouTube also financially rewards content producers based on engagement, which may also encourage production of inaccurate information that is more engaging, despite YouTube's efforts to reduce false information (Matamoros-Fernández et al., 2021). As the design of the algorithms and interfaces of globally dominant social media platforms maximise emotional engagement, we regard social media as a primary site of datafied emotion worldwide.

Monetising Emotion and Deception

The *economics of emotion* monetises deception online in two main ways. The first way is through a service contract from digital influence mercenaries to exploit social media's affordances to achieve a paying client's strategic influence objectives. The second way is by attracting users' attention through deceptive content and then selling that user attention to advertisers. We discuss both contract-based and advertising-based means of generating revenue in this section.

The practices of electioneering, lobbying and information warfare are increasingly outsourced to 'digital influence mercenaries', namely, paid individuals or companies with skills relevant to digital influence campaigns (Forest, 2022). For a fee from paying customers seeking to exert digital influence, these manipulation service providers are increasing and prospering, according to a report from the North Atlantic Treaty Organisation (NATO) (Bay et al., 2020). They add a key service within systems characterised by Howard (2020) as 'lie machines'. NATO draws attention to the immense scale of this increasingly global and interconnected industry, with hundreds of providers generating an infrastructure for social media manipulation software, generating fictitious accounts and providing mobile proxies. For instance, a European service provider will likely depend on Russian manipulation software and infrastructure providers who, in turn, use contractors from Asia for much of the manual labour (Bay et al., 2020).

The second way that the *economics of emotion* monetises deception online is by attracting users' attention through deceptive content (such as via fake news websites) and then leveraging visitor attention to sell as advertising opportunities to advertisers. A key underpinning mechanism has been use of cookie-based behaviourally targeted ads (Bakir & McStay, 2018, 2020). Online behavioural advertising underpinned by advertising technology ('adtech') tracks people's online behaviour (for instance, by planting cookies on users' computers to collect identifying information about a device and software thereon) and serves ads based on what people do online. While advertising spaces are ultimately owned by the web publisher (such as a news website), they are effectively outsourced and rented to entities called 'ad networks' (such as Google's DoubleClick), which offer advertisers a massive range of websites to exhibit their ads, allowing them to reach potentially large, profiled, audiences. Furthermore, programmatic techniques (termed 'programmatic' by the industry) have

allowed advertisers to *automatically* target consumers drawing on even wider varieties of data sources, based on algorithmically obtained metrics. The process often involves real-time bidding, where a potential advertiser (through automated methods) sees information about a person (such as type of device used, websites visited and search queries) and bids for the opportunity to display the ad to a person (Information Commissioners Office, 2019, June 20). This also provides opportunity to use automated means to create (and target) ads, personalising the ad for identified audiences. Such automation of the ad space buying process has resulted in ads for brands such as Honda, Thomson Reuters and Disney appearing on websites and YouTube videos promoting extremist ideologies such as neo-Nazi content. Similarly, if the user looks at a fake news site, the ads will appear there (Bakir & McStay, 2018). This programmatic arrangement produces a financial incentive for fake news provision, motivating deceptive content due to the fact that content can be highly attention-grabbing because it is not beholden to truth. Revenue is, in turn, generated by impressions (namely, how many times an ad is served and judged to have been seen) and click-throughs (the act of clicking on an ad to reach other content owned by the advertiser) (McStay, 2016).

Indeed, a vital driver of false information online is the desire to make money from *civic bodies* undergoing strong conflicted emotions. For instance, journalists traced a significant amount of the fake news upsurge on Facebook during the 2016 US presidential election campaign to students in Veles, Macedonia, who mostly created fake news stories for money rather than propaganda: their experiments with left-leaning content simply underperformed compared to pro-Trump content. In December 2019, an investigative press report highlighted how a small group of Israeli administrators commercially harvest Islamophobic hate from fake news posts on a network of Facebook pages from 21 far-right outlets in Australia, Austria, Canada, Israel, the UK and the USA, with a combined one million followers. This network funnels audiences to ten ad-heavy websites masquerading as news sites, thereby enabling the administrators to profit from the traffic (Knaus et al., 2019, December 5).

In the first study to systematise the auditing process of fake news revenue flows, its analysis (conducted in 2021) of 1044 unique, popular fake news sites (with millions of monthly visitors) and 1368 real news websites shows that well-known legitimate ad networks, such as Google, Index Exchange and AppNexus, still have a direct advertising relation with over 40% of these fake news websites and a re-seller advertising relation with

more than 60% of them. The entities who own fake news websites also operate other types of websites for entertainment, business and politics, indicating that owning a fake news website is part of a broader business operation (Papadogiannakis et al., 2022). Indeed, five years after commercially oriented fake news online was recognised as problematic, a report in 2021 estimates that American household brands still fund false information online by buying programmatic ads. It examined 7500 websites, finding that for every $2.16 in digital advertising revenue sent to legitimate newspapers, $1 goes to false information websites (Businesshala, 2021, August 5).

However, as of 2022, the behavioural advertising environment is undergoing significant change with leading browsers such as Apple's Safari now blocking by default third-party tracking. Arguably more significant given Google's centrality to the online advertising industry, Google and its browser Chrome no longer allow cookies and related identifiers to collect user data, in effect stopping selling of web ads targeted to *individual* users' browsing habits. The idea instead is to assemble groups of similar generalised interests. Under Google's 'Topics' programme, the Chrome browser determines top interests for that week based on browsing history, which Google says are kept for three weeks and then deleted. These 'Topics' are selected on a person's device without use of external servers. When a person visits a site of one of Google's client publishers, Topics are shared with the site and its advertising partners. Google also says that Topics will exclude sensitive categories, such as gender, religion or race (Goel, 2022). The key difference is that the specific sites visited by a person are no longer shared across the web with hard-to-identify third parties. Yet, general interest targeted advertising may still fund questionable publishers.

Moreover, as explained in Chap. 1, people worldwide are most concerned about false information on Facebook. As such, the role of Facebook's business model in incubating false information online warrants further scrutiny. Across the second decade of the twenty-first century, Facebook increasingly deployed machine-learning models to maximise user engagement. This created faster, more personalised, feedback loops that led to increasingly extreme, false content being shared. Central to this is Facebook's News Feed. This is a constantly updated, personally customised scroll of friends' photos, posts and links to news stories. It accounts for most of the time Facebook's users spend on the platform. Based on insights derived from in-app behaviour, and that

collected from usage of other apps and the web, the company sells that user attention to advertisers on Facebook and Instagram, accounting for nearly all of its $86 billion in revenue in 2020. A proprietary algorithm controls what appears in each user's News Feed, deciding a post's position based on predictions about each user's preferences and tendencies, ensuring that engaging material appears near the top. This is enabled by machine learning. Unlike traditional algorithms, which are hard-coded by engineers, Facebook's machine-learning algorithms train on input data to learn correlations within that data. The trained algorithm (known as a machine-learning model) then automates future decisions. As these algorithms could be trained to predict who would like or share what posts in a person's News Feed, this enabled Facebook to then give those posts more prominence. By mid-2016, Facebook had trained over a million machine-learning models, including models for image recognition, ad targeting and content moderation.

Internal Facebook documents leaked in 2021 shed light on this opaque, evolving process. In 2009, the News Feed ranking algorithm was relatively straightforward, prioritising signals such as 'Likes', clicks and comments to decide what to amplify. However, seeking to grow user engagement, the ranking algorithm became ever more sophisticated so that by 2021, it could take in over 10,000 different signals to predict a user's likelihood of engaging with a single post (Oremus et al., 2021, October 26). For instance, it considers users' friends, what kind of groups they joined, what pages they 'Liked', which advertisers have paid to target them, what types of stories drive conversation, how many long comments posts generate, whether a video is live or recorded, whether comments were made in plain text or with cartoon avatars, the computing load that each post requires and the strength of the user's Internet signal (Hagey & Horwitz, 2021, September 15; Merrill & Oremus, 2021, October 26; Oremus et al., 2021, October 26).

This increasing complexity of the News Feed ranking algorithm arose because of Facebook's desire for continued growth, given new competitors and shifting user behaviour, alongside the rise of machine learning that could predict what content would resonate with which user (Levy, 2020). In 2012, as Facebook was preparing for its initial public offering (the process of offering shares of a private corporation to the public in a new stock issuance), its goal was to increase revenue and take on Google, which then had most of the online advertising market (Hao, 2021, March 11). At the time, Facebook's News Feed ranking algorithm prioritised

'Likes', clicks and comments, and had led to publishers, brands and users learning how to craft 'clickbait' content with misleading, teaser headlines. Realising that users were growing wary of clickbait, Facebook recalibrated its algorithm in 2014 and 2015 to downgrade clickbait and focus on new metrics, such as amount of time spent on the site. In 2016, it added a ranking signal to measure a post's value based on the amount of time users spent with it. In 2017, Facebook added another ranking signal for video: completion rate (videos that keep people watching to the end are shown to more people) (Newberry, 2022). By 2017, under an internal point system used to measure its success, the algorithm assigned Reaction emoji ('Love', 'Haha', 'Wow', 'Sad' and 'Angry') five times the weight of a simple 'Like' (Oremus et al., 2021, October 26); and a significant comment, message, reshare or RSVP was assigned 30 times the weight of a 'Like'. Additional multipliers were added depending on whether interactions were between members of a group, friends or strangers (Hagey & Horwitz, 2021, September 15). These fed into Facebook's algorithmic change in 2018 that prioritised meaningful social interactions, namely, 'posts that spark conversations and meaningful interactions'. Posts from friends, family and Facebook groups were prioritised over organic content from pages. Brands would now need to earn more engagement to signal value to the algorithm. In 2019, Facebook prioritised high-quality, original video that keeps viewers watching longer than one minute. Facebook also prioritised content from 'close friends' (those that people engage with the most). In 2020, the algorithm also started to evaluate the credibility and quality of news articles to promote substantiated news rather than false information (Newberry, 2022).

Throughout, these algorithms were creating faster, more personalised feedback loops for tailoring each user's News Feed to increase engagement. The same algorithm produces different results for each user because it learns from their individual behaviours. Facebook found that, for the most politically oriented one million American users, nearly 90% of content that Facebook shows them is about politics and social issues. However, those groups also received the most misinformation, especially users associated with mostly right-leaning content, who were shown 1 misinformation post out of every 40 (Oremus et al., 2021, October 26). Indeed, Facebook's data scientists confirmed in 2019 that posts that sparked 'Angry', 'Wow' and 'Haha' Reaction emoji were disproportionately likely to include misinformation, toxicity and low-quality news (Merrill & Oremus, 2021, October 26). In giving outsize weight to emotional

reactions and posts that sparked interactions, this generated and consolidated communities sharing false, extremist information (Hao, 2021, March 11; Oremus et al., 2021, October 26). In the midst of the fake news furore following the 2016 US presidential election, and increasingly disturbing evidence of proliferation of extremist hate speech worldwide, the first downgrade to the Angry emoji weighting came in 2018, when Facebook cut it to four times the value of a 'Like', keeping the same weight for all other emojis. In April 2019, Facebook created another mechanism to demote content receiving disproportionately angry reactions. In 2020, in efforts to improve the friend ecosystem while reducing virality and its associated problems, Facebook cut the weight of all the Reactions to one and a half times that of a 'Like'. In September 2020, Facebook finally stopped using the Angry Reaction as a signal of what users wanted, cutting its weight to zero: its weight remained zero in 2021. As a result, users began to get less false information, less 'disturbing' content and less 'graphic violence', company data scientists found. At the same time, Facebook boosted 'Love' and 'Sad' to be worth two 'Likes' (Merrill & Oremus, 2021, October 26).

Increasing engagement on a platform is not inherently bad. As a Facebook staffer pointed out in the leaked documents, anger-generating posts might be essential to protest movements against corrupt regimes (Merrill & Oremus, 2021, October 26). However, increasing engagement with otherwise rare, extremist and false information is highly problematic. Furthermore, Facebook's own research also showed that content that is hateful, divisive and polarising was what kept people on its platform (Pelley, 2021, October 4). Ultimately, should such decisions about optimising emotions for financial gain that affect the *civic body* be left to platforms? Regardless of whether it is overall better or worse for the *civic body*, Facebook's (and Google's) optimisation of emotions and engaging content are a fundamental, but untransparent, part of their business model.

Destroying the Business Model for Real News and Fuelling False Information

Inadvertently, as well as fuelling fake news and extremist content, the economic decisions made by digital platforms have devastated journalism's business model in three ways, all of which harm the *civic body*.

Firstly, news sites are reliant on black boxed algorithms of globally dominant social media and search engine platforms to drive content to their site. This is because, increasingly, more people worldwide use search engines and social media as their main source of news. Globally, across 46 countries and five continents surveyed in 2021, only 25% of participants prefer to start their news journeys with a news website or app with most starting elsewhere such as social media, search, aggregators and email. Facebook was the most used social network for accessing news everywhere except Africa. In Africa, 60% used Facebook to access news (this was also the highest Facebook figure across the five continents), but 61% used Meta-owned WhatsApp (Newman et al., 2021). Such 'distributed discovery' means that news organisations have less control over how people find their news (Cornia et al., 2016). As such, appealing to platforms' algorithms became vital for the economic survival of news outlets. Accordingly, Facebook's algorithmic change in 2018 towards prioritising content from friends and family also hurt online publishers. In the first half of 2018, US-based outlet ABC News lost 12% of its traffic compared with the prior six months, BuzzFeed lost 13% and Breitbart lost 46%. To combat such audience loss, misleading headlines in mainstream news outlets seek clickbait audiences to generate Facebook shares for Internet traffic and advertising income. For instance, following the algorithmic change, Jonah Peretti, BuzzFeed's chief executive, wrote to Facebook that his staff felt 'pressure to make bad content' including material exploiting 'racial divisions', 'fad/junky science' and 'extremely disturbing news' (Hagey & Horwitz, 2021, September 15). Also, the only viable way for news to stay free to users is to attract massive audiences (to sell to advertisers). Hence, free sites will often be aggressively populist, such as British outlet, MailOnline, or those funded primarily for propaganda such as Breitbart or RT (formerly Russia Today) (Rapacioli, 2018, p. 92).

The second way in which digital platforms' economic decisions have devastated journalism's business model is in depriving news sites of advertising funds. Targeted digital advertising revenue is largely controlled by a Google-Meta duopoly. Hence, advertisers go to Google and Meta (rather than to news sites) for cheap, targeted advertising. This greatly reduces the amount of advertising income that news websites receive, making it far harder to fund their journalism, despite experimentation with paywalls, donations and digital subscriptions (McChesney, 2016; Nielsen & Fletcher, 2020; Nielsen & Ganter, 2017). Since the 2016 fake news furore captured public and political attention, Google and Facebook began to

voluntarily pay publishers around the world to sponsor news-related projects. However, the amounts have been determined by the platforms and, being voluntary payments, are subject to change depending on the platforms' strategic priorities (Benton, 2022).

Thirdly, it is hard to find consumers who will pay for news given that much information is available for free on social media. Across 40 countries surveyed by YouGov in February 2020, most people, especially young adults, do not pay for online news, a trend observable since social media's onset two decades prior and continuing today (Newman et al., 2022). For instance, across 2019, the proportion of people who paid for online news averaged only 26% in Nordic countries, 20% in the USA, 8% in Japan and 7% in the UK (Newman et al., 2020, pp. 21–22).

At best, this contraction of income flowing into news organisations damages product quality. It increases newsrooms' reliance on (free) press releases rather than (expensive) original reporting (Davies, 2008). Rather than fostering in-depth journalism, it leads to newsroom strategies that focus on the immediacy of 'breaking' news events such as sensational crime and disasters, seeking audience engagement and responding to real-time feedback from analytics companies (Usher, 2018). It also leads to contractions in news provision, generating local news deserts (Curran, 2022; Starr, 2020). In countries that do not subsidise their public service news, this damages overall news quality (McChesney, 2016). At worst, people bypass news sites altogether, getting all their information from social media, a situation that can greatly damage the *civic body*.

Such situations are particularly problematic in poor countries where Facebook has been incentivising use of its platform. Through Internet.org, Facebook partners with telecommunications companies who, through 'zero-rating' policies, make several stripped-down web services (including Facebook) freely available through a mobile app (without tapping into users' mobile data plans). Most charitably, this fulfils the social mission of Meta's Chief Executive Officer, Mark Zuckerberg, of 'connecting the world' where Internet penetration is low (Zuckerberg, 2013, August 21). Less charitably, this helps grow Facebook's international user base while damaging competitors, potentially inspiring future paid use of Facebook when users' financial situation improves. First launched in 2013, and renamed 'Free Basics' in 2015, by 2019, it was available in 65 countries, including 30 African nations. This makes many people in poorer countries entirely reliant on Facebook for information access, eschewing paid-for content (including reputable news outlets). Unfortunately, many of these

countries are characterised by weak public sphere institutions; lack government regulation to protect and educate citizens about false, emotive information; and are in countries where Facebook has been slower to introduce content moderating tools (Hempel, 2018, May 17; Nothias, 2020).

To summarise, the *economics of emotion* finances fake news websites and extremist content online. It involves both contract-based and advertising-based means of generating revenue across digital platforms. It leads to more emotionalised presentation of online news; greatly damages the economic viability, and hence quality, of news; and leads to many users relying on free, but false, information online. As such, the *economics of emotion* is an important incubator of false information online. So too is the *politics of emotion*, the subject of the next section.

The Politics of Emotion

The *politics of emotion* is the phenomenon of optimising datafied emotional content for political gain (Bakir & McStay, 2020). Appealing to a *civic body*'s emotions are long-standing practices among politicians seeking election and nation states conducting information warfare. As we show below, such practices are super-charged in the digital media ecology, exploiting digital platforms' profiling and optimisation affordances. Depending on their own priorities, dominant digital platforms may (or may not) intervene to moderate harmful content.

While for decades, opinion polling allowed political parties to merge broad demographic data with psychographic insights on how to craft emotionally resonant messages, the targeting is now fiercely more granular (see Chaps. 5 and 6). As noted earlier, because of the *economics of emotion*, social media platforms favour emotionality: mainstream platforms such as Facebook surface posts that are emotionally engaging rather than neutral and niche platforms, such as 4Chan, encourage offensive content to get noticed. Indeed, the *politics of emotion* led to complaints to Facebook by major political parties in Poland and Spain in 2019 (Hagey & Horwitz, 2021, September 15). A leaked Facebook report states that the political parties feel strongly that Facebook's algorithmic change (prioritising Meaningful Social Interactions) 'forced them to skew negative in their communications on Facebook… leading them into more extreme policy positions' (Pelley, 2021, October 4). Facebook researchers wrote in their internal report that Polish parties complained that it made 'political debate on the platform nastier' because the parties were now incentivised to

attract reshares, achieved by tapping into anger, with similar complaints from political parties in Taiwan and India (Hagey & Horwitz, 2021, September 15).

False political information on social media makes us angry and less analytical. Barfar's (2019) analysis of user comments on nearly 2100 political posts from popular sources of political disinformation on US Facebook in 2018 finds that compared to true news, political disinformation received significantly less analytic responses and is filled with more anger and incivility (whereas true news elicits more anxiety). This tallies with research by Facebook's data scientists, discussed earlier, which confirmed that posts sparking the 'Angry' reaction emoji were disproportionately likely to include misinformation, toxicity and low-quality news (Merrill & Oremus, 2021, October 26). Unsurprisingly, then, hate speech features in political disinformation worldwide. For instance, the 2019 Indonesian national elections were characterised by misinformation, populism and rampant use of religion, racial and divisive issues by their followers (Neyazi & Muhtadi, 2021). Instructively, leaked Facebook documents from February 2019, not long before India's General Election, show how a dummy account set up to understand the experience of a new, young, female adult user in Facebook's largest market was flooded with pro-Modi propaganda and anti-Muslim hate speech. Although Hindi and Bengali are respectively the fourth- and seventh-most spoken languages worldwide, Facebook only introduced hate speech classifiers in Hindi in 2018 and Bengali in 2020; systems for detecting violence and incitement in Hindi and Bengali were not added until 2021 (Zakrzewski et al., 2021, October 24).

It is not just domestic actors that exploit the *politics of emotion* but international actors strategically applying power in the information domain. Acts of warfare themselves, such as invading another country, unleash raw emotions. When seeded with disinformation, the emotional charge resists debunking or fact-checking, a phenomenon long observed by military historians (Rid, 2021). There are also more subtle ways of applying power in the information domain. Depending on contexts, intergovernmental military alliance, NATO, has numerous terms for this including information warfare, psychological operations, influence operations, strategic communications, computer Network operations and military deception. Russia takes a more integrated view of informational power, covering the full range of practices above, while also applying 'information-psychological warfare' to both wartime and peacetime conflicts (Giles, 2016, p. 9). According to a NATO report, Russia uses

information warfare deploying technical, cognitive and emotional facets to covertly introject distorted facts and 'emotional impressions' on policy-makers in attempts to influence decisions (Giles, 2016, p. 21). Tactics include targeting politicians on social media and the comments sections of major online news outlets and manipulating polls in Western media, for instance, to skew survey results on whether sanctions against Russia were supported following its invasion of Ukraine in 2022 (The Guardian, 2022, May 1). Countries engaging in information warfare also seek to generate ontological insecurity, or intense anxiety, using covert means to attack citizens' sense of being. Examples include destabilising the national narrative, or sense of home, that individuals are embedded within, and fracturing their sense of self by turning factions upon each other (Bolton, 2021). Race-baiting disinformation is an old tactic used in information warfare, deployed, for instance, by the USSR's main security agency, the KGB, in the 1960s to stir up trouble in Black and Jewish communities in American cities (Rid, 2021). As Bolton (2021, p. 134) puts it, such tactics subvert, 'existing frameworks for managing anxiety around existential questions: eroding certainty over where threats reside (existence), undermining the stability of established belief systems (meaninglessness), and curbing positive subgroup recognition (condemnation)'.

A study by the Australian Strategic Policy Institute (a government-funded defence and strategic policy think tank) into elections and referenda held between 2016 and 2019 in 97 free or partly free countries (as defined by Freedom House, a non-profit, majority US government-funded research and advocacy organisation) finds evidence for cyber-enabled foreign interference targeting 20 countries. It largely (allegedly) emanates from Russia and China, but occasionally also Iran, the UK and Venezuela. The foreign interference targeted voting infrastructure in five countries (Colombia, Finland, Indonesia, Ukraine and the USA) and voter turnout in North Macedonia and the USA. Across ten countries (France, Israel, Italy, Malta, the Netherlands, North Macedonia, Spain, Taiwan, Ukraine and the USA), the interference also targeted the wider information environment, for instance, creating and spreading disinformation to undermine a candidate and creating fake personae to provide inflammatory commentary on divisive issues. There were also longer-term efforts to erode public trust in governments, political leadership and public institutions identifiable in ten countries (Australia, Brazil, the Czech Republic, Germany, Montenegro, Norway, the Netherlands, Singapore, Ukraine and the USA) (Hanson et al., 2019). Facebook also regularly reports on

networks of accounts, pages and groups engaged in 'coordinated inauthentic behaviour' targeted at domestic audiences (for instance, in the USA, Georgia, Myanmar and Mauritania) and international audiences (for instance, emanating from Russia and Iran) (Facebook, 2020, April, 2018, August 28).

Deliberate deceptions, whether originating from domestic or international political actors, or non-state actors and digital influence mercenaries, are often recirculated as misinformation, exacerbated by the technological affordances of dominant media systems. Even a global behemoth like Facebook has limits on what resources it will devote to tackling false information online. While long prioritising international growth, Facebook has not safeguarded this by employing sufficient people speaking local languages, thereby damaging its ability to moderate content worldwide (Levy, 2020). A newspaper investigation examining internal documentation leaked by ex-Facebook data scientist, Sophie Zhang, shows how Facebook allows major abuses of its platform in poor, small, non-Western countries while prioritising addressing abuses that attract media attention and negative public relations or that affect the USA and other wealthy countries (where its average revenue per user is higher). For instance, Facebook acted quickly to address disinformation affecting the USA, Poland, South Korea and Taiwan while moving slowly or not at all on cases in Afghanistan, Albania, Iraq, Mexico, Mongolia, Tunisia and much of Latin America (Wong, 2021, April 12). Leaked Facebook documents show that in 2020, Facebook employees and contractors spent over 3.2 million hours searching out, labelling or taking down information that the company concluded was false or misleading, but only 13% of those hours were spent on content from outside the USA (Scheck et al., 2021, September 16) although 90% of Facebook's monthly active users are outside North America (Facebook, 2021).

Like other social media platforms, Facebook's content moderation on hate speech has also proven inadequate to protecting the *civic body*. Facebook reports that its proactive detection methods for hate speech have improved following advances in AI (where automated systems are trained on hundreds of thousands of different examples of violating content and common attacks). However, hate speech online can be hard to identify as it evolves rapidly, with code words and in-jokes for racial and gendered slurs (Ribeiro et al., 2018). For instance, use of triple parentheses around the hate target's name is an anti-Jewish slur on Twitter (Duarte et al., 2017). In the Philippines, gendered online disinformation about Senator Leila de Lima uses the term 'saba queen' rather than her name

(referencing rumours about her having an affair); and hashtags are used to similar effect (Judson et al., 2020, October). Given such difficulties, Facebook also relies on human reviewers to assess nuance and context. In 2020, Zuckerberg stated that Facebook removes 94% of the hate speech it finds before a human reports it (Lima, 2021, October 26). However, a leaked internal study from Facebook in 2021 states: 'we estimate that we may action as little as 3-5% of hate and about 6-tenths of 1% of V & I [violence and incitement] on Facebook despite being the best in the world at it' (Pelley, 2021, October 4).

In some countries, worst-case scenarios prevail. In 2018, the United Nations called out Facebook for allowing hateful posts amplifying ethnic tensions between Buddhist nationalists and Muslim minorities in Myanmar (formerly, Burma). This led to over 9000 Rohingya Muslims killed across 2017 and 800,000 fleeing to Bangladesh to escape genocide in 2018 (Hempel, 2018, May 17). Myanmar was a fragile democracy, having emerged from five decades of military rule in 2011. In 2012, only 1.1% of the population used the Internet, and few had telephones, as the military junta had kept citizens isolated. In 2013, when a quasi-civilian government oversaw telecommunications deregulation, SIM cards became affordable. By 2016, nearly half the population had mobile phone subscriptions (mostly smartphones), and Facebook's app went viral as Myanmar's mobile phone operators adopted zero-rating policies under Free Basics. Yet, Facebook employed only four Burmese speakers as content moderators in 2015, in a digitally illiterate population. Most people in Myanmar do not speak English, yet Facebook's system for reporting problematic posts was then only in English (Stecklow, 2018, August 15). Furthermore, the Burmese language does not always use international standard Unicode online but a unique font difficult for Facebook's system to read (Levy, 2020, p. 437). Facebook's investigation into their role in the genocide found that seemingly independent news, entertainment, beauty and lifestyle pages were linked to the Myanmar military, and celebrity and entertainment accounts pushed military propaganda. Facebook's response across 2018 was to take down the pages, groups and accounts of military officials, organisations and networks that sought to incite the violence (Facebook, 2018, August 28). Yet, in August 2018, Reuters found over 1000 posts, comments, images and videos attacking the Rohingya or other Myanmar Muslims on Facebook in the previous week, some urging extermination (Stecklow, 2018, August 15).

While Facebook claims to have learned lessons from Myanmar, a similar situation emerged in Ethiopia in 2020, where armed groups associated with Ethiopia's government and state media posted inciting viral comments on Facebook against the Tigrayan minority, some calling for Tigrayans to be exterminated. Violence escalated when the government launched an attack on the Tigray capital, Mekelle. In the context of minimal press freedoms, low Internet penetration and fewer Facebook users in Ethiopia (only 6.7 million in 2020 in a population of 115 million) (Internet World Stats, 2021), political ethnic issues dominated discourse on popular Facebook sites in preceding years (Skjerdal & Gebru, 2020). Once again, leaked Facebook internal communications show it did not have enough employees speaking relevant languages to monitor the situation, and AI systems that form the backbone of Facebook's enforcement do not cover most languages used on the site. Facebook claims to have since increased its review capacity in Ethiopian languages and improved its automated systems to stop harmful content (Scheck et al., 2021, September 16; Simonite, 2021, October 25).

Clearly, the *politics of emotion*, and the false information that it propels, is observable worldwide. In practice, democracies vary in their vulnerability to false information based on factors such as extent of domestic and external manipulation, the digital literacy of its citizens and their access to trustworthy information, and willingness of global digital platforms to engage in resource-intensive content moderation (as well as other factors, discussed in Chap. 3).

Conclusion

To explain what incubates contemporary false information in civic bodies, we introduced two concepts. The *economics of emotion* delineates the optimisation of datafied emotional content for financial gain. We explored how it finances digital influence mercenaries, fake news websites and extremist content online; how it leads to more emotionalised presentation of online news; how it greatly damages the economic viability and quality of news; and how it leads to many people relying on free, but false, information online. Our concept of the *politics of emotion* (the phenomenon of optimising datafied emotional content for political gain) demonstrates how the long-standing practice of crafting emotive messages to engage target audiences is super-charged in contemporary informational environments. This exposes citizens to emotive, false information via behavioural targeting on social media platforms, exploited by domestic and international political

actors. This generates affective feedback loops, ranging from intense anxiety to hatred of the other, that are not adequately dealt with by digital platforms' content moderation. Given the varied, and complicated, global picture on vulnerability to false information, in the next chapter we illustrate how the economics and politics of emotion fuel false information in different democracies and under different affective contexts.

References

Alphabet Inc. (2020). *Form 10-K for the fiscal year ended December 31, 2020.* Retrieved April 13, 2022, from https://www.sec.gov/Archives/edgar/data/1652044/000165204421000010/goog-20201231.htm

Andrejevic, M. (2011). The work that affective economics does. *Cultural Studies, 25*(4–5), 604–620. https://doi.org/10.1080/09502386.2011.600551

Bakir, V., & McStay, A. (2018). Fake News and the economy of emotions: Problems, causes, solutions. *Digital Journalism, 6*(2), 154–175. https://doi.org/10.1080/21670811.2017.1345645

Bakir, V., & McStay, A. (2020). Empathic media, emotional AI and optimization of disinformation. In M. Boler & E. Davis (Eds.), *Affective politics of digital media* (pp. 263–279). Routledge.

Barfar, A. (2019). Cognitive and affective responses to political disinformation in Facebook. *Computers in Human Behavior, 101*, 173–179. https://doi.org/10.1016/j.chb.2019.07.026

Bay, S., Dek, A., Dek, I., & Fredheim, R. (2020). *Social media manipulation 2020. How social media companies are failing to combat inauthentic behaviour online.* NATO Strategic Communications Centre of Excellence. Retrieved April 13, 2022, from https://stratcomcoe.org/publications/social-media-manipulation-report-2020/21

Benton, J. (2022, June 16). Facebook looks ready to divorce the news industry, and I doubt couples counseling will help. *Nieman Lab.* Retrieved June 20, 2022, from https://www.niemanlab.org/2022/06/facebook-looks-ready-to-divorce-the-news-industry-and-i-doubt-couples-counseling-will-help/

Bolton, D. (2021). Targeting ontological security: Information warfare in the modern age. *Political Psychology, 42*(1), 127–142. https://doi.org/10.1111/pops.12691

Businesshala. (2021, August 5). *Big brands are funneling as much as $2.6 billion into misinformation websites per year.* Retrieved April 13, 2022, from https://businesshala.com/big-brands-are-funneling-as-much-as-2-6-billion-into-misinformation-websites-per-year/

Cornia, A., Sehl, A., & Nielsen, R. K. (2016). *Private sector media and digital news.* Reuters Institute for the Study of Journalism. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/research/

files/Cornia%2520-%2520Private%2520Sector%2520Media%2520and%2520Digital%2520News%2520FINAL.pdf

Curran, J. (2022). An end to futility: A modest proposal. In J. Zylinska (Ed.), *The future of media* (pp. 45–58). Goldsmiths Press.

Davies, N. (2008). *Flat earth news: An award-winning reporter exposes falsehood, distortion and propaganda in the global media*. Chatto & Windus.

Duarte, N., Llanso, E., & Loup, A. (2017). Mixed messages? *The limits of automated social media content analysis*. Centre for Democracy and Technology. Retrieved April 13, 2022, from https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf

Facebook. (2018, August 28). *Removing Myanmar military officials from Facebook*. Retrieved April 13, 2022, from https://about.fb.com/news/2018/08/removing-myanmar-officials/

Facebook. (2020). *April 2020 coordinated inauthentic behavior report*. Retrieved April 13, 2022, from https://about.fb.com/wp-content/uploads/2020/05/April-2020-CIB-Report.pdf

Facebook. (2021). *United States Securities and Exchange Commission. Form 10-Q. Facebook Inc*. Retrieved April 13, 2022, from https://www.sec.gov/ix?doc=/Archives/edgar/data/1326801/000132680121000049/fb-20210630.htm

Forest, J. F. (2022). *Digital influence mercenaries. Profits and power through information warfare*. Naval Institute Press.

Giles, K. (2016). *Handbook of Russian information warfare* (Fellowship Monograph Series, No. 9). Research Division NATO Defense College. Retrieved April 13, 2022, from https://krypt3ia.files.wordpress.com/2016/12/fm_9.pdf

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies* (pp. 167–194). MIT Press.

Goel, V. (2022). Get to know the new Topics API for privacy sandbox. *Google*. Retrieved April 13, 2022, from https://blog.google/products/chrome/get-know-new-topics-api-privacy-sandbox/

Hagey, K., & Horwitz, J. (2021, September 15). Facebook tried to make its platform a healthier place. It got angrier instead. *Wall Street Journal*. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215?mod=article_inline

Hanson, F., O'Connor, S., Walker, M., & Courtois, L. (2019). *Hacking democracies: Cataloguing cyber-enabled attacks on elections* (Policy Brief 16). Australian Strategic Policy Institute. Retrieved April 13, 2022, from https://www.aspi.org.au/report/hacking-democracies

Hao, K. (2021, March 11). How Facebook got addicted to spreading misinformation. *MIT Technology Review*. Retrieved April 13, 2022, from https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/

Hempel, J. (2018, May 17). What happened to Facebook's grand plan to wire the world? *Wired*. https://www.wired.com/story/what-happened-to-facebooks-grand-plan-to-wire-the-world/

Hern, A. (2022, March 21). TikTok algorithm directs users to fake news about Ukraine war, study says. *The Guardian*. https://www.theguardian.com/technology/2022/mar/21/tiktok-algorithm-directs-users-to-fake-news-about-ukraine-war-study-says

Howard, P. N. (2020). *Lie machines: How to save democracy from troll armies, deceitful robots, junk news operations, and political operatives.* Yale University Press.

Information Commissioners Office. (2019, June 20). *Update report into adtech and real time bidding.* Retrieved April 13, 2022, from https://ico.org.uk/media/about-the-ico/documents/2615156/adtech-real-time-bidding-report-201906.pdf

Internet World Stats. (2021). *Internet users statistics for Africa*. Retrieved April 13, 2022, from https://www.internetworldstats.com/stats1.htm

Iqbal. (2022, January 11). *Twitter revenue and usage statistics.* Retrieved April 13, 2022, from https://www.businessofapps.com/data/twitter-statistics/

Judson, E., Atay, A., Krasodomski-Jones, A., Lasko-Skinner, R., & Smith, J. (2020, October). *The contours of state-aligned gendered disinformation online*. Demos. Retrieved June 23 2022, from https://demos.co.uk/project/engendering-hate-the-contours-of-state-aligned-gendered-disinformation-online/

Knaus, C., McGowan, M., Evershed, N., & Holmes, O. (2019, December 5). Inside the hate factory: How Facebook fuels far-right profit. *The Guardian*. https://www.theguardian.com/australia-news/2019/dec/06/inside-the-hate-factory-how-facebook-fuels-far-right-profit

Larson, R. (2020). *Bit tyrants: The political economy of Silicon Valley.* Haymarket Books.

Levy, S. (2020). *Facebook: The inside story.* Penguin, Random House.

Lima, C. (2021, October 26). A whistleblower's power: Key takeaways from the Facebook Papers. *Washington Post*. https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/

Matamoros-Fernández, A., Gray, J. E., Bartolo, L., Burgess, J., & Suzor, N. (2021). What's "up next"? Investigating algorithmic recommendations on YouTube across issues and over time. *Media and Communication, 9*(4), 234–249. https://doi.org/10.17645/mac.v9i4.4184

McChesney, R. W. (2016). Journalism is dead! Long live journalism?: Why democratic societies will need to subsidise future news production. *Journal of Media Business Studies, 13*(3), 128–135. https://doi.org/10.1080/16522354.2016.1184919

McNamee, R. (2019). *Zucked: Waking up to the Facebook catastrophe*. Harper Collins

McStay, A. (2014). *Privacy and philosophy: New media and affective protocol.* Peter Lang.

McStay, A. (2016). *Privacy and the media*. Sage.

McStay, A. (2018). *Emotional AI: The rise of empathic media*. Sage.

Merrill, J. B., & Oremus, W. (2021, October 26). Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation. *Washington Post*. https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/

Newberry, C. (2022). *How the Facebook algorithm works in 2022 and how to make it work for you*. Retrieved April 13, 2022, from https://blog.hootsuite.com/facebook-algorithm/

Newman, N., Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Reuters Institute digital news report 2020*. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Newman, N., Fletcher, R., Schulz, A., Andı, S., Robertson, C. T., & Nielsen, R. K. (2021). *Reuters Institute digital news report 2021*. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K. (2022). *Reuters Institute digital news report 2022*. Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Neyazi, T., & Muhtadi, B. (2021). Selective belief: How partisanship drives belief in misinformation. *International Journal of Communication, 15*, 1286–1308. https://ijoc.org/index.php/ijoc/article/viewFile/15477/3382

Nielsen, R. K., & Fletcher, R. (2020). Democratic creative destruction? The effect of a changing media landscape on democracy. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 139–162). Cambridge University Press.

Nielsen, R. K., & Ganter, S. A. (2017). Dealing with digital intermediaries: A case study of the relations between publishers and platforms. *New Media & Society, 20*(4), 1600–1617. https://doi.org/10.1177/1461444817701318

Nothias, T. (2020). Access granted: Facebook's free basics in Africa. *Media, Culture & Society, 42*(3), 329–348. https://doi.org/10.1177/0163443719890530

Oremus, W., Alcantara, C., Merrill, J. B., & Galocha, A. (2021, October 26). Facebook under fire: How Facebook shapes your feed. *Washington Post*. https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/

Papadogiannakis, E., Papadopoulos, P., Markatos, E. P., & Kourtellis, N. (2022). *Who funds misinformation? A systematic analysis of the ad-related profit routines of fake news sites*. Preprint retrieved from https://doi.org/10.48550/arXiv.2202.05079

Pelley, S. (2021, October 4). Whistleblower: Facebook is misleading the public on progress against hate speech, violence, misinformation. *60 Minutes*. https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-misinformation-public-60-minutes-2021-10-03/

Rapacioli, P. (2018). *Good Sweden, bad Sweden: The use and abuse of Swedish values in a post-truth world*. Volante.

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. F., & Meira, W., Jr. (2018). Characterizing and detecting hateful users on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1). Retrieved April 13, 2022 from https://ojs.aaai.org/index.php/ICWSM/article/view/15057

Rid, T. (2021). *Active measures: The secret history of disinformation and political warfare*. Profile Books.

Rogers, R. (2018). Otherwise engaged: Social media from vanity metrics to critical analytics. *International Journal of Communication, 12*(732942), 450–472. Retrieved April 13, 2022, from https://ijoc.org/index.php/ijoc/article/view/6407/2248

Scheck, J., Purnell, N., & Horwitz, J. (2021, September 16). Facebook employees flag drug cartels and human traffickers. The company's response is weak, documents show. *Wall Street Journal*. https://www.wsj.com/articles/the-facebook-files-11631713039?mod=bigtop-breadcrumb

Simonite, T. (2021, October 25). Facebook is everywhere; its moderation is nowhere close. *Wired*. https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/

Skjerdal, T., & Gebru, S. (2020). Not quite an echo chamber: Ethnic debate on Ethiopian Facebook pages during times of unrest. *Media, Culture & Society, 42*(3), 365–379. https://doi.org/10.1177/0163443719895197

Stark, L. (2018). Algorithmic psychometrics and the scalable subject. *Social Studies of Science, 48*(2), 204–231. https://doi.org/10.1177/0306312718772094

Stark, L., & Crawford, J. (2015). The conservatism of emoji: Work, affect, and communication. *Social Media + Society, 1*(2), 1–11. https://doi.org/10.1177/2056305115604853

Starr, P. (2020). The flooded zone: How we became more vulnerable to disinformation in the digital era. In W. L. Bennett & S. Livingston (Eds.), *The disinformation age: Politics, technology and disruptive communication in the information age* (pp. 67–91). Cambridge University Press. https://doi.org/10.1017/9781108914628

Statista Research Department. (2022, February 18). *Meta: advertising revenue worldwide 2009–2021*. Retrieved April 13, 2022, from https://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/#:~:text=In%202021%2C%20Meta%20(formerly%20Facebook,of%20the%20social%20network's%20revenue

Stecklow, S. (2018, August 15). Inside Facebook's Myanmar operation. Hatebook. *Reuters*. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

The Guardian. (2022, May 1). 'Troll factory' spreading Russian pro-war lies online, says UK. *The Guardian*. https://www.theguardian.com/world/2022/may/01/troll-factory-spreading-russian-pro-war-lies-online-says-uk?CMP=Share_iOSApp_Other

Tuters, M., & Hagen, S. (2020). (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society, 22*(12), 2218–2237. https://doi.org/10.1177/1461444819888746

Tuters, M., Jokubauskaitė, E., & Bach, D. (2018). Post-truth protest: How 4chan cooked up the Pizzagate bullshit. *M/C Journal, 21*(3). https://doi.org/10.5204/mcj.1422

Usher, N. (2018). Breaking news production processes in US metropolitan newspapers: Immediacy and journalistic authority. *Journalism, 19*(1), 21–36. https://doi.org/10.1177/1464884916689151

Vaidhyanathan, S. (2018). *Antisocial media: How Facebook disconnects us and undermines democracy*. Oxford University Press.

Wong, J. C. (2021, April 12). Revealed: The Facebook loophole that lets world leaders deceive and harass their citizens. *The Guardian*. https://www.theguardian.com/technology/2021/apr/12/facebook-loophole-state-backed-manipulation

Zakrzewski, C., De Vynck, G., Masih, N., & Mahtani, S. (2021, October 24). How Facebook neglected the rest of the world, fueling hate speech and violence in India. *Washington Post*. https://www.washingtonpost.com/technology/2021/10/24/india-facebook-misinformation-hate-speech/

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

Zuckerberg, M. (2013, August 21). Is connectivity a human right? *Meta*. Retrieved April 13, 2022, from https://about.fb.com/news/2013/08/mark-zuckerberg-is-connectivity-a-human-right/?

CHAPTER 3

Affective Contexts Worldwide

Introduction

People worldwide are very concerned about false information, especially across social media platforms, as perpetuated by domestic politicians (see Chap. 1). To better understand the mechanics of how such false information challenges the health of the *civic body*, this chapter explores the *economics of emotion* (the optimisation of datafied emotional content for financial gain) and the *politics of emotion* (the optimisation of datafied emotional content for political gain) under different affective contexts worldwide. We start by examining elections in the USA, given its increasingly politically affectively polarised population and long experience of emotive electoral disinformation on social media. We next turn to the Philippines, given its affective patronage democracy, clientelism and extremely high social media usage. We follow this with Sweden, a country that should be resilient to false information given its strong civic institutions, but that has witnessed a breakdown of consensus culture initiated by the emergence of right-wing populist nationalist political parties and supporting online media. These examples provide a grounded sense of the scale and dynamics of false information media systems. They highlight the importance of understanding specificities of affective contexts, and their intersections with international information flows such as information warfare, ideological struggles and resources for content moderation by global platforms.

V. Bakir, A. McStay, *Optimising Emotions, Incubating Falsehoods*,
https://doi.org/10.1007/978-3-031-13551-4_3

USA: Affectively Polarised Elections

American scholars pioneered the concept of political party identification, defining it as an intense psychological attachment to a political group (Campbell et al., 1960). A study of US partisans across five decades (1960–2010) confirms that partisan identities are primarily affective attachments (Iyengar et al., 2012). Studies also show that the USA is highly politically polarised and that this long predates the social media industry (Barrett et al., 2021). This is true whether one considers affective polarisation (namely, an emotional attachment to in-group partisans and hostility towards out-group partisans (Iyengar & Westwood, 2015)) or ideological polarisation (namely, the degree to which people disagree about political issues (Arguedas et al., 2022)). While there is little comparative work across countries on these forms of polarisation, several studies show that both forms of polarisation increased in the USA across the past four decades, rising more than in other countries (Boxell et al., 2020; Heltzel & Laurin, 2020; Draca & Schwarz, 2021). Explanations for the origins of such polarisation include rising income inequality, elite polarisation, demographic changes and strong political/cultural initiatives in the Democratic Party since President Barack Obama, such as tighter gun control and same-sex marriage (Böttcher & Gersbach, 2020). Explanations also include media influence. For instance, the five-decade study of US partisans (1960–2010) finds that affective polarisation is reinforced following exposure to prolonged media-based presidential negative campaigns (Iyengar et al., 2012). Also relevant is that surveys on news polarisation in the USA from 2016 and 2022 show that the USA is highly polarised compared to most other countries and has no large centrist media outlet (Fletcher, 2022).

It is in this highly polarised country, with very high Internet penetration (90% in 2022) and online news consumption (in 2022, 67% accessed news online, of which 42% were via social media (Jenkins & Graves, 2022)), that the social media and search engine platforms incubating global false information online are located. The USA relies heavily on self-regulation of its media industries. Neither the US Federal Communications Commission nor other federal regulators present formal rules on content that Internet platforms can carry. Section 230 of the Communications Decency Act of 1996 protects Internet platforms from the threat of private liability for content they host. This protection was an effort to promote rapid growth of Internet platforms and placed the burden of content

curation on the platforms themselves (Fukuyama & Grotto, 2020, p. 209). This lack of regulation allowed platforms to design their products in ways that further their business interests: namely, achieving massive scale and advertising income by maximising audience engagement (becoming *emotional by design*). The bigger the platform, the greater the network externalities that make it indispensable to users (so-called network effects) and the greater the capacity to extract data from users that enable the platforms to develop AI systems and target advertising more efficiently (McNamee, 2019; Starr, 2020). Furthermore, because their products are free to users, this protected platforms like Facebook and Google from anti-trust regulation in the USA (McNamee, 2019).

Given this lack of regulation and high polarisation, it is unsurprising that use of social media to spread emotive disinformation to manipulate elections has long been evident in the USA. In 2006, evidencing the *politics of emotion*, political blogs tried to influence American elections by gaming search engines to push web pages carrying negative content to the top of relevant search results. Using link bombing techniques ('Googlebombing'), website masters and bloggers use the anchor text to associate an obscure, negative term with a public entity. In 2010, a study of Twitter in the Massachusetts Senate race found a 'Twitter-bomb', namely, an organised effort to spread false information about the Democratic candidate through anonymous Twitter accounts targeting users interested in the topic (Metaxas & Mustafaraj, 2010; Mustafaraj & Metaxas, 2017). In the 2016 presidential campaign, pro-Donald Trump tweets were more likely to go viral than those in support of his rival, Hillary Clinton, as Trump's gut-feeling tweets were more authentic (Enli, 2017) and because they were amplified by far-right pro-Trump outlets such as Fox News (Benkler et al., 2017; Faris et al. 2017, August 16). 'The donald' subreddit was monitored by the Trump campaign, passing the most powerful content onto the campaign's social media team (Moore, 2018, p. 29). Such amplification was enabled by the affordances of social media platforms that enabled those passionate about politics to organise. For instance, in 2015, Twitter introduced a group direct message function which led to some group direct messages turning into pro-Trump, invite-only 'rooms' accommodating 50 people, like 'Patriots United'. Many rooms had accompanying hashtags to track members' tweets as they propagated (Musgrave, 2017, August 9). Trump's election team also took advantage of Facebook's bespoke guidance on how to run campaigns successfully on Facebook, whereas Clinton's team did not (Levy, 2020,

p. 350). Beyond social media amplification, such emotive messaging is also amplified through the USA's highly polarised mainstream media, with the right-wing media ecosystem less reliant than the left-wing on professionally sourced facts or fact-checking (Faris et al., 2017, August 16; Benkler et al., 2017).

The *economics of emotion* is also in play. Facebook's ad auctions reward advertisers who target people who most want to see the ad, for instance, costing the advertiser less to advertise to such audiences (Levy, 2020, p. 350); and in 2016, Trump won Facebook ad auctions due to the likely engagement his content generates among target audiences compared with Clinton, taking shape in racism, misogyny and antagonism (Jutel, 2021). Similarly, Mustafaraj and Metaxas (2017) demonstrate that infiltration was successfully used on Facebook to spread fake news during the 2016 US presidential election, for financial benefit through online advertising. A year later, Silverman et al. (2017, August 8) documented the growing universe of US-focused, hyperpartisan websites and Facebook pages, many run from outside the USA, motivated by profit-seekers (also see Forest, 2022).

Such media ecologies are exploitable by foreign actors seeking to wage information warfare, for instance, by encouraging dissent via targeted attacks that play on existing societal and cultural fissures. During the 2016 presidential election, the foreign military intelligence agency of the General Staff of the Armed Forces of the Russian Federation (GRU) fuelled disinformation, seeking to influence the very bonds of society (Bolton, 2021). Tactics included hacking-and-dumping campaigns, fake online personas on social media and disseminating propaganda (Howard et al., 2018; Jamieson, 2018; McFaul & Kass, 2019). In terms of hacking-and-dumping campaigns, Russian cyber agents stole data from both the Republican and Democratic parties, then releasing only data stolen from the Democrats through fictitious online personae (DCLeaks and Guccifer 2.0) and through websites including WikiLeaks. This represented the first time that a foreign government had tried to steal data from American politicians and then publish it to influence an election. In terms of fake online personae, such content was created by Russian company, the Internet Research Agency, linked to the Kremlin (McFaul & Kass, 2019; Rid, 2021). In a 2015 exposé, *The New York Times* estimates that the Internet Research Agency's then approximately 400 employees created production-line content for every popular social network: Facebook, Twitter, Instagram, LiveJournal (popular in Russia), VKontakte (Russia's version

of Facebook) and comment sections of Russian news outlets (Chen, 2015, June 2). In September 2017, Facebook revealed that it had closed 470 fake Internet Research Agency-controlled accounts and pages that had bought $100,000 in advertising (over 3000 ads) pushing divisive issues (such as race, gay rights, police shootings and immigration) between 2015 and 2017, reaching at least 29 million Americans (Hatmaker, 2017, November 1; Shane 2017). Twitter disclosed that 3814 accounts were operated by the Internet Research Agency, reaching about 1.4 million people (McFaul & Kass, 2019). Big data analysis of the Internet Research Agency's Twitter activity in the US presidential election identifies five handle categories: 'Right Troll' (propagating nativist and right-leaning populist messages); 'Left Troll' (propagating socially liberal messages, focusing on cultural identity); 'News Feed' (presenting themselves as US local news aggregators); 'Hashtag Gamer' (where users add a hashtag to a tweet and then answer the implied question, such as '#WasteAMillionIn3Words Donate to #Hillary') and 'Fearmonger' (spreading disinformation about fabricated crisis events) (Linvill & Warren, 2020). Russia's Internet Research Agency activities were designed to interfere in elections by campaigning for African American voters to boycott elections or follow the wrong voting procedures; encouraging extreme right-wing voters to be more confrontational; and spreading sensationalist, conspiratorial, false political news to voters (Howard et al., 2018; Padda, 2020). Of course, in the context of overall spending on digital advertising in the 2016 election cycle ($1.4 billion), and overall bot and spam activity online, the amounts identified as Russian interference are tiny, and hence their impact may also be irrelevant (Boyd-Barrett, 2020). It does, however, highlight how the *politics of emotion* are marshalled in information warfare efforts.

In the USA, then, a country characterised by high political and news media polarisation, fake news stories and 'dark ads' (online ads only seen by the recipient) are readily fuelled by partisans, partisan outlets, mainstream press, social media, and domestic and foreign political actors. As Heltzel and Laurin (2020) observe, although fewer than 10% of Americans identify as *extremely* liberal or conservative, this very polarised minority pervades political discourse. News stories cover their views more often, and because both liberal and conservative extremists use negative, angry language to condemn opponents that make them feel threatened (Frimer et al., 2018), their messages on polarising issues containing moral-emotional words are more likely to spread through social networks (Brady et al. 2017; also see Mac & Silverman, 2021, February 21). Perhaps

influenced by this extreme and amplified political polarisation, the USA ranks lowest in media trust (at 26%) among news consumers surveyed across 46 countries in 2022, with only 14% of those on the right-wing trusting most news most of the time (the figure is 39% on the left-wing) (Newman et al., 2022). We turn now to the Philippines, where social media usage is even more pervasive than in the USA and where the affective nature of politics takes form in affective clientelism.

The Philippines: Affective Clientelism

The Philippines is a weak democracy, experiencing centuries' old socio-economic inequalities and a large gap between rich elites and poor masses. Emerging from a system of patron-client relations established during the Spanish colonial period (1521–1898), it has been described as a patronage democracy where parties and candidates mainly rely on contingent distribution of material benefits, or patronage, to mobilise voters. While liberal-democratic in name since the 1986 People Power Revolution removed dictator Ferdinand Marcos (whose dictatorship spanned 1972–1986), politics in the Philippines is clientelistic in practice, with the patron and strongman leadership linking political elites to the electorate. Political parties are candidate-centred coalitions of provincial bosses, political machines and local clans, anchored on clientelistic, parochial, personal inducements rather than on issues, ideologies or party platforms. Through patronage, presidents build alliances among political elites, including legislators and other state agencies, local politicians, warlords and clans (Teehankee & Calimbahin, 2020). Collective clientelism is the norm: it involves strong affective components, providing certain types of 'public goods' to specific groups in exchange for votes from group members. The loyalty bought by collective clientelism operates alongside coercion, including private and public use of violence at all levels of the politico-economic elite. There is a high level of legitimacy for leadership that credibly fills promises of both 'good' patronage and strongmanship (Kreuzer, 2020).

While there is a tradition of freedom of speech, traditional media outlets are owned by oligarchic families and new wealth, and the Philippines ranks high in terms of violence against media practitioners (Chua, 2022; Teehankee & Calimbahin, 2020). In 2016, a press freedom index compiled by Reporters without Borders (an international non-profit, non-governmental organisation whose stated aim is safeguarding the right to freedom of information) ranked the Philippines poorly (138 out of 180

countries), and it has since maintained a similar figure (Reporters without Borders, 2020). With high internet penetration (82% in 2022) (Newman et al., 2022), the reach of television to access news in the Philippines is declining (from 66% in 2020 to 60% in 2022), with a shift to social media (73% use it to access news in 2022) and online sites of traditional media (Chua, 2022). Almost half of the Philippines' 103 million citizens are highly active social media users; in 2020, Filipinos spent over 9 hours daily online, the highest usage in the world, well above the global average of 6 hours 43 minutes (Llamas, 2020). Access to Facebook is provided free with all smartphones (via Facebook's Free Basics), but Filipinos incur data charges when visiting other websites, including newspapers. Consequently, millions of citizens rely on social media for news, consuming partisan opinion masquerading as fact. In 2020, only 27% of Filipinos say that they trust news media overall, this figure rising to 37% in 2022 (Chua, 2022; Newman et al., 2020). This political and media milieu, with its strong affective components, provides fertile ground for false information, especially during elections.

The 2016 presidential elections in the Philippines (won by Rodrigo Duterte on a populist platform with record voter turnout) marked an increase in use of social media platforms, with curated content managed by professionals who amplified their message in an unregulated, cost-effective manner (Teehankee & Calimbahin, 2020). Unsurprisingly, the *politics of emotion* is evident in an ethnographic study across 2016–2017 that finds 'the architects of networked disinformation' to be a common part of Filipino political campaigns at national and local levels. Campaign strategists from boutique advertising and public relations agencies mobilise populist sentiment across the political spectrum, relying on the promotional labour of digital influencers on social media, and fake account operators who manually operate fake profiles to infiltrate community groups and news pages to generate 'volatile virality' (Ong & Cabañes, 2018, p. 8). This involves opening up spaces for discontent to hijack sentiments and sow public division; silencing political dissent; cyberbullying and 'slutshaming' influencers (especially women); using 'signal scrambling' (to dampen virality of opposing campaigns' hashtag by using similar but syntactically different decoy hashtags and seeding these to split the original hashtag's community); and engaging in historical revisionism (retelling sordid political histories as fairy tales of a golden age) (Ong & Cabañes, 2018). The *politics of emotion* is also evident in a study conducted by Demos (a British cross-party, independent think tank) and

US-based National Democratic Institute (a non-partisan, non-governmental organisation that aims to increase the effectiveness of democratic institutions in developing countries). Its focus on gendered disinformation on Twitter in the Philippines finds that stories told to discredit and discourage female participation in public life seek to engender anger, disgust and disdain in the third-party reader, and fear and shame in the second person target (Judson et al., 2020, October).

Well before Duterte's election, numerous fake news sites and partisan blogs supported him, with fake endorsements from celebrities and leaders like Pope Francis (such as 'chosen by God') (Syjuco, 2017, October 23). Notably, in January 2016, Facebook sent three employees to train the various presidential candidates and their staffs on how best to use Facebook. One month before the election, Duterte occupied 64% of all election-related conversations on Facebook pages in the Philippines, despite being vastly outspent by rivals (Vaidhyanathan, 2018, p. 193). Duterte's popularity levels increased as president, even as his administration eroded the separation of powers and rule by law to silence critical media and government opponents. His core, populist message was one of discipline, order and submission to the top strongman's commands (Kreuzer, 2020), including pronouncements in a violent war on drugs, unity of long-established power blocs through patronage and charges of fake news towards his critics (Ragragio, 2020).

In 2021, an investigative story in Rappler highlighted the role played in spreading false information by the *economics of emotion*. Its investigation of digital marketing group, Twinmark Media Enterprises, shows that several Filipino celebrities and influencers were paid hundreds of thousands to millions of pesos across 2017 and 2018 to unknowingly or indirectly amplify false information and government propaganda, before Facebook banned the agency in January 2019 for coordinated inauthentic behaviour. The strategy involved Twinmark paying influencers and popular meme and celebrity fan pages to share content from Twinmark-owned websites to increase engagement. The agency also has its own pages. Facebook users that follow the influencers or popular pages see the posts and are led to Twinmark websites, where they are served money-generating ads, false information or propaganda (Elemia & Gonzales, 2021, February 27). It is only when such behaviour affects Facebook's perceived priorities that action is taken. When ex-Facebook data scientist, Sophie Zhang, uncovered a network of fake accounts creating low-quality, scripted fake engagement for politicians in the Philippines in October

2019, Facebook left it to languish. But when a tiny subset of that network began creating an insignificant amount of fake engagement on Trump's page in February 2020, Facebook moved quickly to remove it (Wong, 2021, April 12).

While affective clientelism and extremely high social media usage are features of the Philippines, the predominance of social media, the neglect by globally dominant social media platforms (in terms of content moderation) and the weakness of mainstream news are common themes in many parts of the world suffering from false information online. However, even countries with strong, independent media institutions are not immune, as the following example from Sweden shows.

Sweden: Alt-Right Erosion of Consensus Culture

Sweden is a strong democratic state. Its secular, liberal society is based on knowledge, education, a strong welfare state, national unity and a deep, consensus-driven political culture (Andersson, 2009). It regularly tops all global rankings for good places to live with a reputation for gender equality, environmental concern, technological prowess and democratic design (Rapacioli, 2018). Reporters without Borders' (2021) press freedom index ranks Sweden as the third most independent and free press in the world out of 180 countries (it has been in the top ten since ranking began in 2013). It has strong press freedoms, with law enforcement actively combatting attacks against journalists. Its public service media is funded through taxation, and the government subsidises local news. Swedes' trust in news media is comparable to the global average: in 2020, four in ten Swedes express a general trust in the news, and 50% did so in 2021 and 2022, with trust much higher for news sources regularly used. However, with very high Internet penetration (96% in 2022), there are very low levels of trust for news found in social media (Newman et al., 2020, 2022; Westlund, 2021, 2022). Furthermore, according to Microsoft's (2021) Digital Civility Index, Sweden is quite uncivil online. It ranked only as 15th most civil out of 22 countries surveyed in 2021. As such, Sweden has both strengths and fissures in its resilience to emotive, false information online.

Furthermore, Sweden's consensus culture has been damaged by the emergence of far-right populist nationalist Sweden Democrats (Sverigedemokraterna [SD]). Founded in 1988, it crossed a threshold to become elected to parliament in 2010 and now forms the country's third

largest parliamentary party. Alongside Sweden's smaller, far-right parties, Sweden Democrats often nostalgically position the 1940s and 1950s as a 'golden age' of 'Swedish democracy, socio-economic wellbeing and ethnic homogeneity and cohesion' while accusing political opponents of eroding these phenomena through liberal immigration policies among other things (Merrill, 2020). Sweden Democrats drastically altered Sweden's dynamics of affective polarisation: by 2014–2015, extremely negative sentiment towards Sweden Democrats was found from Members of Parliament and voters from all other parties (Reiljan & Ryan, 2018). This erosion of consensus is reflected in the growth of Sweden's alternative right-wing media that has taken root despite Sweden's tradition of strong, independent media institutions (Reporters without Borders, 2020).

The rise of extreme right-wing politics, political populism and White supremacy movements has been accompanied by the growth of online alternative media with a far-right political agenda, fuelled by the *politics of emotion*. A cross-national analysis of right-wing alternative media use in Germany, Austria and Finland (countries with similar mainstream media systems, where right-wing populist parties have had electoral success) finds a comparatively high prevalence of right-wing alternative online media in Sweden. With regard to audience characteristics, the strongest predictors of 'alt-right' media use are political interest, a critical stance towards immigration, a sceptical assessment of news quality and distrust in public service broadcasting. Use of social media as a primary news source also increases likelihood of alt-right news consumption (Schulze, 2020). Highlighting the *economics of emotion*, there are also fake news websites with names almost identical to trusted local news websites (such as http://www.thelocal.com) circulating 100% fabricated stories (Rapacioli, 2018).

Sweden's consensus culture has been a constant point of reference and model for the European Left. However, the European and American liberal Right view Sweden as a dystopian, cradle-to-grave society that strangles individual freedom (Andersson, 2009). As such, Sweden's own network of right-wing alternative news sites regularly feed right-wing partisan outlets abroad (Rapacioli, 2018, p. 56). A study of far-right English-language media circulated mainly within transatlantic networks finds what Titley (2019) calls 'Taboo News' about Sweden. Structured by its antagonistic positioning in relation to the 'mainstream', it validates itself as covering news which will not be reported, or which is being actively suppressed, by a 'politically correct' public culture of 'fake news'. Such international

right-wing partisan outlets use racialising and Islamophobic discourses about Muslim immigrants to portray Sweden as a dystopian future to be averted: a failed social experiment in immigration and multiculturalism.

Conclusion

This chapter has exposed specific ways in which the *economics of emotion* and *politics of emotion* incubate false information in different affective contexts to harm the global *civic body*. Our three country-specific examples highlight the importance of understanding specificities of cultures of emotion, as well as their intersections with international phenomena of information warfare (in the case of the USA), global platform neglect (in the case of the Philippines) and ideological struggles (in the case of Sweden).

In the highly politically polarised USA, use of social media to spread emotive disinformation to manipulate elections has been apparent since 2006 and shows no signs of abating. Studies have uncovered disinformation techniques on dominant digital platforms; the importance of political supporters in social media amplification; and foreign and domestic actors promoting fake news and propaganda on social media to further polarise the USA, to spread conspiracies and for economic gain. As platforms constantly tweak their algorithms and alter their affordances, these are exploitable by those seeking to spread viral messages hidden from mainstream view. Notably, platforms have so far avoided content regulation in the USA, allowing them to design their products to maximise audience engagement while failing to protect quality of information flows. That these platforms are *emotional by design* as well as central to everyday life makes it hard for governments worldwide to enact legislation to curb platform power.

Consequently, in the Philippines, a 'patronage democracy' where strongman politicians provide public goods in return for votes, false information flourishes. Disinformation techniques rely on professional campaign strategists to orchestrate and pay digital influencers, and fake account operators to manually operate fake profiles to infiltrate community groups and news pages to generate 'volatile virality'. As in the USA, they open up spaces for discontent to hijack people's sentiments, exacerbate existing divisions and silence political dissent. Such false information takes root because millions of Filipinos rely on social media for news, but are neglected by dominant, US-based social media platforms in terms of providing resources for content moderation. Even Sweden, a country with a

tradition of consensus culture, strong, independent media institutions, and broad trust in mainstream news, is not immune to false information online. The rise of extreme right-wing politics since the late 1980s and political populism has led to an active alt-right media particularly concerned about immigration, providing fodder for transatlantic far-right media.

It is clear, then, that false information online manifests in varied affective contexts worldwide, driven by the *economics of emotion* and *politics of emotion* conducted across digital platforms. Recognising that this media ecology is highly complex with multiple stakeholders, in Part II we will focus on how the *civic body* can be strengthened to protect against affect-driven false information delivered via profiled targeting. It is to these core characteristics of false information, affect/emotions/mood and profiling/targeting that we now turn.

References

Andersson, J. (2009). Nordic nostalgia and Nordic light: The Swedish model as Utopia 1930–2007. *Scandinavian Journal of History, 34*(3), 229–245. https://doi.org/10.1080/03468750903134699

Arguedas, A. R., Robertson, C. T., Fletcher, R., & Nielsen, R. K. (2022). *Echo chambers, filter bubbles, and polarisation: A literature review*. Reuters Institute and the Royal Society. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review

Barrett, P. M., Hendrix, J., & Sims, J. G. (2021). *Fueling the fire: How social media intensifies US. political polarization –and what can be done about it*. NYU/Stern. Retrieved April 13, 2022, from https://www.stern.nyu.edu/experience-stern/faculty-research/fueling-fire-how-social-media-intensifies-u-s-political-polarization-and-what-can-be-done-about-it

Benkler, Y., Faris, R., Roberts, H., & Zuckerman, E. (2017). Study: Breitbart-led right-wing media ecosystem altered broader media agenda. *Columbia Journalism Review*. Retrieved April 13, 2022, from https://www.cjr.org/analysis/breitbart-media-trump-harvard-study.php

Bolton, D. (2021). Targeting ontological security: Information warfare in the modern age. *Political Psychology, 42*(1), 127–142. https://doi.org/10.1111/pops.12691

Böttcher, L., & Gersbach, H. (2020). The great divide: Drivers of polarization in the US public. *EPJ Data Science, 9*, 32. https://doi.org/10.1140/epjds/s13688-020-00249-4

Boxell, L., Gentzkow, M., & Shapiro, J. (2020). *Cross-country trends in affective polarization* (Working paper). National Bureau of Economic Research. Retrieved April 13, 2022, from https://doi.org/10.3386/w26669

Boyd-Barrett, O. (2020). *Russiagate. Disinformation in the age of social media*. Routledge

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences USA, 2017*(114), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Campbell, A., Converse, P., Miller, W., & Stokes, D. (1960). *The American voter*. Wiley.

Chen, A. (2015, June 2). The agency. *The New York Times*. https://www.nytimes.com/2015/06/07/magazine/the-agency.html

Chua, Y. T. (2022). Philippines. In N. Newman, R. Fletcher, C. T. Robertson, K. Eddy, & R. K. Nielsen (Eds.), *Reuters Institute digital news report 2022* (pp. 142–143). Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Draca, M., & Schwarz, C. (2021). *How polarized are citizens? Measuring ideology from the ground-up* (QAPEC Discussion Papers 07). Quantitative and Analytical Political Economy Research Centre. Retrieved April 13, 2022, from https://warwick.ac.uk/fac/soc/economics/research/centres/qapec/discussionpapers/manage/07_-_qapec_draca.pdf

Elemia, C., & Gonzales, G. (2021, February 27). Stars, influencers get paid to boost Duterte propaganda, fake news. *Rappler*. https://www.rappler.com/newsbreak/investigative/celebrities-influencers-get-paid-to-boost-duterte-propaganda-fake-news/

Enli, G. (2017). Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election. *European Journal of Communication, 32*, 50–61. https://doi.org/10.1177/0267323116682802

Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., & Benkler, Y. (2017, August 16). *Partisanship, propaganda, and disinformation: Online media and the 2016 U.S. presidential election*. Berkman Klein Center for Internet and Society at Harvard University. Retrieved April 13, 2022, from https://cyber.harvard.edu/publications/2017/08/mediacloud

Fletcher, R. (2022). Have news audiences become more polarised over time? In N. Newman, R. Fletcher, C. T. Robertson, K. Eddy, & R. K. Nielsen (Eds.), *Reuters Institute digital news report 2022* (pp. 38–41). Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Forest, J. F. (2022). *Digital influence mercenaries*. Naval Institute Press.

Frimer, J. A., Brandt, M. J., Melton, Z., & Motyl, M. (2018). Extremists on the left and right use angry, negative language. *Personality and Social Psychology Bulletin, 45*, 1216–1231. https://doi.org/10.1177/0146167218809705

Fukuyama, F., & Grotto, A. (2020). Comparative media regulation in the United States and Europe. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 199–219). Cambridge University Press.

Hatmaker, T. (2017, November 1). *Here's how Russia targeted its fake Facebook ads and how those ads performed*. Techcrunch. https://techcrunch.com/2017/11/01/list-russian-ads-facebook-instagram/

Heltzel, G., & Laurin, K. (2020). Polarization in America: Two possible futures. *Current Opinion in Behavioral Sciences, 34*, 179–184. https://doi.org/10.1016/j.cobeha.2020.03.008

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2018). *The IRA, social media and political polarization in the United States, 2012–2018*. Retrieved April 13, 2022, from https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/12/The-IRA-Social-Media-and-Political-Polarization.pdf

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science, 59*(3), 690–707. https://doi.org/10.1111/ajps.12152

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly, 76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Jamieson, K. H. (2018). *Cyberwar. How Russian hackers and trolls helped elect a president what we don't, can't, and do know*. Oxford University Press.

Jenkins, J., & Graves, L. (2022). United States. In N. Newman, R. Fletcher, C. T. Robertson, K. Eddy, & R. K. Nielsen (Eds.), *Reuters Institute digital news report 2022* (pp. 112–113). Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Judson, E., Atay, A., Krasodomski-Jones, A., Lasko-Skinner, R., & Smith, J. (2020, October). *The contours of state-aligned gendered disinformation online*. Demos. Retrieved June 23 2022, from https://demos.co.uk/project/engendering-hate-the-contours-of-state-aligned-gendered-disinformation-online/

Jutel, O. (2021). Digital teleologies: Blockchain, affect and digital teleologies. In M. Boler & E. Davis (Eds.), *Affective politics of digital media* (pp. 101–115). Routledge.

Kreuzer, P. (2020). *A patron-strongman who delivers. Explaining enduring public support for President Duterte in the Philippines*. Peace Research Institute Frankfurt. Retrieved April 13, 2022, from https://www.hsfk.de/fileadmin/HSFK/hsfk_publikationen/Prif0120.pdf

Levy, S. (2020). *Facebook: The inside story*. Penguin, Random House.

Linvill, D. L., & Warren, P. L. (2020). Troll factories: Manufacturing specialized disinformation on Twitter. *Political Communication, 37*(4), 447–467. https://doi.org/10.1080/10584609.2020.1718257

Llamas, C. (2020). *We Are Social report: Philippines tops internet and social media use in 2020.* Retrieved April 13, 2022, from https://www.marketing-interactive.com/we-are-social-report-philippines-tops-internet-and-social-media-use-in-2020

Mac, R., & Silverman, C. (2021, February 21). "Mark changed the rules": How Facebook went easy on Alex Jones and other right-wing figures. *Buzzfeed News.* https://www.buzzfeednews.com/article/ryanmac/mark-zuckerberg-joel-kaplan-facebook-alex-jones

McFaul, M., & Kass, B. (2019). Understanding Putin's intentions and actions in the 2016 U.S. Presidential Election. In M. McFaul (Ed.), *Securing American elections.* Stanford Cyber-Policy Centre. Retrieved April 13, 2022, from http://cs.brown.edu/courses/csci1800/sources/2019_06_06_Stanford_SecuringAmericanElections.pdf

McNamee, R. (2019). *Zucked: Waking up to the Facebook catastrophe.* Harper Collins.

Merrill, S. (2020). Sweden then vs. Sweden now: The memetic normalisation of far-right nostalgia. *First Monday, 25*(6). https://doi.org/10.5210/fm.v25i6.10552

Metaxas, P. T., & Mustafaraj, E. (2010). From obscurity to prominence in minutes: Political speech and real-time search. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line.* Retrieved April 13, 2022, from https://repository.wellesley.edu/cgi/viewcontent.cgi?article=1008&context=computersciencefaculty

Microsoft. (2021). *Digital Civility Index.* Retrieved April 13, 2022, from https://www.microsoft.com/en-us/online-safety/digital-civility

Moore, M. (2018). *Democracy hacked: Political turmoil and information warfare in the digital age.* OneWorld Publishing.

Musgrave, S. (2017, August 9). I get called a Russian Bot 50 times a day. *Politico.* http://www.politico.com/magazine/story/2017/08/09/twitter-trump-train-maga-echo-chamber-215470

Mustafaraj, E., & Metaxas, P. T. (2017). *The fake news spreading plague: Was it preventable?* Preprint retrieved from http://arxiv.org/abs/1703.06988

Newman, N., Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Reuters Institute digital news report 2020.* Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K. (2022). *Reuters Institute digital news report 2022.* Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Ong, J. C., & Cabañes, J. V. A. (2018). *Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines.* Retrieved April 13, 2022, from https://doi.org/10.7275/2cq4-5396

Padda, K. (2020). Fake news on Twitter in 2016 U.S. presidential election: A quantitative approach. *The Journal of Intelligence, Conflict, and Warfare, 3*(2), 1–24. https://doi.org/10.21810/jicw.v3i2.2374

Ragragio, J. L. D. (2020). Strongman, patronage and fake news. Anti-human rights discourses and populism in the Philippines. *Journal of Language and Politics.* https://doi.org/10.1075/jlp.20039.rag

Rapacioli, P. (2018). *Good Sweden, bad Sweden: The use and abuse of Swedish values in a post-truth world*. Volante.

Reiljan, A., & Ryan, A. (2018). Affective and ideological polarisation in Swedish party system (1985-2015): Voter and elite level analysis. In *ECPR General Conference*, Universität Hamburg, Hamburg, 22–25 August 2018. Retrieved April 13, 2022, from https://ecpr.eu/Events/Event/PaperDetails/42588

Reporters without Borders. (2020). *2020 World Press Freedom Index.* Retrieved April 13, 2022, from https://rsf.org/en/ranking/2020

Reporters without Borders. (2021). *2021 World Press Freedom Index.* Retrieved April 13, 2022, from https://rsf.org/en/ranking#

Rid, T. (2021). *Active measures: The secret history of disinformation and political warfare.* Profile Books.

Schulze, H. (2020). Who uses right-wing alternative online media? An exploration of audience characteristics. *Politics and Governance, 8*(3), 6–18. https://doi.org/10.17645/pag.v8i3.2925

Shane, S. (2017, November 1). These are the ads Russia bought on Facebook in 2016. *The New York Times.* https://www.nytimes.com/2017/11/01/us/politics/russia-2016-election-facebook.html

Silverman, C., Lytvynenko, J., Vo, L. T., & Singer-Vine, J. (2017, August 8). Inside the partisan fight for your newsfeed. *Buzzfeed News.* https://www.buzzfeed.com/craigsilverman/inside-the-partisan-fight-for-your-news-feed/

Starr, P. (2020). The flooded zone: How we became more vulnerable to disinformation in the digital era. In W. L. Bennett & S. Livingston (Eds.), *The disinformation age: Politics, technology and disruptive communication in the information age* (pp. 67–91). Cambridge University Press. https://doi.org/10.1017/9781108914628

Syjuco, M. (2017, October 23). Fake news floods the Philippines. *The New York Times.* https://www.nytimes.com/2017/10/24/opinion/fake-news-philippines.html

Teehankee, J. C., & Calimbahin, C. A. A. (2020). Mapping the Philippines' defective democracy. *Asian Affairs: An American Review, 47*(2), 97–125. https://doi.org/10.1080/00927678.2019.1702801

Titley, G. (2019). Taboo news about Sweden: The transnational assemblage of a racialized spatial imaginary. *International Journal of Sociology and Social Policy, 39*(11/12), 1010–1023. https://doi.org/10.1108/IJSSP-02-2019-0029

Vaidhyanathan, S. (2018). *Antisocial media: How Facebook disconnects us and undermines democracy*. Oxford University Press.

Westlund, O. (2021). Sweden. In N. Newman, R. Fletcher, A. Schulz, S. Andı, C. T. Robertson, & R. K. Nielsen (Eds.), *Reuters Institute digital news report 2021* (pp. 104–105). Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf

Westlund, O. (2022). Sweden. In N. Newman, R. Fletcher, C. T. Robertson, K. Eddy, & R. K. Nielsen (Eds.), *Reuters Institute digital news report 2022* (pp. 38–41). Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Wong, J. C. (2021, April 12). Revealed: The Facebook loophole that lets world leaders deceive and harass their citizens. *The Guardian*. https://www.theguardian.com/technology/2021/apr/12/facebook-loophole-state-backed-manipulation

CHAPTER 4

The Nature and Circulation of False Information

Introduction

This chapter focuses on the nature of disinformation (false information spread with intent to deceive) and misinformation (false information spread without specific deceptive intent), inquiring into processes that increase their circulation online. As befits any study of media systems, it addresses interconnections between technologies, media forms, wider media and political environments, people and impacts. It opens with a discussion on the role of deception in citizen-political communications. This highlights the long-standing debate on whether political leaders should lie to their citizens, addressing evidence of such activity in the areas of national security and election campaigns. It then discusses the nature and scale of two key forms of contemporary disinformation: fake news and deepfakes. Widening the focus beyond intentionally deceptive forms to false information in general, the chapter then examines the dynamics of spreading false information online, discussing why people engage with such processes.

Deception in Citizen-Political Communications

The role of deception in citizen-political communications has a long lineage. These span deliberations by Plato in 380 BC and Machiavelli in 1532 over whether leaders should lie to their publics. Plato advocated that

V. Bakir, A. McStay, *Optimising Emotions, Incubating Falsehoods*,
https://doi.org/10.1007/978-3-031-13551-4_4

states should be ruled by wise, intelligent, philosopher kings, because most people were too ignorant and irrational for democracy to be a good form of government (Plato, 2007 [381 B.C.]). Plato also advocated that philosopher king rulers would often need to deceive the people for their own good in the form of 'noble lies' (Plato, 2007 [381 B.C.]). Centuries later, informed by the context of Italy's city-states, Machiavelli (2003 [1532]) described how deception and treachery were essential components of successful governance.

This debate was reinvigorated by early-twentieth-century American mass communication theorists, some of whom viewed propaganda as essential to mass democracy. For them, propaganda enabled organised management and manipulation of the newly enfranchised masses by trained elites according to the elite's vision of the public interest: for instance, where truth would threaten political stability, the safety of an army, a diplomatic negotiation, or (more self-interestedly) advantage in an election. Propaganda was deemed necessary given the sheer scale of modern governance and the fact that most people would hear about governmental decisions through mass media rather than being able to directly experience or influence these decisions (Lasswell, 1936, 1971 [1927]; Lippman, 1922, 1993 [1927]). Lippman (1922), for instance, saw the public as incapable of fully understanding the complexity and 'truth' of political reality, not least because of lack of time but also because people view any facts through their own subjective mental constructs and cultures, as well as through journalists' partial lenses, which in turn is influenced by news owners' ideological and financial motivations.

Conversely, those sceptical of elite power argue that deception should be limited, with strict criteria for when it might be justified (Bakir et al., 2018a). Bok (1999, p. 20) argues against deception as it erodes social trust by depriving people of the ability to make choices for themselves according to the best information available. Bok suggests that: 'If duplicity is to be allowed in exceptional cases, the criteria for these exceptions should themselves be openly debated and publicly chosen. Otherwise government leaders will have free reign to manipulate and distort the facts and thus escape accountability' (Bok, 1999, p. 170). Indeed, there is evidence of long-standing, systematic, political and commercial efforts in liberal democracies to persuade and influence populations through deception (Bakir et al., 2018b) and through disinformation networks of think tanks, lobbyists and financiers (MacLean, 2020; Oreskes & Conway, 2011; Oreskes et al., 2020). Unsurprisingly, numerous surveys in the USA, UK,

European Union and Australia since the 1990s indicate that government officials, industry officials and journalists rank lowest out of society's key institutions that people trust to tell the truth. Indeed, during the first decade of the twenty-first century when social media was still in its infancy, explanations for observably declining trust in news and media included the increased use of professionalised communications from public relations; political and commercial intentions to manufacture public opinion rather than represent citizens' views to power; and opaque practices of digital surveillance and commodification of users (Bakir & Barlow, 2007). All these practices continue today.

More recently, since fears that 'fake news' influenced the 2016 US presidential election and 'Brexit' referendum in the UK, scholars from diverse disciplines have become intensely interested in disinformation. However, prior to these events, such disinformation was rarely addressed. Illustrating this point, until relatively recently, intelligence agencies were amongst the more likely purveyors of state-generated disinformation. Yet, systematic reviews of germane disciplines (Journalism, Media, History and International Relations) regarding such secretive attempts to influence the media find few studies (Bakir, 2015, 2019 [2018]). Nonetheless, these systematic reviews have synthesised knowledge on disinformation techniques used by intelligence agencies and their interlocking political, military and corporate networks (so-called 'intelligence elites' (Bakir, 2019 [2018])), and their success in secretly influencing the media and wider civil society. The systematic reviews find recurring examples of disinformation targeted at foreign audiences, but also sometimes domestic publics. Most studies focus on World War II and the Cold War, often featuring Israel, the UK and the USA, but with disinformation primarily originating in the Union of Soviet Socialist Republics (USSR) (Bakir, 2019 [2018]). Disinformation operations become known in the USSR as 'active measures', a term describing techniques to strategically disrupt policies and relations of opposition governments while strengthening allies. It involves a disorienting mix of true facts and fakes, methodically leveraged by large bureaucracies (intelligence agencies) to resonate with a targeted community's emotions, to covertly widen rifts and tensions, to undermine trust in specific institutions and to destabilise other states' relations with their publics and each another. Such activities involving 'slow-moving, highly skilled, close range, labor-intensive psychological influence' were professionalised by the USA's Central Intelligence Agency following World War II, with the USSR achieving a well-resourced

disinformation bureaucracy by the 1970s, maintained until the USSR's dissolution in 1991 (Rid, 2021, p. 7).

Current scholarship on digital disinformation, much of it based in information science, would benefit from consulting this literature, as it enriches our understanding of disinformation techniques currently evident across social media where active measures have become 'high tempo, low-skilled, remote and disjointed' (Rid, 2021, p. 7). For instance, large influence operations on social media often use established media outlets as camouflage, a practice originating in information operations during the Cold War (Rid, 2021). Current examples of such usage (supplied by non-profit organisation EU DisinfoLab, which is funded by private and government sources, and aims to be fact-based, balanced and objective) include France's online news-site, France Soir (resting on its past credibility before it went bankrupt); Italy's registered online newspaper, Gasp.news (which has the same owner and publishing team as anti-science outlet, Oltre.tv, a COVID-19 false information super-spreader); and exploiting factional media such as American cable news channel, Fox News and Russian state-owned news agency, Sputnik (EU DisinfoLab, 2021, October 13).

Under President Vladimir Putin (2000–2008, 2012–), Russia has several key disinformation strategies, these identified by security organisations such as NATO, Partnership for Peace Consortium of Defense Academies and Security Studies Institutes, and EU East Stratcom Task Force (which is part of the Strategic Communications and Information Analysis Division of the European External Action Service and was set up to address Russia's disinformation campaigns). Identified strategies include troll factories to, for instance, spread as many conflicting messages and conspiracy theories as possible to persuade audiences that the truth cannot be found (Abrams, 2016; EU East Stratcom Taskforce, 2017, January 19; Waszczykowski, 2015). Disinformation operations also involve passing on genuine hacked and leaked data to whistleblowing websites such as WikiLeaks, the truthful content flanked by little lies about the data's provenance or the publisher's identity (Rid, 2021). Other disinformation tactics include setting up ostensibly independent investigative platforms abroad to propagate alternative narratives, with content then amplified in Russian-controlled news outlets (EU DisinfoLab, 2021, October 13).

While contemporary Russian cyber-enabled foreign interference targets the USA and Europe, China targets Indo-Pacific nations during elections, according to a study by the Australian Strategic Policy Institute (a

government-funded defence and strategic policy think tank) (Hanson et al., 2019). For instance, its identity-grievance disinformation campaigns have sought to undermine political trust and amplify partisan discord in Taiwan (Nisbet & Kamenchuk, 2019; Tatsumi et al., 2020). China is believed to have started directly targeting Western audiences during Hong Kong's pro-democracy protests in 2019, according to the Stimson Center (an American non-profit and non-partisan think tank with diverse funders including defence and security contractors) (Tatsumi et al., 2020). According to the Council on Foreign Relations (an American non-profit, non-partisan think tank specialising in US foreign policy and international relations and funded by individuals, corporations and foundations), alongside Russia, China also spreads disinformation on the origins of COVID-19 and effectiveness of other actors' responses to it (for instance, highlighting and misrepresenting democracies' failures) (Kurlantzick, 2020). This provides attractive content for conspiracy theorists and far-right extremists in the West. However, unlike Russia, China appears more interested in promoting a positive image of itself as a responsible global leader, rather than merely disparaging others (Brovdiy, 2020).

There has been growing attention to Russia and China in analysis of contemporary disinformation, but liberal democracies also engage in such activities during times of conflict. For instance, Kaufmann (2004) analyses how the US administration of George W. Bush distorted the public record on the intelligence-assessed threat posed by Saddam Hussein's weapons of mass destruction to justify invading Iraq in 2003: favourable analyses were selectively publicised, while contrary information was suppressed. Briant (2015) shows how, since '9/11', the close Anglo-American relationship is used to overcome restrictions on propaganda for counterterrorism purposes, exploiting the two countries' different capabilities and the UK's weaker legislative restrictions. Snowden's leaks in 2013 into (primarily) US and UK mass surveillance tools also show that British intelligence agencies possesses multiple tools for online covert action to manipulate social media (Bakir, 2015). Boyd-Barrett's (2020) examination of 'RussiaGate' discourse following the 2016 election of Trump finds multiple examples of pro-NATO propaganda manoeuvres that resulted in a focus on Russia's deceptive tactics while deflecting attention from those of the USA and UK. Indeed, multiple studies across liberal democracies highlight challenges currently faced by journalists when dealing with 'intelligence elites' (for a review, see Bakir, 2019 [2018]). Journalists' challenges include recognising disinformation being funnelled through

the press and recognising attempts to shape the narrative through selective, authorised intelligence-based leaks, declassification and message coordination between various intelligence agencies. When intelligence whistleblowers come forward to the press (itself a rare occurrence), journalists must deal with intelligence elite secrecy-based strategies of influence aimed at protecting intelligence sources and methods (such as highly limited public admissions and vague statements about press inaccuracy and lack of context while denying journalists' requests for information to explain the leaks) (Bakir, 2016 [2013], 2019 [2018], Lashmar, 2013). Journalists also have to deal with being surveilled, which compromises source anonymity and may deter whistleblowing (Lashmar, 2017). Added to this, contemporary strategies of influence targeted at whistleblowers are primarily one of silencing, using techniques of identifying, prosecuting, threatening and harassing those who whistleblow (with unauthorised leaks) to the press (Bakir, 2019 [2018]).

Despite deploying deception in security issues, democratic ideals value honesty in public life. This is because democracies are unlikely to accurately express the people's will unless the people have sufficient information on which to base their vote. As Perloff (2018) observes, voters are not customers purchasing a marketed political product. Rather, voters are citizens, the fundamental constituents of a democratic society, whose ideas and objectives elected leaders must represent and channel into legitimate public policy rather than deceive (also see Habermas, 1996 [1962]). Despite such ideals, deception is a long-observed tactic in democracies during elections. For instance, the deceptive intent of political campaigners and lobbyists, SCL Elections and associated companies, is documented in their efforts to influence elections worldwide, through distortion of facts, creation of online 'sock puppets' (fake identities), amplification of deceptive sources and microtargeting voters to suppress and direct their vote. According to the UK's Parliamentary Inquiry into Fake News and Disinformation, 17 countries were affected across most continents (Digital, Culture, Media and Sport Committee, 2018, July 24, p. 54). During post-election violence in Kenya (2017) and Zimbabwe (2018), people posted videos and pictures of past election violence as if they were current events (Ndlela, 2020), and political disinformation in Zimbabwe threatened to undermine the credibility of its electoral process (Mare & Matsilele, 2020). In Nigeria's 2019 elections, candidates paid social media influencers typically less than US$14 per month to spread false information about opponents (Mano, 2020). A study of 14,684 Facebook ads

published by five major national political parties during two Spanish general elections in 2019 finds that 2% of the ads contained disinformation: these largely came from new parties and on a highly polarised issue (the country's cohesion and autonomy of each of its regions) (Cano-Orón et al., 2021).

Another problematic dimension for citizen-political communications is unearthed by more recent studies on gender-based disinformation, namely, the dissemination of false or misleading information attacking women (especially political leaders, journalists and public figures), basing the attack on their identity as women (Sessa, 2020, December 4). According to a report from the Center for Democracy and Technology (an American non-profit organisation whose stated aims include enhancing freedom of expression globally and stronger legal controls on government surveillance), various studies find that gendered disinformation seeks to maintain the status quo of gender inequality, undermine women's ability to participate in representative politics and create more polarised electorates (Thakur & Hankerson, 2021). For instance, one such study by Demos (a British cross-party, independent think tank) and US-based National Democratic Institute (a non-partisan, non-governmental organisation that aims to increase the effectiveness of democratic institutions in developing countries) examines state-aligned (right-wing) gendered disinformation on Twitter in Poland and the Philippines emanating from politically influential people. It finds evidence of emotive disinformation campaigns attacking women and using gendered narratives to undermine women who oppose or criticise the state. In these cases, gendered disinformation plays on existing stereotypes and tropes to try to convince people that women in public life are devious, stupid, overly sexual, in need of protection, immoral and ultimately unfit for public life. While women fight back through counterspeech (to challenge the gendered disinformation), such disinformation has 'chilling effects' on politically active women's social media engagement in Kenya, Indonesia and Colombia, as well as self-censorship of female journalists (Judson et al., 2020, October). A study from the Wilson Center (an American non-partisan forum funded by government and donations) on the perspectives of female politicians, media workers and civil society leaders from 30 countries and 5 continents indicates that such disinformation campaigns may make women who are interested in politics reconsider their ambitions while also diverting and sapping the energy of women leaders (Di Meco, 2019).

To summarise, deception in citizen-political communications has a long lineage across liberal democracies and autocracies, especially in areas of national security and election campaigns but also in gendered participation in civic life. A wide range of deceptive techniques have been documented by scholars, security organisations and think tanks, as well as arising challenges to journalism. Traditionally the domain of intelligence agencies and well-resourced organisations or countries, deception is now propagated by diverse actors in the contemporary media ecology, at speed and at scale. We turn now to two key deceptive media forms that have attracted recent concern: fake news and deepfakes.

Contemporary Forms of Deception

We examine two important, contemporary forms of disinformation: one that rose to notoriety in recent years (fake news) and one that could further erode our trust in the indexical link between reality and audiovisual recordings (deepfakes).

Fake News

Fake news emerged as a key form of disinformation in 2016, the nature and scale of which have since been much debated. The news media are often a focus of efforts to persuade and influence because of their professional commitment to accuracy, facticity and in some cases impartiality and objectivity (Gelfert, 2018). These professional commitments can be traced in the USA to the turn of the twentieth century, responding to sensational 'yellow journalism' of the 1890s and 'muckraking' in the early 1900s. Such professional commitments are intended to confer credibility and truth to the news, with claims to impartiality drawing attention away from the fact that all news is constructed and hence presents a selective, partial view (Winston & Winston, 2021). Yet, affectively charged (sensationalised, horrible, wondrous), deceptive (slanted, biased, opinionated, misleading and sometimes wholly fabricated) news has long been with us, the product of professional persuaders, propagandists, partisans and audience-maximising, competitive, news outlets.

This is evident in the history of the growth of the mass-market press in the USA and Europe (Dentith, 2017; Habermas, 1996 [1962], Winston & Winston, 2021). For instance, Winston and Winston (2021, p.17) describe how 'strange newes' (sic) (sensationalised accounts often

containing a kernel of truth about unusual and, at the time, poorly understood, meteorological, psychological, biological and astronomical events) made regular appearances in sixteenth-century printed pamphlets and broadsheets (the forerunners of recognisable newspapers). They also note observations by seventeenth-century cultural critics about the mendacity of the printed press. Several centuries later, in 1925, *Harper's Magazine* published an article titled 'Fake news and the public', decrying how emerging technologies made it hard to separate fact from rumours (Grinberg et al., 2019). Affect and deception are especially apparent in more recent news forms like cable TV (Hermida, 2016) and are highly evident online.

Since 2016 brought the term 'fake news' to popular attention, scholars have parsed this term. Before this date, the term was reserved for parody news outlets that used satire as political commentary (Baym, 2005). Shortly after Donald Trump took presidential office in the USA in 2016, there was a large increase in usage of the term 'fake news' in political discourse. Analysis of Trump's Twitter discourse (across 2016–2017) finds that Trump uses fake news accusations to demonstrate allegiance to Fox News and as a cover for his own spreading of false information that he frames as truth (Ross & Rivers, 2018). Indeed, the term 'fake news' is increasingly used worldwide as a rhetorical device used by the powerful to crush dissent (Dentith, 2017) or to justify state censorship policies (Newman et al., 2018; RSF, 2017).

Given this state of affairs, Gelfert (2018) argues that that the term 'fake news' should be reserved for cases of deliberate presentation of (typically) false or misleading claims as news, where these are misleading by design. As Egelhofer and Lecheler (2019) observe, many other definitions of fake news similarly stipulate that fake news: (1) is, either wholly or partly, deliberately fabricated (i.e. low in facticity); (2) has intention to deceive; and (3) has the appearance of a genuine news story (also see Mustafaraj & Metaxas, 2017; Nelson & Taneja, 2018). Egelhofer and Lecheler (2019) argue that all three characteristics should be present before labelling something as 'fake news'. Horne and Adal (2017) show that fake news items are shorter and less informative than real news, use less complex and more personal language, and are likely to have longer titles containing the article's main claim. An analysis of 150 real and fake news articles using an AI application to test for differences in emotional appeal finds that fake news story titles are substantially more negative than real news titles. Furthermore, the text body of fake news is substantially higher in

displaying specific negative emotions, such as disgust and anger, and lower in displaying positive emotions such as joy (Paschen, 2019).

Scholars have tried to quantify the extent of fake news circulating and spreading throughout the *civic body*. A big data study that measures trends in diffusion of content across social media platforms finds a mixed picture (Allcott et al., 2019). It examines 569 fake news websites and 9540 fake news stories on Facebook and Twitter between 2015 and 2018. It finds that user interactions with false content rose steadily on both Facebook and Twitter through the end of 2016, but then fell sharply on Facebook while continuing to rise on Twitter. However, the study also shows that the absolute level of interaction with false content remains high.

Several studies investigate how many citizens visited fake news sites in the USA (Guess et al., 2018; Grinberg et al., 2019; Nelson & Taneja, 2018) and in France and Italy (Fletcher et al., 2018). They find that fake news (narrowly defined) is consumed by only a small proportion of people. For instance, Grinberg et al.'s (2019) study of exposure to, and sharing of, political fake news by registered voters on Twitter during the 2016 US presidential election finds that engagement with fake news sources was extremely concentrated. Only 1% of individuals accounted for 80% of fake news source exposures, and just 0.1% accounted for nearly 80% of fake news sources shared. They observe that most political exposures, across all political groups, still came from popular non-fake news sources. Similarly, a study in France and Italy finds the actual audience of fake news sites to be limited compared to the audience of established news sites (Fletcher et al., 2018).

Looking beyond specific national levels finds overall rising levels of fake news online. Vargo et al.'s (2018) computational approach to investigate fake news in the online landscape from 2014 to 2016 uses the Global Knowledge Graph of the Global Database of Events, Language, and Tone (GDELT) as its data source. On a daily basis, GDELT ingests all news-like content globally from online sources including Google News and identifies people, locations, themes, emotions, narratives and events. This allows researchers to computationally analyse real, fake and fact-checking-oriented news content. The study found 60 fake news websites and 171,365 stories from fake news websites. It confirms that content generated from fake news sites was on the rise during that period and that fake news has a relatively stable ability to influence the entire mediascape.

Deepfakes (and Shallowfakes)

While fake news attracted popular concern from 2016 onwards, this was quickly followed by concern over deepfakes. Audiovisual material has long been held as a bastion of trusted evidence (where seeing is believing), but this is challenged by the rise of synthetic media (viz. media that is enabled or modified by AI). While computationally manipulated audiovisual content has been possible in Hollywood and boutique production outfits since the 1990s (landmark films include *Jurassic Park* (1993) and *Avatar* (2009)), its expense was prohibitive. As such, creating manipulated videos for political propaganda was rare (Langguth et al., 2021). However, since 2017, 'deepfakes' have emerged. These use 'deep learning', a type of machine learning that uses layers of algorithms called 'neural networks' to sort through visual data to make predictions (Paris & Donovan, 2019).

In 2017, a Canadian AI start-up called Lyrebird released its voice imitation algorithm (a form of 'speech synthesis') that mimics a real person's speech and shifts its emotional cadence, based on snippets of real-world audio (Lomas, 2017, April 25). The firstly widely known deepfakes appeared in 2017, when a Reddit user called Deepfakes uploaded videos with faces of famous actresses over the faces of pornographic actresses. In 2018, a deepfake video of former US President Barack Obama harshly criticising then US President Trump went viral (Pérez Dasilva et al., 2021). Since then, deepfakes across the world have been recorded, many for pornographic purposes but also for artistic, creative commentary and satire spurred by development of accessible, easy-to-use apps. For instance, Reface superimposes a face onto existing gifs; FaceApp allows users to age and contort a facial image; and Zao draws on a film and TV clip library to enable voice modulation and faceswaps. Deepfakes have also been used for advocacy purposes. For instance, Pakistani climate-change initiative, Apologia Project, depicts world leaders apologising from the year 2032 for their previous inaction on environmental crises, with its rhetorical power deriving from the seeming sincerity of the leaders' remorse and knowledge that more could have been done on their watch. Deepfakes have brought back deceased victims of injustice to demand change, as with murdered Mexican journalist Javier Valdez Cárdenas, who was 'brought back' by the group Propuesta Civica to call for an end to state-backed violence against the press (Ajder & Glick, 2021).

As Langguth et al. (2021) explain, the first generation of deepfake software available from 2017 required a huge number of training images to

function properly. Consequently, these programs were impractical for creating manipulated videos of an average person (Pérez Dasilva et al., 2021). For that reason, most deepfake videos created for entertainment purposes featured famous actors of which many images are publicly available (Langguth et al., 2021). However, a second generation of deepfake software no longer needs many training images to function properly. Instead, they use Generative Adversarial Networks (GANs), namely, algorithms designed to replace human faces or voices in thousands of images and videos to make them more realistic (Li et al., 2018).

The adversarial dimension of a GAN is that it is not just one network but two. Its goal is to trick itself into not being able to tell the difference between what is real and fake. A GAN includes a 'generator' that will eventually learn to produce convincing output (the deepfake) and the 'discriminator' that exists to test whether what is coming from the generator is fake or real. If deemed to be fake, this is reported back to the generator that will keep trying to fool the discriminator that, eventually, will not be able to tell the difference between real and fake (or, e.g. a photo of a real person or a photo of a synthetic 'person' that has never existed). The same principle of GAN self-deception applies to deepfakes in that a system will have data about the real person (such as Donald Trump), but if asked to say something that the real person has not said, the same process of self-deception applies: try, fail, amend and repeat until the discriminator is deceived.

As GANs improve (especially through increased computation), future creations of deepfakes can only become a more significant problem in spreading false information (Vizoso et al., 2021). Indeed, deepfakes have already been used for disinformation purposes. For instance, Rana Ayyub, a female Indian journalist and critic of India's ruling Bharatiya Janata Party (BJP), fell victim to a deepfake campaign that used face replacement in a pornographic video to discredit her (European Science Data Hub, 2019, December 4). Deepfakes have been used to make fake accounts appear more authentic: in December 2019, Facebook identified and removed a network using AI-generated photos to conceal their fake accounts (Bickert, 2020, January 6). Another example is Peace Data, an apparent global news organisation set up by Russia's Internet Research Agency to heighten discord ahead of the 2020 US presidential elections. The staff, editor and editorial assistants on Peace Data's website were fake personas with AI-generated profile pictures. Russian trolls used these fake personas to contact freelance journalists, paying them to write articles for the website

(EU DisinfoLab, 2021, October 13). The first weaponised use of deepfakes during an armed conflict appeared in March 2022 during Russia's invasion of Ukraine. The poor-quality deepfake video emerged on Facebook, YouTube and Twitter and was also posted to Telegram and Russian social network, VKontakte. It shows Ukrainian president, Volodymyr Zelensky, urging his country's troops to surrender to invading Russian forces (Simonite, 2022, March 17).

Given the recency of the phenomenon, there are few studies on audience's responses to deepfakes, but they find limited capacity to recognise this new deceptive form, especially when the content presented is neutral rather than suspiciously out of character. An experimental study on people's ability to distinguish wholly synthetic faces from real faces finds that synthetic faces are indistinguishable from, and regarded as more trustworthy than, real faces and that this realism of synthetic faces extends across race and gender (Nightingale & Farid, 2022). Vaccari and Chadwick (2020) find that a deepfake from 2018 with inflammatory, but unlikely, content (Obama calling President Trump a 'dipshit') deceives 16% of respondents in a nationally representative UK survey, with more people (33%) feeling uncertain than misled, but only 51% recognising the statement as untrue. An experimental study on the British public's ability to distinguish more neutral deepfakes from ordinary videos finds that individuals are no more likely to notice anything out of the ordinary when exposed to a deepfake video of neutral content compared to a control group who viewed only authentic videos. Although content warnings (where participants are told that at least one of the videos they are to see is a deepfake) improve capacity for detection among participants, most are still unable to identify the deepfake (Lewis et al., 2022). With even fewer studies on the impact of deepfakes on political attitudes, Dobber et al.'s (2020) online experiment is instructive. It constructed a political audiovisual deepfake to study effects on Dutch participants' political attitudes (n = 278). Only a small fraction of the sample recognised the deepfake as a manipulated video, and attitudes towards the depicted politician are significantly lower after seeing the deepfake.

Looking to the future, synthetic media could lead to more convincing online fake profiles. It could elicit more visceral, emotional and empathic responses than text-based media, putting words that we want to hear into the mouths of political leaders and resurrecting the dead to deliver powerful demands (as already deployed by advocacy groups). With people's difficulty in detecting deepfakes, even in highly digitally literate societies such

as the UK and the Netherlands, its appeal is obvious for those with malign intent, seeking to spread emotionally disturbing disinformation.

However, it is not so much deepfakes but the idea of them that appears to have been the main locus of disinformation to date. Increasingly realistic synthetic media can provide malign actors the opportunity to avoid being held accountable, by suggesting that anything, even the audiovisual record, can be fake. Since 2017, US president Trump claimed repeatedly that the Access Hollywood recording from 2005 in which he bragged about grabbing women's genitals was inauthentic. In 2019, a poorly made video of a New Year's address by the allegedly incapacitated Gabonese president, Ali Bongo, was declared a deepfake and part of a cover-up by opposition leader Bruno Ben Moubamba. The video was not a deepfake, but its weaponisation contributed to growing unrest and an attempted military takeover (Ajder & Glick, 2021).

Despite several prominent examples of deepfakes, or their invocation, currently, it is 'shallowfakes' (Langguth et al., 2021), also called 'cheapfakes' (Paris & Donovan, 2019), that are more prevalent. In such fakes, videos claim to show something different from what they actually show or edit authentic video material to misrepresent the situation filmed, with manipulation occurring at a level that does not require AI. According to a report from Data and Society (an independent, non-profit, US research organisation, originally funded by Microsoft but now funded by numerous foundations, that seeks to ground evidence-based public debate about emerging technology), an example is the fake video that went viral in 2019 of US House Speaker Nancy Pelosi slurring her words after meeting with President Trump. The video was simply re-encoded at reduced speed, giving the impression of slurred speech. TikTok has a time filter, making such cheapfakes easy (Paris & Donovan, 2019). Brennen et al.'s (2020) content analysis of 225 pieces of English-language COVID-19 information rated false or misleading by fact-checkers in 2020 finds that there were no examples of deepfakes: rather, the manipulated content includes cheapfakes. For instance, one video includes images of bananas edited into a news segment to suggest that bananas can prevent or cure COVID-19.

Dynamics of False Information Online

To appreciate how and why false information online is spread, we must move beyond intentional deception and specific forms of disinformation (such as fake news and deepfakes) to consider false information more

broadly. This includes *misinformation*, namely, false information spread without intent to deceive, scholastically defined as 'that which contradicts the best expert evidence available at the time' (Vraga & Bode, 2020, p. 136). While the focus of this section is the contemporary digital environment, the spread of false information in other media environments is also instructive in highlighting continuities and disjunctures. To that end, we open this section by reporting on a study of scientific misinformation before social media.

Sleigh (2021) analyses the spread of misinformation on the harms of fluoridation of drinking water in the UK in the 1960s, as the government considered whether to fluoridate water supplies to improve people's teeth. False information spread through the activity of campaigning journalists, women's groups and establishment figures. They generated letters to local newspapers and distributed pamphlets from a pressure group that itself relied upon a network of a few relatively wealthy people with finances and confidence to research, write and print. It drew on anti-fluoridation campaigns already running in the USA and Australia, and its network of local British branches generated abundant local rumours. While making diverse medical and medical-related false claims, a persistent claim was that fluoride is a 'cumulative poison'. Ultimately, at the heart of this activity, and growing throughout the period studied, was protest at the lack of democratic process and fear of authoritarianism. The response from authorities was inadequate to address the concerns of the anti-fluoridation side: the authorities' response mixed science and politics, also presenting science as an unrealistically monolithic method and entity while failing to engage in conversations about legitimate doubt.

Many of these features are apparent in contemporary misinformation spread in digital environments. This includes network organisation (the interweaving of global and local activity, and initiation by a small core of well-resourced claims-makers); affective claims unsupported by science; and inadequate political responses. What has changed, however, is the scale and virality of the spread of false information, as well as involvement of bots in its propagation. While still in its infancy, we also know more about why people spread false information online. We document these changed features below.

Scale

Although the scale of contemporary fake news (narrowly defined) is relatively small, big data studies find that false information online is more prevalent. For instance, a study examining 1000 randomly selected Twitter status updates mentioning 'antibiotic(s)' in 2009 finds that 700 contained medical misinformation or malpractice (Scanfeld et al., 2010). Shao et al.'s (2016) analysis of online misinformation on Twitter across 2015 to 2016 finds that misinformation is produced in much larger quantities than fact-checking content. An analysis of 673 English-language tweets trending on Twitter using common COVID-19 terms and hashtags on 27 February 2020 finds that 25% included false information on COVID-19. The study also finds that misinformation is as likely to spread and engage users as the truth (Kouzy et al., 2020). An analysis in March 2020 of the most viewed COVID-19 YouTube videos finds over 25% of the top videos contained misleading information, totalling 62 million views worldwide (Li et al., 2020). Four months into the pandemic, around a third of people surveyed in six countries (Argentina, Germany, South Korea, Spain, the UK and the USA) said they had seen 'a great deal' of false or misleading information about COVID-19 on social media and messaging apps (Newman et al., 2020). This prevalence on social media contrasts sharply with false information in traditional news environments. An analysis of 38 million articles published in English-language traditional media around the world finds that just under 3% of the overall COVID-19 conversation comprised misinformation, itself largely driven by President Trump (Evanega et al., 2020).

Virality

Big data studies demonstrate that false information is contagious online. In a study of the differential diffusion of all verified true and false news stories distributed on Twitter from 2006 to 2017, Vosoughi et al. (2018) find that falsehood diffuses significantly farther, faster, deeper and more broadly than the truth in all categories of information. They find that false stories inspire fear, disgust and surprise in replies, whereas true stories inspire anticipation, sadness, joy and trust. Other Twitter-based big data studies similarly show that misinformation spreads faster and more widely across Twitter, with fact-checking content typically lagging that of misinformation or false rumours by 10–20 hours (Shao et al., 2016; Zubiaga

et al., 2016) and that low-credibility content is equally or more likely to spread virally as fact-checked articles (Shao et al., 2018).

Social media platforms can reduce the virality of false information, but rarely seem to do so. For instance, in preparation for the 2020 US presidential elections, Twitter temporarily introduced forms of friction in the weeks leading up to election day to discourage spread of false information. One such form was to encourage (via prompts) 'Quote Tweets' instead of 'Retweets', so that people would add commentary when amplifying content, thereby giving them an extra moment to consider why and what they were adding to the conversation. Since making the change, Twitter reported that Retweets and Quote Tweets combined decreased by 20% and that the change slowed the spread of misleading information by virtue of an overall reduction in the amount of sharing on the platform. However, reducing user engagement goes against the business models of dominant social media platforms, and soon after the election, Twitter stopped this form of friction, re-enabling standard Retweet behaviour (Gadde & Beykpour, 2020, November 12).

Spreaders

Studies show that false information on social media is spread both by software robots ('bots') designed to amplify the reach (Shao et al., 2017) and by humans (Mustafaraj & Metaxas, 2017; Vosoughi et al., 2018), especially by partisans, politicians, celebrities, public figures, certain demographics and journalists.

A 'bot' is a computer software program designed to execute commands, protocols or routine tasks. Bots are often used to flood social media networks with false information and can amplify marginal voices and ideas by inflating the number of likes, shares and retweets they receive, creating an artificial sense of momentum or relevance. Bots exploit our cognitive and social biases by, for example, creating the appearance of popular grassroots campaigns ('astroturfing') to manipulate attention and target influential users to induce them to reshare false information. Bots were deployed by government actors as early as 2011 (in Syria) and 2012 (in Argentina) (Bradshaw & Howard, 2017). In Qatar, across the 2017–2018 Gulf crisis, in which Qatar was blockaded by Saudi Arabia, Egypt, Bahrain and the United Arab Emirates, Twitter bots were used to disseminate propaganda that demonised Qatar and its government, reflecting the demands of the blockading countries. For instance, the bots manipulated Twitter trends in

the Gulf (these then amplified by mainstream Western news outlets) and fabricated evidence of popular hostility to the Qatari regime among ordinary Qataris (Jones, 2019). According to a report by the Atlantic Council (a US non-partisan think tank that seeks to galvanise American leadership to shape solutions to global challenges), in Mexico, a country dangerous to life for both journalists and politicians, disinformation in the 2018 elections (Andrés Manuel López Obrador v. Ricardo Anaya) deployed mainly political bots to spread specific electoral messages and drown out organic conversation about political candidates. Commercial bots were also hired for financial and political gain: on Facebook and Twitter, this consisted of commercial groups that coordinated large-scale responses to posts in return for payment (Pérez Argüello & Barojan, 2019). The amplification of low-credibility sources by bots is demonstrated by Shao et al.'s (2017) analysis of 14 million messages spreading 400,000 articles on Twitter during and following the 2016 US presidential election: they show that bots are particularly active in amplifying low-credibility sources before an article goes viral, targeting users with many followers through replies and mentions. They argue that people are vulnerable to such manipulation, retweeting bots who post low-credibility content just as much as they retweet other humans.

Various studies (largely US-dominated) examine politicians, partisans and partisan media sharing false information (Benkler et al., 2018; Cano-Orón et al., 2021; Gorrell et al., 2019; Guess et al., 2019; Ross & Rivers, 2018; Vargo et al., 2018). For instance, Benkler et al.'s (2018) network analysis of four million stories relating to the US presidential election and national politics from 2015 to 2018 finds that the right-wing ecosystem is more insular and skewed towards the extreme, where even leading news organisations (Fox and Breitbart) do not observe truth-seeking norms. Their audiences have become used to receiving belief-consistent news and abandon outlets that insist on facts when these are inconsistent with partisan narratives (Benkler, 2020, p. 49). Moving beyond the USA, a big data study of 'leavers' and 'remainers' on Twitter in the run-up to the 2016 'Brexit' referendum shows uptake of misleading claims from the 'Leave' campaign was high, dwarfing any evidence of Russian influence (Gorrell et al., 2019).

US-based studies conducted during presidential campaigns find that certain types of people are more likely to share false information on social media: namely, conservative, far-right, older and politically engaged (Grinberg et al., 2019; Guess et al., 2019). Grinberg et al. (2019) suggest

that heightened engagement by older adults could result from cognitive decline, lack of digital media literacy, stronger motivated reasoning, or cohort effects. Moving beyond the USA, a nationally demographically representative British study in 2018 finds that of those who shared news about politics at least once a month, 43% acknowledge that some shared was false or exaggerated; 29% admitted to unintentionally sharing false news; and 17% admitted to knowingly sharing disinformation. Those younger than 45 and those over 65 are more likely to share false news than middle-aged groups, as are those with higher levels of interest in politics (Chadwick & Vaccari, 2019). In China, statistics from popular messaging app WeChat show that rural citizens are more apt to share fake news than city residents (Deng, 2019, January 22).

In terms of spreading false COVID-19 information, public figures play a prominent role. For example, Brennen et al.'s (2020) study of 225 pieces of English-language COVID-19 information rated false or misleading by fact-checkers and published in early 2020 finds that in terms of sources, politicians, celebrities and other prominent public figures made up just 20% of the claims but accounted for 69% of social media engagement. Facebook's own research shows that a small number of posters and commenters were responsible for a large amount of anti-vaccine content: of nearly 150,000 posters in Facebook groups disabled for COVID-19 misinformation, 5% were producing half of all posts (Schechner et al., 2021, September 17).

Significantly, journalists themselves are not immune from sharing false information online. Despite a long history of fact-checking within the journalistic profession, and despite being on guard, many American and British journalists report being tricked by false information: 80% of 803 survey respondents (conducted in 2018) admitted to believing false information at some point, although most state that this occurrence is rare (Persen & Woolley, 2021).

Why Share False Information Online?

Why do people share false information online? To date, this has not been studied extensively, and it is difficult to assess motivation from examining content alone. When people are directly asked, reasons for sharing false information include a desire to 'troll', political partisanship and belief that the information is true (Brennen et al., 2020). Surveys of 552 journalism students in a Spanish university finds that over half said they have shared a

piece of fake news 'by mistake' (26%), 'to play a joke' (24%), 'it looked legitimate' (19%), 'it had a shocking headline' (7%), 'it said what I would like to happen' (6%), 'I agreed with the information' (3%) and 'to spread a rumour' (2%) (Tejedor et al., 2021). A national survey in China in 2018 finds that respondents who encountered and shared fake news more frequently exhibited higher levels of information overload (Tang et al., 2021). Taking a more ecological approach, a big data study suggests that disinformation thrives when news outlets fail to cater to people's interests. Analysis of Italy's entire supply of news (fake news vetted by fact-checkers and general news published online and offline) for COVID-19 compared to news demand (captured via Google Trends data for Italy) across December 2019 to August 2020 finds that the fake news supply is more reactive than general news to people's interests and thrives when there is a mismatch between what people are interested in and what news outlets provide (Gravino et al., 2021).

Studies suggest that people are quite ineffective at recognising deception online (Lewis et al., 2022; Rubin et al., 2016, pp. 7–8). Various reasons have been proposed for this. Firstly, most people show an inherent truth bias: they tend to assume that the information they receive is true and reliable. Secondly, some people are very receptive to ideas that they do not fully understand. Thirdly, confirmation bias, where people unwittingly seek or interpret information in ways that conform with their existing beliefs or hypotheses, can cause people to see only what they want to see. Indeed, pre-existing attitudes seem to be a common factor in reasons for sharing rumours and conspiracy theories (Douglas et al., 2019, p. 18; Greenhill & Oppenheim, 2017). A fourth reason for being ineffective at recognising deception is our reliance on others to make credibility assessments. For instance, Metzger et al.'s (2010) US-based focus group data show that most users rely on others to make credibility assessments, often via group-based tools. In a related vein, Sterrett et al.'s (2019) survey experiment of American adults finds that people's trust in news on social media is strongly related to who shares it: even if it comes from an unknown outlet (and hence is potentially false), they are willing to share it if it comes from a trusted public figure. As with all media effects scholarship, however, the results are not clear-cut (Tamul et al., 2019). Not least, this is because every component of the information transmission channel (the source, content, form, channel and individual attributes of the recipient) can influence credibility.

Conclusion

The question of whether leaders should lie to their publics has proven of enduring interest in studies on citizen-political communications. The debate was reinvigorated by early-twentieth-century mass communication theorists who viewed propaganda, manipulation and deception as essential to democracy and managing large populations in the public interest. Against deception by leaders, others point out its erosion of social trust and democratic foundations, and demand clear and publicly accepted rules for any exceptions to truth-telling: these exceptions should be occasional rather than routine. Unfortunately, scholars worldwide have observed many techniques of governmental and political deception that have long been in play for national security reasons, during elections and even during periods of governing. However, once mainly the domain of intelligence agencies and well-resourced organisations or countries, deception is now propagated by diverse actors in the contemporary media ecology, making it harder for authorities to subdue the spread of false information online.

Two important forms of contemporary disinformation are fake news and deepfakes. Since it was recognised as a key form of online disinformation in 2016, attracting widespread political concern, the affective and deceptive nature of fake news online has been clarified by scholars. Studies find that the scale of fake news varies across platforms and time and is relatively small, but nonetheless damaging to the health of the *civic body* both during elections and pandemics. This is because of their wider media agenda-setting effects and because of their high absolute levels of engagement among certain users. Since 2017, deepfakes have emerged from developments in AI and machine learning. Along with the more common deceptive forms of shallowfakes and cheapfakes, deepfakes could further disrupt our confidence in the believability of broadcast content. While still comparatively rare, the rise of deepfake synthetic media is of concern as they are becoming easier to produce, be this via GANs (requiring far less training data) or via apps (producing formulaic, but widely available, deepfakes such as faceswaps). Their appeal to the architects of disinformation lies in that they are hard for people to recognise as false, especially when content presented is neutral rather than suspiciously out of character; they incubate uncertainty; and they can elicit more visceral, emotional and empathic responses than text-based media. The very idea of deepfakes is also being used by the powerful to avoid accountability and for political gain, by suggesting that anything can be fake.

Considering online false information more broadly (beyond deliberate attempts to deceive), it is hard to assess its scale as social media platforms have sole, proprietary access to their data. Nonetheless, big data studies suggest that false information online is prevalent. Studies show that false information during election campaigns worldwide is spread both by humans and bots: and, unsurprisingly, studies from the USA and UK show that false information is shared mainly by politicians, partisan media, partisans and the politically engaged. More user studies across different countries are needed, but so far we know that false information spreaders tend to be older, conservative and on the far-right (in the USA); older and younger people interested in politics (in the UK); and rural citizens (in China). In terms of COVID-19 false information, politicians, celebrities and other prominent public figures account for most of the total social media engagement. Even American and British journalists themselves are not immune from sharing false information online, despite a long history of fact-checking within the journalistic profession and despite being on guard.

The question of why people share false information online has not been studied extensively, but reasons uncovered include a desire to 'troll' and spread rumours, political partisanship, a belief that the information is true, ineffectiveness at recognising deception, for fun, a reliance on mental short cuts to evaluate credibility, trusting the news on social media if it comes from a trusted public figure, a mismatch between what people are interested in and what news outlets provide and congruence with individuals' pre-existing worldviews. We are bad at recognising deception due to our pre-existing biases and attitudes, and reliance on others to make credibility assessments.

Whether the circulation of false information is a newly problematic situation is open to question, as trust in the news to tell the truth has been under strain for decades across the world and fake news is recognisable from the birth of the printed newspaper centuries ago. We also note that misinformation predating social media was spread through features we would recognise today. This includes network organisation (the interweaving of global and local activity, and initiation by a small core of well-resourced claims-makers); affective claims unsupported by science; inadequate political responses; and sharing by prominent public figures, politicians, partisans, the politically engaged, partisan media and, sometimes, journalists. What has changed, however, is the scale and virality of the spread of false information, as well as involvement of bots deployed by

state actors and during elections. It is worth remembering Chap. 1's findings: globally, people regard domestic politicians as by far the most responsible for false and misleading information online (far more than foreign governments or journalists), with the greatest concern over stories where facts are twisted to push particular agendas. It would seem, then, that those who advocate that rulers should not be deceptive, because of its erosion of social trust and democratic foundations, were right.

As subsequent chapters will show, our concern with the detrimental impacts on the *civic body* from contemporary false information arises from its reach and prevalence when combined with the energising force of affect and emotion (Chap. 5) delivered via the profiling and targeting of audiences' datafied emotion (Chap. 6).

References

Abrams, S. (2016). Beyond propaganda: Soviet active measures in Putin's Russia. *Connections: The Quarterly Journal, 15*(1), 5–31. https://doi.org/10.11610/Connections.15.1.01

Ajder, H., & Glick, J. (2021). *Just joking! Deepfakes, satire and the politics of synthetic media*. WITNESS and MIT Open Documentary Lab. Retrieved April 13, 2022, from https://cocreationstudio.mit.edu/just-joking/

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research and Politics, 6*(2). https://doi.org/10.1177/2053168019848554

Bakir, V. (2015). News, agenda-building and intelligence agencies: A systematic review of the field from the discipline of journalism, media and communications. *International Journal of Press/Politics, 20*(2), 131–144. https://doi.org/10.1177/1940161214566693

Bakir, V. (2016). *Torture, intelligence & sousveillance in the war on terror*. Routledge. (Original work published 2013).

Bakir, V. (2019). *Intelligence elites & public accountability: Relationships of influence with civil society*. Routledge. (Original work published 2018).

Bakir, V., & Barlow, D. (2007). *Communication in the age of suspicion: Trust and the media*. Palgrave Macmillan.

Bakir, V., Herring, E., Miller, D., & Robinson, P. (2018a). Lying and deception in politics. In J. Meibauer (Ed.), *The Oxford handbook of politics and lying* (pp. 529–540). Oxford University Press.

Bakir, V., Herring, E., Miller, D., & Robinson, P. (2018b). Organized persuasive communication: A new conceptual framework for research on public relations, propaganda and promotional culture. *Critical Sociology, 45*(3), 311–328. https://doi.org/10.1177/0896920518764586

Baym, G. (2005). The daily show: Discursive integration and the reinvention of political journalism. *Political Communication, 22*(3), 259–276. https://doi.org/10.1080/10584600591006492

Benkler, Y. (2020). A political economy of the origins of asymmetric propaganda in American media. In W. L. Bennett & S. Livingston (Eds.), *The disinformation age: Politics, technology and disruptive communication in the information age* (pp. 43–66). Cambridge University Press. https://doi.org/10.1017/9781108914628

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bickert, M. (2020, January 6). Enforcing against manipulated media. *Meta*. Retrieved April 13, 2022, from https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

Bok, S. (1999). *Lying: Moral choice in public and private life*. Harvester.

Boyd-Barrett, O. (2020). *Russiagate. Disinformation in the age of social media*. Routledge.

Bradshaw, S., & Howard, P. N. (2017). *Troops, trolls and troublemakers: A global inventory of organized social media manipulation*. Working paper no. 2017.12 (pp. 1–37). Oxford Internet Institute. Retrieved April 13, 2022, from https://ora.ox.ac.uk/objects/uuid:cef7e8d9-27bf-4ea5-9fd6-855209b3e1f6

Brennen, J. S., Simon, F., Howard, N. P., & Nielsen K. R. (2020). *Types, sources, and claims of COVID-19 misinformation*. Reuters Institute. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation

Briant, E. L. (2015). Allies and audiences: Evolving strategies in defense and intelligence propaganda. *The International Journal of Press/Politics, 20*(2), 145–165. https://doi.org/10.1177/1940161214552031

Brovdiy, Y. (2020). *Disinformation in times of COVID-19: Reinforcing the responses of the European Union and the United States*. College of Europe Policy Brief, # 5. Retrieved April 13, 2022, from https://www.coleurope.eu/research-paper/disinformation-times-covid-19-reinforcing-responses-european-union-and-united-states

Cano-Orón, L., Calvo, D., López García, G., & Baviera, T. (2021). Disinformation in Facebook ads in the 2019 Spanish general election campaigns. *Media and Communication, 9*(1), 217–228. https://doi.org/10.17645/mac.v9i1.3335

Chadwick, A., & Vaccari, C. (2019). *News sharing on UK social media: Misinformation, disinformation, and correction*. Loughborough University. Retrieved April 13, 2022, from https://www.lboro.ac.uk/media/media/subjects/communication-media-studies/downloads/chadwick-vaccari-o3c-1-news-sharing-on-uk-social-media-1.pdf

Deng, I. (2019, January 22). Tencent's fake news debunkers reached nearly 300 million WeChat users last year. *South China Morning Post*. https://www.scmp.

com/tech/apps-social/article/2183124/tencents-fake-news-debunkers-reached-nearly-300-million-wechat

Dentith, M. R. X. (2017). The problem of fake news. *Public Reason, 8*(1–2), 65–79. https://philpapers.org/archive/DENTPO-31.pdf

Di Meco, L. (2019). *#ShePersisted. Women, politics, & power in the new media world* (pp. 1–58). The Wilson Center. Retrieved June 23, 2022, from https://static1.squarespace.com/static/5dba105f102367021c44b63f/t/5dc431aac6bd4e7913c45f7d/1573138953986/191106+SHEPERSISTED_Final.pdf

Digital, Culture, Media and Sport Committee. (2018, July 24). *Disinformation and 'fake news': Interim Report*. House of Commons 363. Retrieved April 13, 2022, from https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/363.pdf

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2020). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics, 26*(1), 69–91. https://doi.org/10.1177/1940161220944364

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Advances in Political Psychology, 40*, 3–35. https://doi.org/10.1111/pops.12568

Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two- dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association, 43*(2), 97–116. https://doi.org/10.1080/23808985.2019.1602782

EU DisinfoLab. (2021, October 13). *The role of "media" in producing and spreading disinformation campaigns*. Retrieved April 13, 2022, from https://www.disinfo.eu/publications/the-role-of-media-in-producing-and-spreading-disinformation-campaigns/

EU East StratCom Task Force. (2017, January 19). *Means, goals and consequences of the pro-Kremlin disinformation campaign*. ISPI. Retrieved April 13, 2022, from http://www.ispionline.it/it/pubblicazione/means-goals-and-consequences-pro-kremlin-disinformation-campaign-16216

European Science Data Hub. (2019, December 4). *Deepfakes, shallowfakes and speech synthesis: Tackling audiovisual manipulation*. European Parliamentary Research Service. Retrieved April 13, 2022, from https://sciencemediahub.eu/2019/12/04/deepfakes-shallowfakes-and-speech-synthesis-tackling-audiovisual-manipulation/

Evanega, S., Lynas, M., Adams, J., & Smolenyak, K. (2020, October 1). *Coronavirus misinformation: Quantifying sources and themes in the COVID-19 'Infodemic'*. Retrieved April 13, 2022, from https://int.nyt.com/data/documenttools/evanega-et-al-coronavirus-misinformation-submitted-07-23-20-1/080839ac0c22bca8/full.pdf

Fletcher, R., Cornia, A., Graves, L., & Nielsen, R. K. (2018). *Measuring the reach of "fake news" and online disinformation in Europe*. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/Measuring%20the%20reach%20of%20fake%20news%20and%20online%20distribution%20in%20Europe%20CORRECT%20FLAG.pdf

Gadde, V., & Beykpour, K. (2020, November 12). *An update on our work around the 2020 US Elections.* Retrieved April 13, 2022, from https://blog.twitter.com/en_us/topics/company/2020/2020-election-update.html

Gelfert, A. (2018). Fake news: A definition. *Informal Logic, 38*(1). https://doi.org/10.22329/il.v38i1.5068

Gorrell, G., Bakir, M. E., Roberts, I., Greenwood, M. A., Iavarone, B., & Bontcheva, K. (2019). Partisanship, propaganda and post-truth politics: Quantifying impact in online debate. *Journal of Web Science, 7*. https://doi.org/10.34962/jws-84

Gravino, P., Prevedelloa, G., Gallettia, M., & Loreto, V. (2021). Assessing disinformation through the lens of news supply and demand during the COVID-19 outbreak. *Nature Human Behaviour*. Preprint retrieved from https://doi.org/10.21203/rs.3.rs-577571/v1

Greenhill, K. M., & Oppenheim, B. (2017). Rumor has it: The adoption of unverified information in conflict zones. *International Studies Quarterly, 61*, 660–676. https://doi.org/10.1093/isq/sqx015

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science, 363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guess, A., Nyhan, B., & Reifler, J. (2018). *Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign*. Retrieved April 13, 2022, from https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances, 5*(1). https://doi.org/10.1126/sciadv.aau4586

Habermas, J. (1996). *The structural transformation of the public sphere: An inquiry into a category of Bourgeois Society* (T. Burger & F. Lawrence, Trans.). Polity Press. (Original work published 1962).

Hanson, F., O'Connor, S., Walker, M., & Courtois, L. (2019). *Hacking democracies: Cataloguing cyber-enabled attacks on elections, Policy Brief 16.* Australian Strategic Policy Institute. Retrieved April 13, 2022, from https://www.aspi.org.au/report/hacking-democracies

Hermida, A. (2016). Trump and the triumph of affective news when everyone is the media. In D. Lilliker, E. Thorsen, D. Jackson, & A. Veneti (Eds.), *US election analysis 2016: Media voters and the campaign early reflections from leading academics* (p. 76). Centre for the Study of Journalism, Culture and

Community. Bournemouth University. Retrieved April 13, 2022, from http://eprints.bournemouth.ac.uk/24976/1/US%20Election%20Analysis%202016%20-%20Lilleker%20Thorsen%20Jackson%20and%20Veneti%20v1.pdf

Horne, B. D., & Adal, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *The Workshops of the Eleventh International AAAI Conference on Web and Social Media AAAI Technical Report.* https://ojs.aaai.org/index.php/ICWSM/article/view/14976/14826 759

Jones, M. O. (2019). The Gulf information war| propaganda, fake news, and fake trends: The weaponization of Twitter bots in the Gulf Crisis. *International Journal of Communication, 13.* https://ijoc.org/index.php/ijoc/article/view/8994/2604

Judson, E., Atay, A., Krasodomski-Jones, A., Lasko-Skinner, R., & Smith, J. (2020, October). *The contours of state-aligned gendered disinformation online.* Demos, London. Retrieved June 23 2022, from https://demos.co.uk/project/engendering-hate-the-contours-of-state-aligned-gendered-disinformation-online/

Kaufmann, C. (2004). Threat inflation and the failure of the marketplace of ideas: The selling of the Iraq War. *International Security, 29*(1), 5–48. http://www.mitpressjournals.org/doi/abs/10.1162/0162288041762940

Kouzy, R., Jaoude, J. A., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., & Baddour, K. (2020). Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter. *Cureus, 12*(3). https://doi.org/10.7759/cureus.7255

Kurlantzick, J. (2020). *How China ramped up disinformation efforts during the pandemic.* Council on Foreign Relations. Retrieved April 13, 2022, from https://www.jstor.org/stable/resrep29835

Langguth, J., Pogorelov, K., Brenner, S., Filkukova, P., & Schroeder, D. (2021). Don't trust your eyes: Manipulation of visual media in the age of deepfakes. *Frontiers in Political Communication.* https://doi.org/10.3389/fcomm.2021.632317

Lashmar, P. (2013). Urinal or conduit? Institutional information flow between the UK intelligence services and the news media. *Journalism, 14*(8), 1024–1040. https://doi.org/10.1177/1464884912472139

Lashmar, P. (2017). No more sources? The impact of Snowden's revelations on journalists and their confidential sources. *Journalism Practice, 11*(6), 665–688. https://doi.org/10.1080/17512786.2016.1179587

Lasswell, H. D. (1936). *Politics: Who gets what, when, how.* Whittlesey House.

Lasswell, H. D. (1971). The theory of political propaganda. *The American Political Science Review, 21*, 627–631. https://doi.org/10.2307/1945515. (Original work published 1927).

Lewis, A., Vu, P., Duch, R. M., & Chowdhury, A. (2022). *Do content warnings help people spot a deepfake? Evidence from two experiments.* The Royal Society.

Retrieved April 13, 2022, from https://royalsociety.org/-/media/policy/projects/online-information-environment/do-content-warnings-help-people-spot-a-deepfake.pdf

Li, Y., Chang, M.-C., & Lyu, S. (2018). In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. https://ieeexplore.ieee.org/document/8630787

Li, H. O.-Y., Bailey, A., Huynh, D., & Chan, J. (2020). YouTube as a source of information on COVID-19: A pandemic of misinformation? *British Medical Journal of Global Health, 5*(1–6), e002604. https://gh.bmj.com/content/5/5/e002604.info

Lippman, W. (1922). *Public opinion*. Free Press.

Lippman, W. (1993). *The phantom public*. Transaction Publishers. (Original work published 1927).

Lomas, N. (2017, April 25). Lyrebird is a voice mimic for the fake news era. *TechCrunch*. https://techcrunch.com/2017/04/25/lyrebird-is-a-voice-mimic-for-the-fake-news-era/

Machiavelli, N. (2003). *The prince* (G. Bull, Trans.). Penguin. (Original work published 1532).

MacLean, N. (2020). "Since we are greatly outnumbered": Why and how the Koch network uses disinformation to thwart democracy. In W. L. Bennett & S. Livingston (Eds.), *The disinformation age: Politics, technology and disruptive communication in the information age* (pp. 120–149). Cambridge University Press. https://doi.org/10.1017/9781108914628

Mano, W. (2020). Alternative responses to presidential tweets on elections in Africa: A new counter-power? In M. Ndlela & W. Mano (Eds.), *Social media and elections in Africa, Volume 1: Theoretical perspectives and election campaigns* (pp. 61–74). Palgrave Macmillan, Springer Nature.

Mare, A., & Matsilele, T. (2020). Hybrid media system and the July 2018 elections in "post-Mugabe" Zimbabwe. In M. Ndlela & W. Mano (Eds.), *Social media and elections in Africa, Volume 1: Theoretical perspectives and election campaigns* (pp. 147–176). Palgrave Macmillan, Springer Nature.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication, 60*(3), 413–439. https://doi.org/10.1111/j.1460-2466.2010.01488.x

Mustafaraj, E., & Metaxas, P. T. (2017). The fake news spreading plague: Was it preventable? In *Proceedings of the 2017 ACM on web science conference* (pp. 235–239). https://doi.org/10.1145/3091478.3091523

Ndlela, M. N. (2020). Social media algorithms, bots and elections in Africa. In M. Ndlela & W. Mano (Eds.), *Social media and elections in Africa, Volume 1: Theoretical perspectives and election campaigns* (pp. 13–37). Palgrave Macmillan, Springer Nature.

Nelson, J. L., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New Media & Society, 20*(10), 3720–3737. https://doi.org/10.1177/1461444818758715

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2018). *Reuters Institute digital news report 2018*. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/digital-news-report-2018.pdf

Newman, N., Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Reuters Institute digital news report 2020*. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *PNAS, 119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

Nisbet, E. C., & Kamenchuk, O. (2019). The psychology of state-sponsored disinformation campaigns and implications for public diplomacy. *The Hague Journal of Diplomacy, 14*(1–2), 65–82. https://doi.org/10.1163/1871191X-11411019

Oreskes, N., & Conway, E. M. (2011). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Press.

Oreskes, N., Conway, E. M., & Tyson, C. (2020). How American businessmen made us believe that free enterprise was indivisible from American democracy: The National Association of Manufacturers' propaganda campaign 1935–1940. In W. L. Bennett & S. Livingston (Eds.), *The disinformation age: Politics, technology and disruptive communication in the information age* (pp. 95–119). Cambridge University Press. https://doi.org/10.1017/9781108914628

Paris, B., & Donovan, J. (2019). *Deepfakes and cheap fakes: The manipulation of audio and visual evidence*. Data & Society. Retrieved April 13, 2022, from https://datasociety.net/wp-content/uploads/2019/09/DSDeepfakesCheapFakesFinal-1-1.pdf

Paschen, J. (2019). Investigating the emotional appeal of fake news using artificial intelligence and human contributions. *Journal of Product & Brand Management, 29*(2). https://www.emerald.com/insight/content/doi/10.1108/JPBM-12-2018-2179/full/html

Pérez Argüello, M. F., & Barojan, D. (2019). Mexico. In L. Bandeira, D. Barojan, R. Braga, J. L. Peñarredonda, & M. F. Pérez Argüello (Eds.), *Disinformation in democracies: Strengthening digital resilience in Latin America* (pp. 20–29). Atlantic Council. Retrieved April 13, 2022, from https://www.atlanticcouncil.org/in-depth-research-reports/report/disinformation-democracies-strengthening-digital-resilience-latin-america/

Pérez Dasilva, J., Meso Ayerdi, K., & Mendiguren Galdospin, T. (2021). Deepfakes on Twitter: Which actors control their spread? *Media and Communication, 9*(1), 301–312. https://doi.org/10.17645/mac.v9i1.3433

Perloff, R. M. (2018). *The dynamics of political communication: Media and politics in a digital age*. Routledge.

Persen, K. A. C., & Woolley, S. C. (2021). Computational propaganda and the news: Journalists' perceptions of the effects of digital manipulation on reporting. In M. Boler & E. Davis (Eds.), *Affective politics of digital media* (pp. 245–260). Routledge.

Plato. (2007). *The republic* (H. D. P. Lee & D. Lee, Trans.). Penguin Classics. (Original work published 381 B.C.)

Rid, T. (2021). *Active measures: The secret history of disinformation and political warfare*. Profile Books.

Ross, A. S., & Rivers, D. J. (2018, April–June). Discursive deflection: Accusation of "fake news" and the spread of mis- and disinformation in the tweets of President Trump. *Social Media + Society, 112*. https://doi.org/10.1177/2056305118776010.

RSF. (2017). *Predators of press freedom use fake news as a censorship tool*. Retrieved April 13, 2022, from https://rsf.org/en/news/predators-press-freedom-use-fake-news-censorship-tool

Rubin, V. L., Conroy, N. J., Chen, Y., & Cornwell, S. (2016). Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 7–17). http://www.aclweb.org/website/oldanthology/W/W16/W16-0802.pdf

Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control, 38*(3), 182–188. https://doi.org/10.1016/j.ajic.2009.11.004

Schechner, S., Horwitz, J., & Glazer, E. (2021, September 17). How Facebook hobbled Mark Zuckerberg's bid to get America vaccinated. *Wall Street Journal*. https://www.wsj.com/articles/facebook-mark-zuckerberg-vaccinated-11631880296

Sessa, M. G. (2020, December 4). *Misogyny and misinformation: An analysis of gendered disinformation tactics during the COVID-19 pandemic*. EU DisinfoLab. Retrieved June 23, 2022, from https://www.disinfo.eu/publications/misogyny-and-misinformation:-an-analysis-of-gendered-disinformation-tactics-during-the-covid-19-pandemic/

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on World Wide Web* (pp. 745–750). https://arxiv.org/abs/1603.01511

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (2017). *The spread of fake news by social bots*. Preprint retrieved from http://arxiv.org/abs/1707.07592

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications, 9*, 4787. https://doi.org/10.1038/s41467-018-06930-7

Simonite, T. (2022, March 17). A Zelensky deepfake was quickly defeated. The next one might not be. *WIRED*. https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/

Sleigh, C. (2021). *Fluoridation of drinking water in the UK, c.1962-67. A case study in scientific misinformation before social media*. The Royal Society. Retrieved April 13, 2022, from https://royalsociety.org/-/media/policy/projects/online-information-environment/oie-water-fluoridation-misinformation.pdf

Sterrett, D., Malato, D., Benz, J., Kantor, L., Tompson, T., Rosenstiel, T., Sonderman, J., & Loker, K. (2019). Who shared it? Deciding what news to trust on social media. *Digital Journalism, 7*(6), 783–801. https://doi.org/10.1080/21670811.2019.1623702

Tamul, D. J., Ivory, A. H., Hotter, J., & Wolf, J. (2019). All the President's tweets: Effects of exposure to Trump's "fake news" accusations on perceptions of journalists, news stories, and issue evaluation. *Mass Communication and Society, 23*(3), 301–330. https://doi.org/10.1080/15205436.2019.1652760

Tang, S., Willnat, L., & Zhang, H. (2021). Fake news, information overload, and the third-person effect in China. *Global Media and China, 6*(4), 492–507. https://doi.org/10.1177/20594364211047369

Tatsumi, Y., Kennedy, P., & Li, J. (2020). *Taiwan security brief: Disinformation, cybersecurity, & energy challenges*. Stimson Centre. Retrieved April 13, 2022, from https://www.stimson.org/2019/disinformation-cybersecurity-and-energy-challenges/

Tejedor, S., Portalés-Oliva, M., Carniel-Bugs, R., & Cervi, L. (2021). Journalism students and information consumption in the era of fake news. *Media and Communication, 9*(1), 338–350. https://doi.org/10.17645/mac.v9i1.3516

Thakur, D., & Hankerson, D. L. (2021). *Facts and their discontents: A research agenda for online disinformation, race, and gender*. Center for Democracy & Technology. https://osf.io/3e8s5/.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1). https://doi.org/10.1177/2056305120903408

Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society, 20*(5), 2028–2049. https://doi.org/10.1177/1461444817712086

Vizoso, A., Vaz-Álvarez, M., & López-García, X. (2021). Fighting deepfakes: Media and internet giants' converging and diverging strategies against hi-tech misinformation. *Media and Communication, 9*(1), 291–300. https://doi.org/10.17645/mac.v9i1.3494

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Vraga, E., & Bode, V. (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication, 37*(1), 136–144. https://doi.org/10.1080/10584609.2020.1716500

Waszczykowski, W. (2015). *The battle for the hearts and minds: Countering propaganda attacks against the Euro-Atlantic community*. NATO Parliamentary Assembly, Committee on the Civil Dimension of Security. Retrieved April 13, 2022, from https://connections-qj.org/article/battle-hearts-and-minds-countering-propaganda-attacks-against-euro-atlantic-community

Winston, B., & Winston, M. (2021). *The roots of fake news: Objecting to objective journalism*. Routledge.

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS One, 11*(3), e0150989. https://doi.org/10.1371/journal.pone.0150989

CHAPTER 5

Feeling-Into the Civic Body: Affect, Emotions and Moods

Introduction

Affect, emotions and moods all play an important role in social and political life. They motivate, excite, colour experience, are core to communication, help us perceive value and inform our judgements (including those of a moral sort). This chapter accounts for the energising role of feelings in relation to false information throughout the *civic body*. Using feelings as a catch-all term to describe affects, emotions and moods, as well as reactions to stimuli we may not be aware of, we start by charting the trajectory of the role of feelings in understanding citizen-political communications. Their persuasive importance was recognised millennia ago and this has been recognised anew in recent decades with the advent of neuroscience and the understanding that emotions are important for decisions and judgements. Many studies address how governments can try to best manage public feeling, and hence behaviour, and we highlight three main mechanisms: discursive, decision-making based and datafied.

Claims that we live in a post-truth condition are prevalent, with appeals to emotion and personal belief argued to be more influential in shaping public opinion than objective facts. While the relative importance of emotion and facts in everyday life is difficult to ascertain, we demonstrate that the media from which people would normally derive their facts (namely, news media and social media) have become more emotionalised and affective. We suggest that we live in an informational environment that is

V. Bakir, A. McStay, *Optimising Emotions, Incubating Falsehoods*,
https://doi.org/10.1007/978-3-031-13551-4_5

sub-optimal for a healthy *civic body*. We exemplify this by examining challenges faced by governments in managing their population's feelings during the COVID-19 pandemic where uncertainty, anxiety and false information proliferate.

Feelings in Citizen-Political Communications

The role of affect, emotions and moods in understanding citizen-political communications has been uneven across the centuries. Their importance for persuasion was recognised several millennia ago, as well as by those engaged in twentieth-century social engineering and propaganda. Although challenged by Enlightenment-oriented discourse and much communications research, their importance has been refreshed in recent decades with the advent of identity-based conceptions of the political, as well as neuroscientific understanding that emotions underpin our decisions. In this section, we chart this trajectory, but first, we define our terms.

The distinction between affect and emotion has long been debated (Döveling et al., 2011). Writing on the 'passions' goes back at least as far as Aristotle in the *Nicomachean Ethics*. The word 'emotion' came into use in the English language in the seventeenth and eighteenth centuries, but the idea of *emotions* as mental states that can be systematically studied only emerged in the nineteenth century (Dixon, 2012). Although emotions have been studied across diverse disciplines, we are apparently 'not much closer on reaching consensus on what emotions are than we were in Ancient Greece' (Scarantino, 2018, p. 37). We recognise that emotion has physiological qualities, but also that emotion is not synonymous with these. Emotion has episodic, experiential, intentional, communicative, historical, cultural and social qualities too. A reasonable working definition is that emotions are 'internal states that arise following appraisals (evaluations) of interpersonal or intrapersonal events that are relevant to an individual's concerns' which in turn 'promote certain patterns of response' (Cowen et al., 2019). *Affect*, again studied by diverse disciplines, is broadly understood as feelings that are less fully formed and a 'general property of experience that has at least two features: pleasantness or unpleasantness (valence) and degree of arousal' (Barrett et al., 2019, p. 51). Important, too, is the more literal definition of affect, to be moved by something, potentially without a person or group being aware. Affect is part of every waking moment of life and not specific to instances of emotion, although all emotional experiences have affect at their core. A *mood* is a longer-term condition, and drawing on Heidegger's (2011 [1962])

phenomenology of moods and experience, McStay (2018, p. 164) defines moods as 'a way of being-in-the-world. They represent an attunement, that characterises "being-there" and the disclosure of how and what things are'.

The powerful role of feelings in persuasive communications was recognised in the era of classical Greek Democracy (5 BC) by Aristotle. He advocated that rhetors use 'pathos' (appeals to emotion), as well as 'ethos' (appeals to the speaker's character and personality) and 'logos' (appeals to rationality) (Aristotle, 1991). The contagious nature of emotion has also long been observed. David Hume, for example, in *A Treatise of Human Nature* (1978 [1739]) spoke of affections passing from one to another. Likewise, Gustave Le Bon's *The Crowd* (2016 [1896]) spoke of emotional contagion as undermining individual rational thought, exaggeration of sentiment, impulsiveness, force, destruction and absence of critical spirit.

As the early twentieth century ushered in the rise of mass communication technologies alongside expanded electorates, American scholars reflected on how these could be combined with psychological research to manage populations and engineer consent (Bernays, 1928a, b; Lasswell, 1936). For instance, Bernays applied the concept of the subconscious mind (pioneered by his psychoanalyst uncle, Sigmund Freud) to management of mass communications, blending the idea of subconscious messaging with theories of crowd psychology and herd instinct (Bernays, 1928a). Describing mass psychology and the 'group mind', Bernays posits that: 'In place of thoughts it has impulses, habits, and emotions' (Bernays, 1928a, p. 73). Also influenced by psychoanalytical concepts, Lasswell (1936, p. 317) describes how elite propaganda, which he regarded as necessary to manage the masses, is complicated by the 'changing emotional requirements of the community, moods of submissiveness, moods of self-assertion'. Certainly, studies of propaganda frequently analyse how it is designed to bypass rational thought (Quaranto & Stanley, 2021).

Despite these forays into mass media's influence on public feeling, early communication research largely emphasised cognitive aspects such as recall, learning, thoughts and beliefs, with emotions regarded as mere 'noise' (Konijn & ten Holt, 2011). Indeed, until the 1990s, information processing models assumed that affect and cognition were two antagonist forces, with cognition the pre-eminent force and affect something requiring control (Spezio & Adolphs, 2013 [2007]). However, neuroscience has since challenged presumptive claims that devalue emotional processes in decision-making (Barfar, 2019), with the influential works of António Damásio (1994, 2010) suggesting that emotions can enhance information

processing. Damásio (1994) showed in *Descartes' Error* that people with brain damage that makes them incapable of experiencing emotion or detecting it in others cannot function rationally: they cannot feel what decisions will make them (or others) happy or unhappy.

Within contemporary communications and cultural studies scholarship, the 'affective turn' has also become more pronounced in recent decades (Clough & Halley, 2007; Döveling et al., 2011; Bösel & Wiemer, 2020). As McStay (2013) notes, it was Deleuze's (1988 [1970]) monograph on Spinoza, and his later work with Guattari (Deleuze & Guattari, 2011 [1994]), that helped to return affect to contemporary critical attention within this field. Spinoza (1996 [1677]) opposed mind–body dualism, preferring instead a monism that more directly links mental goings-on with the body. For Spinoza, the mind and body work in parallel (as with neuroscientific accounts) and are indivisible, somehow made of the same substance. Such affective accounts explore drives, motivations, will, emotion, feelings and sensations. For Spinoza, these were central to being human. In Deleuzo-Guattarian terms, affect is an aesthetic activity in the way that artists are interested in generating intense experiences that take the body (including the brain) from one condition to another. McStay (2013, p. 4) argues that such conceptions of affect allow us to analyse media in terms of attention, attraction, stimulation, sensation, context and corporeal events.

Despite these interventions, our understanding of the significance of affect, emotion and mood in citizen-political relations has been hampered by the suffusion with Enlightenment principles of liberal democratic theory (the dominant mode of political organisation in Western democracies) (Wahl-Jorgensen, 2019). The Enlightenment path to knowledge idealises rational, dispassionate, autonomous and informed citizens; and it regards passions as something to be controlled and channelled, without recognising the orienting role of affect and emotion (Kant, 1998 [1781]). Following Enlightenment principles, Jurgen Habermas argues for a privileging of the rational in his (now archetypal) democratic ideal of the public sphere. This ultimately seeks a consensus among citizens by enabling all to speak rationally, through listening to each other's viewpoints and agreeing the best way forward (Habermas, 1984).

By contrast, questioning the very desirability of Habermas' rational, consensus-oriented model of deliberative democracy, Chantal Mouffe (2005) regards 'the political' as a space of power and unavoidable conflict between adversaries. In this affective, identity-based conception of the political, people are deeply embedded within their own communities and

passionately attached to their own conceptions of the common good. For Mouffe, a heartfelt clash of perspectives (rather than a universal, rational consensus) is central to democracy (Mouffe, 2005, p. 11) as it produces an agonistic debate where citizens can be heard and choose between real alternatives (Mouffe, 2013) rather than adopting a technocratic consensus. From Mouffe's perspective, communication that is persuasive, passionate and conflictual is desirable, as long as it does not seek to exclude others either discursively or literally. Discursive exclusion could constitute rendering adversaries as enemies through hate speech (that creates an othered object of disgust). Literal exclusion might constitute advocating physical elimination of the other side (as in genocide). Indeed, there is a long tradition of work within feminist, Black and queer scholarship that values the power of public feelings as important sources of knowledge about power, oppression and governance (Blackman, 2022).

As such, the 'emotional public sphere' (Lunt & Pantti, 2007; Richards, 2007) where emotions are expressed, shared and managed draws attention to civic gains from affective, mediated engagement. Often, it is the advent of new media forms that prompts scholarship on their affective affordances. While today it is the affective nature of social media that is attracting widespread attention, and tomorrow it may be the affective nature of biometric forms of emotional AI, in the early 1990s it was the rise of satellite TV and its live imagery of far-flung conflicts that absorbed many scholars of media, politics and tele-diplomacy. Arising, the term 'the CNN effect' was coined to describe how media influence foreign policy by evoking audience responses through concentrated, emotionally based coverage, which pressurises governments to respond (Livingston, 1997, June). Almost two decades later, Papacharissi's (2015) analysis of events on Twitter (such as the 2011 Arab Spring and Occupy) conceptualises an 'affective public' where people use social media platforms to facilitate engagement, shape solidarity and make their voices matter in everyday politics.

These affective affordances play out differently across diverse geopolitical contexts. In a context of authoritarian silence in Tunisia and Egypt, Sumiala and Korpiola (2017) explain the construction of digital solidarities in the circulation and remediation of martyr narratives of the suicide protest of Tunisian fruit seller, Mohammed Bouazizi and the death of a young Egyptian man, Khaled Saeed, after being beaten by police. In a geopolitical context of strong, complex institutions for Internet censorship, Song et al.'s (2016) study of emotional expression on political aspects

of food safety issues on China's microblogging site, Weibo, notes the benefit of forming like-minded clusters around emotions expressed. It concludes that such activity can convey political opinions that resonate, helping to hold the state accountable, which is beneficial in an authoritarian society that values rational social engineering to efficiently achieve order (also see Tong, 2015). Given the importance of feelings in citizen-political communications worldwide, it is unsurprising that scholarship has turned to addressing how governments can best try to influence and manage public feeling within the *civic body*.

Managing Public Feeling: Discourses, Decision-Making and Datafication

We highlight three mechanisms that power-holders use in efforts to manage public feeling and hence behaviour: one is discursive, one is decision-making based and one is datafied.

Managing Discourses

The management of public feeling can be attempted through carefully constructed discourses in public communications. Wahl-Jorgensen (2019) sees the discursive construction of emotion through media texts as carefully staged strategic performances for specific purposes and audiences, driving social and political action. She argues that societies have always been preoccupied with managing emotions, with eras characterised by distinctive 'emotional regimes', namely, normative emotions and ways of expressing them in public. As already noted, early twentieth-century American mass communication scholars explored how mass media could be combined with psychological research to manage the population (Bernays, 1928a; Lasswell, 1936). A century later, political rulers continue to attempt emotional influence of their populations through media. For instance, in China, the 14th Five-Year Plan for National Informatisation issued by the Central Commission for Cybersecurity and Informatization (2021, December 28) states its aim of expanding 'diversified online propaganda platforms and channels' and strengthening 'the propagation of positive energy information' in cyberspace.

Likewise, in political campaigning, the discursive manipulation of the electorate's feelings has long been attempted. During India's 2019 General

Election, the ruling Hindu-nationalist party, the BJP, used humour, wit and sarcasm in its digital campaigns, helping entrench conversations about Hindu nationalism, stretching the boundaries of what could be said in public and creating familiarity with nationalist vocabulary while stirring Hindu majority fears against the 14% Muslim minority (Naumann et al., 2019). Such discursive manipulation during elections has been extensively studied in the USA. A longitudinal study of television advertising in US presidential campaigns across the second half of the twentieth century (1952–1996) finds that emotional appeals are more often dominant than logical or ethical ones (Kaid & Johnston, 2001). Negative campaigning (namely, attacking an opponent) has also long featured in American political campaigning (evident since at least 1800) as well as in other democracies (Fowler et al., 2016) and continues on social media (Haselmayer, 2019). A meta-analysis of 111 studies (mostly from a pre-social media ecology) finds that negative campaigning is more memorable and stimulates knowledge about the campaign yet, less positively, also slightly lowers feelings of political efficacy, trust in government and possibly overall public mood (Lau et al., 2007). Worryingly, experimental work (on American college students) in cognitive psychology shows that negatively valenced false political information tends to be more durable than positive or neutral false information, even after the false information is corrected (Guillory & Geraci, 2016).

Managing Decision-Making

A second mechanism for trying to manage public feeling focuses on people's decision-making processes. The twenty-first century has seen a surge in research suggesting that emotions guide formation of opinions and decisions to take political action (Brader & Wayne, 2016). For instance, analysis of survey data from the 1996 US presidential election shows that voters' opinions of candidates eventually converge with their initial emotional responses (Just et al., 2013 [2007]). American research on campaign ads evidences the importance of enthusiasm and fear in increasing desire to volunteer and vote (Brader, 2006). When people search for information online, campaign-related and experimentally induced fear consistently makes voters more inclined to pay attention to candidates and debates (Valentino et al., 2008). Incivility and negative political speech can lead to higher participation and stimulate voter turnout (Lu & Myrick, 2016).

The importance of emotions and gut feeling in decision-making more generally has been studied both by behavioural economics and cognitive psychology. It has led governments and corporations to design interventions or 'nudges' to help us make better (or different) decisions (The Behavioural Insights Team, 2020). Nudges are 'any aspect of the choice architecture that alters people's behaviour in a predictable way, without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid' (Thaler & Sunstein, 2008, p. 6). Nudges inform people of factual information (such as via warnings, reminders, personalisation, framing, timing and increases in salience); make certain choices easier (via simplification, ease, convenience and active choosing); use the power of default and procrastination (such as default rules on opting in or out); or exploit social influences (for instance, being told what other people do and leveraging social norms) (Sunstein, 2016).

Operationalising nudges, the first Behavioural Insight Team institutionalised in government was in the UK, created in 2010 (Sunstein, 2016). By 2020 its work spanned 31 countries, applying behavioural insights to inform policy and improve public services (The Behavioural Insights Team, 2020). Global digital platforms also embrace nudging to engage and track their users. For instance, Facebook's get-out-the-vote button nudges users to vote. This message displayed in users' Facebook News Feed on election day encourages voting, provides a link to local polling places, shows a clickable 'I Voted' button, and a counter indicating how many other Facebook users reported voting. The button was first used in the USA in 2008 and has since been used in elections and referenda in multiple countries. Facebook's own studies show that this slightly increases voter turnout (Bond et al., 2012; Jones et al., 2017). Yet, even this ostensibly positive nudge has raised concerns about uneven exposure across different parts of the population (as not everyone is on Facebook and not everyone is shown the button). It has also raised concerns about interfering in foreign elections (as Facebook, in all countries except the USA, is a foreign power and legally should not be interfering in their elections) (Grassegger, 2018, April 15). Also problematic are 'dark patterns', namely, design choices that alter users' decision-making for the designer's benefit. For instance, across 2022 Google was sued in the USA over 'deceptive' location tracking policies that make it hard for people to understand (and hence control) when or why Google collects and retains their location. The legal action refers to 'dark patterns' that include complicated

navigation menus, visual misdirection, confusing wording and repeated nudging towards a particular outcome (Wakefield, 2022, January 26).

Managing Datafication (Optimisation)

A third mechanism for attempting to manage public feeling involves gauging citizens' emotions through their datafied behaviour, gleaned via big data and profiling technologies. Across the twenty-first century, big data has been combined with psychological science to optimise and target individual desires and vulnerabilities, whether for political campaigning or governing.

These profiling and targeting technologies are utilised by political campaigners, as will be elaborated in Chap. 6. Several studies evidence the nefarious practices of digital targeting of misleading messages designed to bypass thoughtful deliberation in favour of emotionalised engagement and culture wars. For instance, Kim et al.'s (2018) study of social media ads run by anonymous groups in the 2016 US presidential election demonstrates that they largely focused on divisive issues and that lower-income, White voters in swing states were most likely to be targeted, especially by ads on immigration and race. They found ads run by these groups to be largely misleading, emphasising negative emotions and political attacks. Attempted emotional manipulation of targeted electorates via social media has also been observed during the Catalan referendum for independence from Spain on 1 October 2017. Analysis of nearly four million Twitter posts collected during the referendum finds two polarised groups of Independentists and Constitutionalists. Bots targeted the most influential humans of both groups. They also bombarded Independentists with violent contents, increasing their exposure to negative, inflammatory narratives (for instance, that inspire fight, violence and shame against government and police), so exacerbating social conflict online (Stella et al., 2018).

Beyond political campaigning, some advocate for real-time social media surveillance to enable timely assessment of the public's emotional and behavioural responses to governments' actions. This includes measures to engage the public during crises such as riots and natural disasters (for instance, in the UK, USA and Indonesia), through to full incorporation into epidemic preparedness and response systems for public health communication and control in more centralised countries like China (Chen et al., 2020; Ni et al., 2020). Clearly, feelings deserve serious attention

when considering citizen-political communications. This is especially so when dissecting contemporary false information and wider questions of post-truth.

Post-truth? Assessing Emotionalised Media

According to the dictionary definition of post-truth environments, appeals to emotion and personal belief are argued to be more influential in shaping public opinion than objective facts. Scholars point to multiple practices indicative of a post-truth era (Balaskas & Rito, 2021; Blackman, 2022; Capilla, 2021; Farkas & Schou, 2020), many of them discussed in our book. Empirically we cannot speak to whether emotion is more, or less, important than facts currently. However, we can point to the increasing emotionality of media environments from which citizens draw their facts. Below, we examine the emotionality of two such media forms: news media and social media.

The Emotionality of News

Despite journalism's long-standing ideals of objectivity and privileging of facts over values, emotionality has always been part of the profession (Beckett & Deuze 2016; Peters, 2011), although its extent varies across different types of news genre and outlet, geography, platform and time (Pantti & Wahl-Jorgensen, 2021; Wahl-Jorgensen, 2013). For instance, a century ago, Lippman (1922) (himself a former US journalist) regarded news stories as lacking 'truth' because they are dominated by the emotions and hopes of those working in the news organisation. A hundred years later, Glück's (2021) interviews with Indian and British broadcast journalists within commercial networks and public service broadcasters find that they consider emotionalising elements as indispensable to engaging audiences: Indian producers appear particularly open to interventionist (rather than detached) roles, combining ideas of national development with motivating citizens and government-critical journalistic elements.

Other scholars observe that emotionality is an *increasing* feature of journalism. Pantti and Wahl-Jorgensen's (2011) study of British press coverage (1952–1999) of human-made disasters finds that from the 1980s onwards, there is a shift to a more open emotional regime, which values journalists' individualised emotional expressions, allowing them to raise

structural questions of collective significance. Coward (2013) observes that 'objective' reporters often want to be known for a distinctive personal voice in the 'confessional society'. Indeed, digital native news outlets BuzzFeed News (UK) and Vice News (UK) used subjective, confessional and personalised forms of expression to engage young audiences when reporting the 2017 UK General Election (Dennis & Sampaio-Dias, 2021). In the USA, Benkler (2020) observes the changing political-economic and commercial imperatives fuelling 'outrage' discourse in right-wing media, going back to the repeal by the Federal Communications Commission in 1987 of the fairness doctrine (which had required broadcasters to offer public affairs programming and a balance of viewpoints).

Others blame the digital media ecology itself for the rise in emotionality in news (Al-Rawi, 2020; Beckett & Deuze, 2016; Peters, 2011). For instance, Al-Rawi's (2020) concept of 'networked emotional news' comprises news stories posted on social media that generate quantifiable collective emotional responses due to audiences' strong involvement facilitated by Facebook's Reactions features. Beckett and Deuze (2016) observe that mobile digital media are increasingly personalised and intimate (as in always-on smart devices where personal and public networks interconnect) and that journalism turns to emotion to virally engage news consumers in an increasingly economically competitive news ecology. Not all types of emotion are viral, and there are national differences. An analysis of 9.6 million comments on the *New York Times* website between 2007 and 2013 finds that comments featuring partisan incivility receive the most engagement, but comments with swearing do not drive engagement (Muddiman & Stroud, 2017). However, a study of user comments from 26 news websites in South Korea in 2012 finds that swearing increases interaction with comments, especially for political discussions (Kwon & Cho, 2017).

Expectations of emotionality in news also play out differently in different countries and different demographics. For instance, whereas in the USA only 29% of people surveyed in 2022 think that journalists should be able to express personal opinions as well as reporting news, this figure is far higher in Japan (44%) and Brazil (60%); and across all countries, younger adults are more prone to this view (Newman et al., 2022). The affective nature of news consumption is further evident in a study of 56 Dutch users' news browsing which finds that affective considerations influence their clicking patterns (Kormelink & Meijer, 2017, p. 678).

That audiences want a news experience that accords with their worldview is found in a qualitative study of far-right citizens in Norway. These citizens perceive that mainstream press do not cover the perceived threats of immigration and Islam objectively and are angry that far-right political actors are silenced and ridiculed in the news. They seek alternative news sources that support their worldview (Ihlebæk & Holter, 2021). There is also a growth worldwide in selective news avoidance. Although across 46 countries surveyed in 2022, the most commonly cited reason for news avoidance is the repetitiveness of the news agenda (43%), and other reasons include emotionality, namely, that the news brings down their mood (36%), feeling worn out by the news (29%) and that the news leads to arguments they would rather avoid (17%) (Newman et al., 2022). Conversely, the role of news on social media platforms in eliciting positive emotions is shown by Al-Rawi's (2020) study in 2016 of over 12,000 news items on Facebook pages of mainstream American and British news outlets. This finds that social media readers are emotionally engaged with news that involves positive feelings (especially love). Another study of American mainstream news on Facebook shows how different ways of framing protests influences emoji engagement: for instance, if protests are framed as legitimate, this decreases emotional reactions from audiences (Kilgo & Harlow, 2021).

In terms of studying audiences' emotional relationships with fake news, there are few studies, and these are focused on the USA. Martel et al.'s (2020) experiments into the relationship between experiencing 20 specific emotions and believing fake news find that heightened emotionality at the study's outset predicts greater belief in fake (but not real) news posts. In other words, there are notable increases in belief in fake news as emotionality increases. Their study also finds correlational and causal evidence that audience's reliance on emotion increases their belief in fake news. Another US-based study finds that participants are more likely to believe fake news political headlines that align with their existing beliefs (for instance, liberals are more likely to believe negative news about conservatives); react with more negative emotions to such headlines that attack their party; and are more likely to report intentions to suppress fake news that attacks their own party. Furthermore, participants who reported high levels of emotions are more likely to take actions that would spread or suppress the fake news; and participants who reported low levels of emotions are more likely to ignore or disengage from the spread of false news (Horner et al., 2021).

The Emotionality of Social Media

While social media can support rational, deliberative discourse (Jakob, 2020), more studies point to highly emotional content circulating on social media. This section examines areas of emotionality on social media closely associated with false information: incivility, hate speech and conspiracy theories. Notwithstanding everyday realities of how people receive, understand, negotiate and subsequently circulate content, we observe that many studies highlight the various affordances of social media platforms that are then exploited by architects of disinformation.

Internet scholarship has long noted uncivil behaviour online. This includes 'trolling', an antagonistic rhetorical practice that aims to elicit emotional responses from unwitting or unwilling targets (Phillips, 2015); online shaming through 'viral outrage' (Sawaoka & Monin, 2018); 'oppressive outrage' where marginalised voices are silenced with coordinated harassment (Brady & Crockett, 2018); and 'hate speech', commonly understood to be bias-motivated, hostile, malicious language targeted at people because of their actual or perceived innate characteristics (Sellars, 2016; Siegel, 2020, p. 57). The overall incidence of hate speech in social media appears to be rare, but there are few systematic studies (Siegel, 2020, p. 66). However, hate speech promotes reactions and travels further. This virality is evident in a study on Facebook surrounding Ethiopia's 2015 General Election (Gagliardone et al., 2016). Similarly, a big data study across 2016–2018 on Gab (a site created in 2016 as a free speech alternative to Twitter that mainly attracts alt-right users) finds that content generated by hateful users tends to spread faster, farther and reach a wider audience compared to content from non-hateful users (Mathew et al., 2019). Unsurprisingly given its virality, a cross-national survey of youths and young adults from Finland, Germany, the UK and the USA suggests that many have been incidentally exposed to online hate speech (53% of Americans, 48% of Finns, 39% of Britons and 31% of Germans), especially those who use online social networks often and visit 'dangerous' sites (Hawdon et al., 2017).

Multiple reasons are posited for the rise of online incivility and hate speech, many pointing to the affordances of social media platforms. While their design would not have intended such anti-social behaviour, it is a regular outcome. A notorious example is 4Chan's affordances of anonymity and ephemerality that enable what Tuters and Hagen (2020) call 'memetic antagonism', namely, the use of memes as vehicles for

antagonistically articulating an out-group, unbound by civility. On more mainstream platforms such as Twitter, Ott (2017) suggests that incivility is due to its informality and depersonalisation of interactions with others. Others discuss the emotional disinhibition and lack of social control prevalent online as, for various reasons, including anonymity, participants feel free from social convention (Suler, 2016). Crockett (2017) suggests that digital media may exacerbate expression of moral outrage and viral online shaming in three ways. Firstly, digital media inflate its triggering stimuli (as people are more likely to learn about immoral acts online than in person, as online algorithms promote content most likely to be shared and as people are more likely to share content that elicits moral emotions). Secondly, digital media reduce the costs of online shaming (the tools for quickly expressing outrage online are at our fingertips while hiding the target's suffering). Thirdly, digital media amplify personal benefits from online shaming (such as virtue signalling moral authority to large audiences).

Conspiracy theories also proliferate on social media (Bessi et al., 2015; Zollo et al., 2017). Conspiracy beliefs are attempts to explain the ultimate causes of significant social and political events and circumstances with claims of secret plots by two or more powerful, malevolent actors. Such beliefs are widespread and long-standing in modern Western societies (Allcott & Gentzkow, 2017; Sutton & Douglas, 2020). The affordances of social media are not a dominant explanation for conspiracy theory proliferation. Rather, explanations point to complex psychological, political and social factors, this demonstrated by a review of studies from psychology, political science, sociology, history, information sciences and the humanities (Douglas et al., 2019). For instance, US experiments and surveys show an association of anxiety and personal uncertainty with conspiracy perceptions: as such, increasing anxiety or personal uncertainty levels (potentially induced by disinformation architects) may lead ordinary people (not just the paranoid) to become conspiracy theorists (Radnitz & Underwood, 2017; Miller, 2020). Conspiracy theories appear to provide broad, internally consistent explanations that help people to preserve beliefs in the face of uncertainty and contradiction, helping them see the world as orderly, understandable and predictable following threatening societal events (van Prooijen & Jostmann, 2013). There is also a relationship between conspiracy belief and distrust in governments, authorities and scientists (Jensen et al., 2021; Lindholt et al., 2021; Sutton & Douglas, 2020). Across these studies, the direction of causality remains unclear. More clear-cut, however, is a demographically representative

survey of Americans in 2020 that finds that women are significantly less likely than men to endorse COVID-19 conspiracy theories and that this cuts across political party lines (Cassese et al. 2020).

That conspiracy theories proliferate on social media, then, is not reducible to social media affordances. Yet, these affordances certainly have some bearing. Given that social media are designed to maximise user attention and affect (see Chap. 2), it is unsurprising that conspiracy theories proliferate there. Indeed, a US national online survey (760 adults) into conspiracy beliefs finds that those with heavy reliance on, and trust in, social media news have the highest level of general and COVID-19-related conspiracy beliefs. Furthermore, those who blindly trust social media news are more likely to fall prey to conspiracy theories even if they can identify the false information (Xiao et al., 2021). A UK-wide national survey (May 2020) of over 16-year-olds finds that those who believe in COVID-19 conspiracy theories are far more likely than non-believers to get their information about the virus from social media (Duffy & Allington, 2020). Demography, nation and its media ecology clearly make a difference as an online survey of adults in China, where the information environment is strictly controlled and rumours are banned on social media, shows that social media use was not associated with conspiracy theory endorsement (Su et al., 2021; also see Jensen et al. 2021). An experimental study into how 50 German university students emotionally cope when confronted with an opinion-challenging YouTube clip propagating conspiracy theory disinformation about causes of climate change finds highly varied coping strategies. Of concern is that, for many participants, their climate change problem awareness decreased following exposure to the conspiracy clip (Taddicken & Wolff, 2020).

Indeed, the importance of social media's affordances in proliferating emotional content more generally is indicated in studies that find that expression of emotion is *socially contagious* on social media (meaning that a perceiver's emotions become more similar to others' emotions as a result of exposure to these emotions), with caveats that such causality is difficult to prove (Goldenberg & Gross, 2020; McStay, 2018). Facebook's infamous mood study conducted in 2012 secretly optimised 689,003 people's News Feeds to understand 'emotional contagion' on its platform. (This is the only published study that has manipulated users' emotions without their knowledge on a digital media platform.) When users logged into their Facebook pages, some were shown News Feed content with a greater number of positive words, while others were shown sadder than average

content. After the week of exposure to either more positive or negative content, manipulated users were more likely to post either especially positive or negative status messages. When the experimenters reduced the positive *and* negative content (making News Feeds lacklustre), people reduced the overall amount they posted (Kramer et al., 2014). Certainly, Del Vicario et al.'s (2016) computational, comparative study of Italian Facebook pages' reporting on two polarised communities (scientific and conspiracy) across 2010–2012 shows that in both communities, emotional behaviour (ascertained by sentiment analysis of users' posts) is affected by how often users post comments. More posting of comments resolves in a more negative emotional state; and on average, more active users show a faster shift towards negativity than less active ones. This emotional contagion is also found in studies on Twitter (Brady et al., 2017; Ferrara & Yang, 2015; Goldenberg & Gross, 2020; Stieglitz & Dang-Xuan, 2013). For instance, Brady et al. (2017) find that presence of moral-emotional words in 563,312 tweets on three polarising issues increased their transmission by approximately 20% per word.

Emotional contagion is also found on non-US-based social media platforms. A big data analysis of the discussion network on Chinese microblogging site, Weibo, regarding political aspects of food safety (43,575 posts, June–August 2014) finds that compared with non-emotional posts, emotional posts are more likely to be spread through reposting and that political discussions expressing anger are most likely to generate responses (Song et al., 2016). Another big data study of Weibo in 2017 unpicks the massive-scale network of emotion contagion underpinning the anger of online activism. It finds that this is driven by broadcasters (presenting emotionally neutral posts, but signalling that the Chinese authorities are open to public discussion of the topic); celebrities (whose emotional venting acts as 'emotion initiators', provoking emotion contagion); and micro-celebrities (who act as 'emotion brokers' by connecting diverse subgroups) (Liu & Liu, 2021). Such emotional contagion is not an accident but the result of social media algorithms that are constantly tweaked to optimise engagement.

In short, intense emotions (including incivility, outrage and hate speech) and conspiracies proliferate online at least partly because of the affordances of social media platforms, which as Chap. 2 explained are geared towards eliciting high arousal and viral emotions to further their attention economy. This emotional virality is not just evident on Facebook (the mechanics of which have been revealed by whistleblowers, as

discussed in Chap. 2) but on other US and non-US-based social media platforms. Such affordances are exploitable by the varied architects of disinformation: partisanship can be stoked, and money can be made, from *civic bodies* undergoing strong conflicted emotions.

To summarise, the media from which people would normally derive their facts (namely, news media and social media) have become more emotionalised. Alongside the prominence of false information (see Chap. 4), is it any wonder that claims for a post-truth condition are prevalent? We cannot assess the general accuracy of whether emotion and personal belief play a *greater* role than facts in shaping public opinion, but we do observe that some (US-based) studies show that people *do* prioritise emotion over fact in political arenas. We recommend more studies in different affective contexts across the world and demographically to empirically scrutinise the claim that we live in a post-truth condition. From what we have evidenced in this chapter, we can assert that we live in an informational environment that is sub-optimal for a healthy *civic body*. We exemplify this below by examining challenges faced by governments in managing their population's feelings during the COVID-19 pandemic where uncertainty, anxiety and disinformation prevail.

Affective Challenges in Managing COVID-19: Uncertainty, Anxiety and False Information

During the COVID-19 pandemic, inherently *uncertain* facts raised *anxiety* levels and provided fertile ground for false information worldwide. This made it harder for governments to manage their population's feelings to secure behaviour changes deemed necessary to combat this highly infectious respiratory disease.

COVID-19 ('coronavirus disease 2019') was first reported in Wuhan, China, on 31 December 2019. By March 2020, the virus had spread to over 120 countries, leading the World Health Organization to declare it a pandemic. Governments, to various degrees across the world, and with vastly different resources and states of preparedness, simultaneously mobilised their healthcare systems to cope with an influx of patients requiring prolonged intensive care; attempted to track and curtail the exponential spread of the disease; and instructed citizens to engage in profound and rapid behaviour change including wearing masks, washing hands and engaging in prolonged and repeated lockdowns. Messaging was often

mixed, and some governments played down health risks to keep public confidence in the economy. A year following the outbreak, there had been 2.75 million deaths globally, but several vaccines had been developed. Two years following the initial outbreak, the global death toll was over five million, and vaccine roll-out remained highly uneven worldwide, partly because of lack of supply but also because many refused to take the vaccine (Mallapaty et al., 2021). COVID-19 proved to be an inherently affective issue: alongside high death tolls, absence of cure; onsets of new, more transmissible and potentially vaccine-resistant variants; and extreme behaviour change required to quell the death spikes, increased anxiety and depression were reported across multiple countries (Sigurvinsdottir et al., 2020).

While false information in individual countries takes shape under specific affective contexts (as explored in Chap. 3), COVID-19 adds to this the sociological characteristics of being a 'risk issue' (Beck, 1992). Like other risk issues (such as climate change), it induces systematic, often irreversible harm (such as death and the debilitating condition of 'long COVID'). It also makes it hard for people to find trustworthy, reliable information because of three other core features that breed *uncertainty*. The first of these features is immateriality: risk issues generally remain invisible, giving them an air of unreality. Indeed, across 2020, the visibility of COVID-19 would only become apparent some 2–14 days from infection on manifestation of symptoms (such as loss of smell) or, as many infected people were asymptomatic, on reliable testing for the infection itself or for antibodies (World Health Organisation, 2020a, b, April 17). A second feature breeding uncertainty is reliance on causal interpretations: we only know how COVID-19 is likely to spread because experts have modelled this. For instance, the UK government consulted scientists, as part of its Scientific Advisory Group for Emergencies (SAGE) team, to model different interventions. Experts realised that COVID-19's reproduction rate, if left unchecked, is exponential, meaning that its effects on populations would largely be invisible in initial weeks, but would rapidly spike thereafter, overwhelming health services. A third feature that breeds uncertainty is that people must respond to the risk without an adequate foundation of knowledge. COVID-19 has no cure and, until almost a year after it was first identified, had no vaccine. Also, the virus mutates, producing new variants with resistance to some of the vaccines. As such, there was incomplete understanding of who is most at risk and what will best prevent it. As the pandemic progressed, some of these knowledge gaps closed,

but many remained. These features produce Beck's (1992) 'risk society'—a society that is uninsured and incapable of providing for the uncertainties it faces.

How do people react when living with such uncertainty? Uncertainty is strongly related to information seeking, especially with health information online (Lin et al., 2016). Certainly, COVID-19 saw a substantial increase in consumption for mainstream news media and online sources, evident in all six countries surveyed in 2020 before and after the pandemic took effect (Argentina, Germany, Spain, South Korea, the UK and the USA). Four months after the emergence of the disease, people considered the news media to have done a good job in helping them understand the crisis (60%) and in making clear what they can do to mitigate the impact (65%) (Newman et al., 2020). Of course, negotiating such crisis communication is not straightforward for journalists. In countries with more authoritarian tendencies such as Slovenia, where the governing Slovenian Democratic Party seeks to politically instrumentalise and economically devastate the media, scholars find journalists juggling their facilitative role (in helping the public to understand the health crisis, promoting the official discourse and pointing out false information) with a watchdog role critical of those in power (Pajnik & Hrženjak, 2022).

While in five countries surveyed in 2020 (the UK, Ireland, the USA, Spain and Mexico) public belief in false information about COVID-19 is rare, a substantial proportion views such false information as highly reliable. Furthermore, a small group finds common factual information about the virus highly unreliable (Roozenbeek et al., 2020). More commonly, public health scholarship demonstrates that when the public is exposed to novel or contradictory health information, people experience more uncertainty and disorientation, and decreasingly trust scientists issuing these competing recommendations (Chang, 2015; Clark et al., 2019). More broadly, trust in experts has long been in global decline: as far back as 2005, trust shifted from authorities to peers (Edelman, 2021). Analysis of vaccine misinformation across the twentieth and early twenty-first centuries in the USA, UK and Nigeria helps illuminate public reactions in the face of uncertainty. It finds that lack of trust in the science, government and money-hungry pharmaceuticals, alongside rare but heavily reported vaccine accidents, side effects or dangerous experimental tests, exacerbates the public's vulnerability towards conspiracy theories as they seek explanations (Cabrera-Lalinde, 2022). Even if there is quality, trusted information available, multiple psychological biases that help people reduce

uncertainty may prevent them from acting on public health messages. For instance, we may limit our exposure to conflicting information by defaulting to information channels we deem credible (the 'channel heuristic'), which may generate large variations in beliefs about what is true, especially when issues become politicised, as COVID-19 became in many countries including the USA and Brazil (Dunwoody, 2020; Gramacho et al., 2021, Hamilton & Safford, 2021).

Under such conditions of uncertainty, governments have had to formulate responses that maximise public safety, effectively use finite health services and minimise adverse economic impacts and disruption to people's lives, but also avoid undesirable population responses such as panic or indifference. In liberal democracies, scholars proffered advice on managing populations based on likely affective reactions to COVID-19. Petersen (2020, March 9) advocated that 'optimistic anxiety' (but not insecurity), and telling people the truth about the pandemic, would affect citizens' political behaviour and information seeking in positive ways (regarding compliance with government measures) while averting panic. Fear is a central emotional response during a pandemic. A meta-analysis reports that appealing to fear leads people to change their behaviour if they feel capable of dealing with the threat, but produces defensive reactions when feeling helpless. Furthermore, people often exhibit 'optimism bias': the belief that bad things are less likely to befall oneself than others. Behavioural and social scientists therefore recommended that COVID-19 communication strategies should strike a balance between breaking through optimism bias without inducing excessive anxiety and dread. The study also notes that an emerging sense of shared identity and concern for others arises from the shared experience of being in a disaster and that this feeling can be harnessed by urging 'us' to act for the common good (Bavel et al., 2020).

In the UK, Independent Scientific Pandemic Insights Group on Behaviours (SPI-B) provides independent, expert behavioural science advice to SAGE, which in turn advises government ministers and officials. By February 2021, they found low vaccine take-up in certain groups (such as those shielding, and from deprived socio-economic circumstances, and also from non-White groups). Accordingly, a core SPI-B recommendation in March 2021 was to develop communications from 'a more data-driven approach that moves beyond aggregated headline percentages and flags important disaggregated, nuanced sub-groups, confounders and intersectionality to more efficiently target low uptake and hesitancy in a more tailored manner' (SPI-B, 2021, March 9, p. 14). SPI-B also recommends

that the government should: 'continue messaging about positive effects of behavioural interventions such as face coverings, high vaccine uptake, low vaccine hesitancy, hope and return to longer goals and avoid blame or enforcement' (SPI-B, 2021, March 9, p. 2). Positivity, giving hope and encouragement, and avoiding blame, then, was the UK government's desired emotional regime for managing COVID-19 during 2021.

However, of particular concern to governments trying to change population behaviour is that when health messages are unclear, people are less likely to change behaviour (Chang, 2015; Taber et al., 2015). It was therefore of grave concern to governments and health organisations worldwide that as COVID-19 spread globally, so did emotive, false information. In China, conspiracies circulated that the virus was part of the American trade war, or a biological weapon, or brought into China by American military members. Conversely, in the USA, conspiracies abounded that the virus may have originated in a Chinese lab and was a Chinese bio-weapon (Su et al., 2021). Indeed, a demographically representative online survey of US adults in 2020 found that 52% believe the virus was accidentally released by China and 49% believe it is a Chinese biological weapon (Miller, 2020).

Harmful disinformation about COVID-19 went particularly viral in smaller media markets, where technology companies face lower incentives to take adequate countermeasures, according to a report from East Stratcom Task Force (an organisation set up in 2015 to increase public awareness, understanding of, and resistance to, Russia's disinformation) (EUvsDISINFO, 2020). Given Facebook's research showing a small number of posters and commenters were responsible for much anti-vaccine content, an internal memo from 2 April 2021 saw Facebook reducing the number of comments a person could make on posts from authoritative health sources from 300 to 13 per hour. However, a leaked Facebook memo shows that in the first few months of 2021, about 41% of comments on English-language vaccine-related posts risked discouraging vaccinations; and even authoritative sources of vaccine information were becoming 'cesspools of anti-vaccine comments' (Schechner et al., 2021, September 17).

There are varied actors and motivations behind COVID-19 disinformation. As well as anti-vaccine activists worldwide whose existence predates COVID-19 (Cabrera-Lalinde, 2022), such disinformation is spread by Russian and Chinese state media, aiming to undermine the European Union and its crisis response and to sow confusion about COVID-19's

origins and health implications (according to a report from East Stratcom Task Force (EUvsDISINFO, 2020) and from an American think tank, the Council on Foreign Relations (Kurlantzick, 2020)). By April 2021, a rumour-tracking program from US-based analytics company, Novetta, found that Russia targets African countries to discredit Western vaccines in favour of its own Sputnik V (Hotez, 2021). In an African context where COVID-19 has exposed poor health systems, governments' default response, in line with years of official practices, has been denial, secrecy and false information spread through state-controlled media (Ogola, 2020). Religious actors also spread emotive false COVID-19 information. In a Middle Eastern and North African context, Alimardani and Elswah (2020) find that Islamic misinformation and clickbait on social media became more acute during the pandemic. This took shape in false Hadiths (fabrications of retellings of the Prophet's words and deeds). For instance, a flood of Arabic-speaking YouTube videos prophesised that a divine sound that would take 70,000 souls and leave 70,000 deaf would be heard on the night of the 15th of Ramadan 2020, based on a false Hadith. After being viewed millions of times, their virality and fear led the official Egyptian religious entity, Al-Azhar, to pronounce the videos false. Religious misinformation draws on fear, emotional appeals, or the credibility of religious authority to persuade. It is harder to fact-check and requires a deeper knowledge of religion and its socio-political context to discern.

Such false information harms the *civic body* mentally and physically. A study from a technology company that monitors and disrupts violent extremism online finds that in terms of mental harms, COVID-19 created a spike in online hate speech against China and Jews, with racially linked incitements of violence, hate speech and a rebirth of old conspiracy theories on Twitter (Moonshot, 2020, April). The physical health of the *civic body* was harmed as anti-vaccination attitudes hardened: refusal to get vaccinated effects the individual (a greater risk of getting the disease severely and of resulting hospitalisation) and the community (greater virus transmission and strain on health resources). While vaccines were developed by December 2020, vaccine hesitancy remained a major hurdle across 27 countries surveyed in 2020. On average, only one in three would take the vaccine as soon as possible. Those with poor information hygiene were 11% less likely than those with good information hygiene to say that they would take the vaccine within a year (Edelman 2021; also see Roozenbeek et al., 2020). With a large cross-country survey (conducted across

September 2020 to February 2021) finding large variations in acceptance of an approved COVID-19 vaccine (ranging from 83% in Denmark to just 47% in France and Hungary), the study finds that lack of vaccine acceptance is associated with conspiratorial thinking (namely, that the government is hiding information about the virus and its cures), as well as with lack of trust in authorities and scientists and a lack of concern about COVID-19 (Lindholt et al., 2021).

Conclusion

Feelings have enduring importance in citizen-political communications, fuelling collective identities and solidarities, and helping form opinions and decisions to act. This chapter highlighted three mechanisms that are used in efforts to manage public feeling and hence behaviour: one is discursive, one is decision-making based (often involving 'nudges') and one is datafied (often involving social media platforms and optimisation). We see these mechanisms at play in attempts to gain power (via political campaigning) and in attempts to govern once in office (for instance, in pandemic mitigation).

Our examination of the emotionality of two key media forms from which citizens garner their facts (news and social media) finds these to be highly emotionalised environments and hence fertile grounds for post-truth. Emotionality has always been a part of journalism despite long-standing ideals of objectivity; and it appears to be an increasing feature in the digital ecology given the rise of confessional journalism; changing political-economic imperatives; the personalised, always-on nature of digital media; and use of emotion to virally engage news consumers. Audiences' expectations of emotions in news vary across countries and demographics. Audiences share news for emotional (as well as other) reasons, and studies show that some audiences want a news experience that accords with their worldview, avoiding news exposure that elicits negative feelings while emotionally engaging with news that involves positive feelings. Indeed, studies suggest that as far as fake news and American audiences are concerned, there are notable increases in belief in fake news as audience emotionality increases and that people are more likely to believe fake news political headlines that align with their existing beliefs. While social media can support rational discourse, more studies point to the often highly negative and positive emotions they circulate worldwide. This is enabled by the affordances of social media platforms and their optimisation of

emotions, these then exploited by the architects of disinformation to spread incivility, outrage, hate speech and conspiracy theories. In this media ecology, it is easy to see how and why objective facts *may* have become less influential in shaping public opinion than appeals to emotion and personal belief, but studies on this causal link are mostly lacking.

Our examination of the resulting harms to the *civic body* highlights the challenges it poses to governmental efforts to manage their population's feelings and behaviour during the COVID-19 pandemic where uncertainty, anxiety, false information and conspiracy theories proliferated, where the issues became politicised and where people lack trust in authorities and scientists. While the news in multiple countries is generally regarded as having helped people understand the crisis, the facilitative versus watchdog role of journalists in negotiating such crisis communication is not straightforward, especially in countries with authoritarian tendencies. Both mental harms (online hate speech) and physical harms (reduced vaccine uptake) were evident. We conclude that we live in an informational environment that is sub-optimal for a healthy *civic body*.

Of note for horizon scanners is that although social media platforms have developed and honed the practice of profiling and targeting individual desires and vulnerabilities, they are now being joined by more emergent forms of emotional AI that claim to read and react to emotions through text, voice, computer vision and biometric sensing (McStay, 2018). While the biometric part of this is not yet widespread enough to have had significant empirical impacts on false information, in our final chapter (Chap. 9), we reflect upon near-horizon futures and how emotional AI may further incubate false information. Before such future-gazing, however, we turn next to examine the role of profiling and targeting in incubating false information online.

References

Alimardani, M., & Elswah, M. (2020). Online temptations: COVID-19 and religious misinformation in the MENA Region. *Social Media and Society, 6*(3), 4–7. https://doi.org/10.1177/2056305120948251

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Al-Rawi, A. (2020). Networked emotional news on social media. *Journalism Practice, 14*(9), 1125–1141. https://doi.org/10.1080/17512786.2019.1685902

Aristotle. (1991). *On rhetoric: A theory of civic discourse* (G. A. Kennedy, Trans.). Oxford University Press

Balaskas, B., & Rito, C. (2021). *Fabricating publics: The dissemination of culture in the post-truth era*. Open Humanities Press. http://openhumanitiespress.org/books/download/Balaskas-Rito_2021_Fabricating-Publics.pdf

Barfar, A. (2019). Cognitive and affective responses to political disinformation in Facebook. *Computers in Human Behavior, 101*, 173–179. https://doi.org/10.1016/j.chb.2019.07.026

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Bavel, J. J. V., Baicker, K., Boggio, P. S., et al. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour, 4*, 460–471. https://doi.org/10.1038/s41562-020-0884-z

Beck, U. (1992). *Risk society: Towards a new modernity*. University of Munich.

Beckett, C., & Deuze, M. (2016). On the role of emotion in the future of journalism. *Social Media + Society, 2*(3), 1–6. https://doi.org/10.1177/2056305116662395

Benkler, Y. (2020). A political economy of the origins of asymmetric propaganda in American media. In W. L. Bennett & S. Livingston (Eds.), *The disinformation age: Politics, technology and disruptive communication in the information age* (pp. 43–66). Cambridge University Press. https://doi.org/10.1017/9781108914628

Bernays, E. (1928a). *Propaganda*. Horace Liveright.

Bernays, E. (1928b). *Crystallizing public opinion*. Boni and Liveright.

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PloS One, 10*(2), e0118093. https://doi.org/10.1371/journal.pone.0118093

Blackman, L. (2022). Future faking, post-truth and affective media. In J. Zylinska (Ed.), *The future of media* (pp. 59–80). Goldsmiths Press.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*, 295–298. https://doi.org/10.1038/nature11421

Bösel, B., & Wiemer, S. (2020). *Affective transformations: Politics-algorithms-media*. Meson Press.

Brader, T. (2006). *Campaigning for hearts and minds*. University of Chicago Press.

Brader, T., & Wayne, C. (2016). The emotional foundations of democratic citizenship. In A. J. Berinsky (Ed.), *New directions in public opinion*. Routledge.

Brady, W. J., & Crockett, M. J. (2018). Letter. How effective is online outrage? *Trends in Cognitive Science, 23*(2), 79–80. https://doi.org/10.1016/j.tics.2018.11.004

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Moral contagion in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Cabrera-Lalinde, I. (2022). *How misinformation affected the perception of vaccines in the 20th century based on the examples of the polio, pertussis and MMR vaccines.* Retrieved April 13, 2022, from https://royalsociety.org/-/media/policy/projects/online-information-environment/oie-vaccine-misinformation.pdf

Capilla, P. (2021). Post-truth as a mutation of epistemology in journalism. *Media and Communication, 9*(1), 313–322. https://doi.org/10.17645/mac.v9i1.3529

Cassese, E. C., Farhart, C. E., & Miller, J. M. (2020). Gender differences in COVID-19 conspiracy theory beliefs. *Politics & Gender, 16*(4), 1009–1018. https://doi.org/10.1017/S1743923X20000409

Central Commission for Cybersecurity and Informatization. (2021, December 28). *14th Five-Year Plan for National Informatization*. Retrieved 27 April 2022, from https://digichina.stanford.edu/work/translation-14th-five-year-plan-for-national-informatization-dec-2021/

Chang, C. (2015). Motivated processing: How people perceive news covering novel or contradictory health research findings. *Science Communication, 37*(5), 602–634. https://doi.org/10.1177/1075547015597914

Chen, Q., Min, C., Zhang, W., Wan, G., Ma, X., & Evans, R. (2020, September). Unpacking the black box: How to promote citizen engagement through government social media during the COVID-19 crisis. *Computers in Human Behavior, 110*, 106380. https://doi.org/10.1016/j.chb.2020.106380

Clark, D., Nagler, R. H., & Niederdeppe, J. (2019). Confusion and nutritional backlash from news media exposure to contradictory information about carbohydrates and dietary fats. *Public Health Nutrition, 22*, 3336–3348. https://doi.org/10.1017/S1368980019002866

Clough, P., & Halley, J. (2007). *The affective turn: Theorizing the social*. Duke University Press.

Coward, R. (2013). *Speaking personally: The rise of subjective and confessional journalism*. Palgrave Macmillan.

Cowen, A., Sauter, D., Tracy, J. L., & Keltner, D. (2019). Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest, 20*(1), 69–90. https://doi.org/10.1177/1529100619850176

Crockett, M. J. (2017). Moral outrage in the digital age. *Natural Human Behaviour, 1*, 769–771. https://doi.org/10.1038/s41562-017-0213-3

Damásio, A. (1994). *Descartes' error: Emotion, reason, and the human brain*. Putnam Publishing.

Damásio, A. (2010). *Self comes to mind: Constructing the conscious brain*. Vintage

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo chambers: emotional contagion and group polarization on Facebook. *Nature, 6*, 37825. https://doi.org/10.1038/srep37825

Deleuze, G. (1988). *Spinoza: Practical philosophy*. City Lights. (Original work published 1970).

Deleuze, G., & Guattari, F. (2011). *What is philosophy?* Verso. (Original work published 1994).

Dennis, J., & Sampaio-Dias, S. (2021). "Tell the story as you'd tell it to your friends in a pub": emotional storytelling in election reporting by BuzzFeed News and Vice News. *Journalism Studies, 22*(12), 1608–1626. https://doi.org/10.1080/1461670X.2021.1910541

Dixon, T. (2012). "Emotion": The history of a keyword in crisis. *Emotion Review, 4*(4), 338–344. https://doi.org/10.1177/1754073912445814

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Advances in Political Psychology, 40*(1). https://doi.org/10.1111/pops.12568

Döveling, K., von Scheve, C., & Konijn, E. A. (2011). Emotions and mass media: An interdisciplinary approach. In K. Döveling, C. von Scheve, & E. A. Konijn (Eds.), *The Routledge handbook of emotions and mass media* (pp. 1–12). Routledge.

Duffy, B., & Allington, D. (2020). *Covid conspiracies and confusions: The impact on compliance with the UK's lockdown rules and the link with social media use*. London: The Policy Institute, King's College. Retrieved April 13, 2022, from https://www.kcl.ac.uk/policy-institute/assets/covid-conspiracies-andconfusions.pdf

Dunwoody, S. (2020). Science journalism and pandemic uncertainty. *Media and Communication, 8*(2), 471–474. https://doi.org/10.17645/mac.v8i2.3224

Edelman. (2021). *Edelman Trust Barometer*. Retrieved April 13, 2022, from https://www.edelman.com/sites/g/files/aatuss191/files/2021-03/2021%20Edelman%20Trust%20Barometer.pdf

EUvsDISINFO. (2020). *EEAS special report update: Short assessment of narratives and disinformation around the COVID-19 pandemic (updated 2–22 April)*. Retrieved April 13, 2022, from https://euvsdisinfo.eu/eeas-special-report-update-2-22-april/

Farkas, J., & Schou, J. (2020). *Post-truth, fake news and democracy: Mapping the politics of falsehood*. Routledge.

Ferrara, E., & Yang, Z. (2015). Measuring emotional contagion in social Media. *PLoS One, 10*(11), e0142390. https://doi.org/10.1371/journal.pone.0142390

Fowler, E., Franz, F., Michael, M., & Ridout, T. N. (2016). *Political advertising in the United States*. Routledge, Taylor and Francis.

Gagliardone, I., Pohjonen, M., Beyene, Z., Zerai, A., Aynekulu, G., Bekalu, M., Bright, J., Moges, M. A., Seifu, M., Stremlau, N., Taflan, P., Gebrewolde, T. M., & Teferra, Z. (2016). Mechachal: Online debates and elections in Ethiopia – from hate speech to engagement in social media. Retrieved April 13, 2022, from https://ssrn.com/abstract=2831369 or http://dx.10.2139/ssrn.2831369

Glück, A. (2021). Replacing the public with customers: How emotions define today's broadcast journalism markets. A comparative study between television journalists in the UK and India. *Journalism Studies, 22*(12), 1682–1700. https://doi.org/10.1080/1461670X.2021.1977166

Goldenberg, A., & Gross, J. J. (2020). Digital emotion contagion. *Trends in Cognitive Sciences, xi*(2), 316–328. https://doi.org/10.1016/j.tics.2020.01.009

Gramacho, W., Turgeon, M., Kennedy, J., Stabile, M., & Santos Mundim, P. (2021). Political preferences, knowledge, and misinformation about COVID-19: The case of Brazil. *Frontiers in Political Science, 3*(36). https://doi.org/10.3389/fpos.2021.646430

Grassegger, H. (2018, April 15). Facebook says its 'voter button' is good for turnout. But should the tech giant be nudging us at all? *The Guardian*. https://www.theguardian.com/technology/2018/apr/15/facebook-says-it-voter-button-is-good-for-turn-but-should-the-tech-giant-be-nudging-us-at-all

Guillory, J. J., & Geraci, L. (2016). The persistence of erroneous information in memory: The effect of valence on the acceptance of corrected information. *Applied Cognitive Psychology, 30*(2), 282–288. https://doi.org/10.1002/acp.3183

Habermas, J. (1984). *The theory of communicative action, Volume 1: Reason and the rationalization of society* (T. McCarthy, Trans.). Beacon Press

Hamilton, L. C., & Safford, T. G. (2021). Elite cues and the rapid decline in trust in science agencies on COVID-19. *Sociological Perspectives, 64*(5), 988–1011. https://doi.org/10.1177/07311214211022391

Haselmayer, M. (2019). Negative campaigning and its consequences: A review and a look ahead. *French Politics, 17*, 355–372. https://doi.org/10.1057/s41253-019-00084-8

Hawdon, J., Oksanen, A., & Rasänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior, 38*(3), 254–266. https://doi.org/10.1080/01639625.2016.1196985

Heidegger, M. (2011). *Being and time*. Harper & Row. (Original work published 1962).

Horner, C. G., Galletta, D., Crawford, J., & Shirsat, A. (2021). Emotions: The unexplored fuel of fake news on social media. *Journal of Management Information Systems, 38*(4), 1039–1066. https://doi.org/10.1080/07421222.2021.1990610

Hotez, P. (2021). COVID vaccines: Time to confront anti-vax aggression. *Nature, 592*, 661. https://doi.org/10.1038/d41586-021-01084-x

Hume, D. (1978). *A treatise of human nature*. Oxford University Press. (Original work published 1739).

Ihlebæk, K. A., & Holter, C. R. (2021). Hostile emotions: An exploratory study of far-right online commenters and their emotional connection to traditional and alternative news media. *Journalism, 22*(5), 1207–1222. https://doi.org/10.1177/1464884920985726

Jakob, J. (2020). Supporting digital discourse? The deliberative function of links on Twitter. *New Media and Society, 24*(5). https://doi.org/10.1177/1461444820972388

Jensen, E. A., Pfleger, A., Herbig, L., Wagoner, B., Lorenz, L., & Watzlawik, M. (2021). What drives belief in vaccination conspiracy theories in Germany? *Frontiers in Communication, 6*, 678335. https://doi.org/10.3389/fcomm.2021.678335

Jones, J. J., Bond, R. M., Bakshy, E., Eckles, D., & Fowler, J. H. (2017). Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election. *PLoS One, 12*(4), e0173851. https://doi.org/10.1371/journal.pone.0173851

Just, M. R., Crigler, A. N., & Belt, T. L. (2013 [2007]). Don't give up hope: Emotions, candidate appraisals, and votes. In G. E. Marcus, W. R. Neuman, & M. MacKuen (Eds.), *The affect effect: Dynamics of emotion in political thinking and behavior*. Chicago Scholarship Online. https://doi.org/10.7208/9780226574431-010

Kaid, L. L., & Johnston, A. (2001). *Videostyle in presidential campaigns: Style and content of televised political advertising*. Praeger

Kant, I. (1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Trans.). Cambridge University Press. (Original work published 1781).

Kilgo, D. K., & Harlow, S. (2021). Hearts and Hahas of the public: Exploring how protest frames and sentiment influence emotional emoji engagement with Facebook news posts. *Journalism Studies, 22*(12), 1627–1647. https://doi.org/10.1080/1461670X.2021.1908840

Kim, Y. M., Hsu, J., Neiman, D., Kou, C., Bankston, L., Kim, S. Y., Heinrich, R., Baragwanath, R., & Raskutti, G. (2018). The stealth media? Groups and targets behind divisive issue campaigns on Facebook. *Political Communication, 35*(4), 515–541. https://doi.org/10.1080/10584609.2018.1476425

Konijn, E. A., & ten Holt, J. M. (2011). From noise to nucleus: Emotion as key construct in processing media messages. In K. Döveling, C. von Scheve, & E. A. Konijn (Eds.), *The Routledge handbook of emotions and mass media* (pp. 37–59). Routledge.

Kormelink, T. G., & Meijer, I. C. (2017). What clicks actually mean: Exploring digital news user practices. *Journalism, 19*(5), 668–683. https://doi.org/10.1177/1464884916688290

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(29), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Kurlantzick, J. (2020). *How China ramped up disinformation efforts during the pandemic*. Council on Foreign Relations. Retrieved April 13, 2022, from https://www.jstor.org/stable/resrep29835

Kwon, K. H., & Cho, D. (2017). Swearing effects on citizen-to-citizen commenting online: A large-scale exploration of political versus nonpolitical online news sites. *Social Science Computer Review, 35*, 84–102. https://doi.org/10.1177/0894439315602664

Lasswell, H. D. (1936). *Politics: Who gets what, when, how*. Whittlesey House.

Lau, R. R., Sigelman, L., & Rovner, I. B. (2007). The effects of negative political campaigns: A meta-analytic reassessment. *Journal of Politics, 69*(4), 1176–1209. https://doi.org/10.1111/j.1468-2508.2007.00618.x

Le Bon, G. (2016). *The crowd: A study of the popular mind*. CreateSpace Independent Publishing Platform. (Original work published 1896)

Lin, W.-Y., Zhang, X., Song, H., & Omori, K. (2016). Health information seeking in the Web 2.0 age: Trust in social media, uncertainty reduction, and self-disclosure. *Computers in Human Behavior, 56*, 289–294. https://doi.org/10.1016/j.chb.2015.11.055

Lindholt, M. F., Jørgensen, F., Bor, A., & Petersen, M. B. (2021). Public acceptance of COVID-19 vaccines: Cross- national evidence on levels and individual-level predictors using observational data. *BMJ Open, 11*, e048172. https://doi.org/10.1136/bmjopen-2020-048172

Lippman, W. (1922). *Public opinion*. Free Press.

Liu, N., & Liu, J. (2021). Leading with hearts and minds: Emotion contagion in China's online activism. *Social Movement Studies*. Advance online publication. https://doi.org/10.1080/14742837.2021.2011716

Livingston, S. (1997, June). *Clarifying the CNN effect: An examination of media effects according to type of military intervention*. Cambridge, MA: The Joan Shorenstein Center Research on the Press, Politics, and Public Policy, John F Kennedy School of Government, Harvard University. Retrieved April 13, 2022, from http://genocidewatch.info/images/1997ClarifyingtheCNNEffect-Livingston.pdf

Lu, Y., & Myrick, J. G. (2016). Cross-cutting exposure on Facebook and political participation: Unravelling the effects of emotional responses and online incivility. *Journal of Media Psychology: Theories, Methods, and Applications, 28*(3), 100–110. https://doi.org/10.1027/1864-1105/a000203

Lunt, P., & Pantti, M. (2007). The emotional public sphere: Social currents of feeling in popular culture. In R. Butch (Ed.), *Media and public spheres* (pp. 162–174). Palgrave.

Mallapaty, S., Callaway, E., Kozlov, M., Ledford, H., Pickrell, J., & Van Noorden, R. (2021, December 16). How COVID vaccines shaped 2021 in eight powerful charts. *Nature*. https://www.nature.com/articles/d41586-021-03686-x

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications, 5*(1), 1–20. https://doi.org/10.31234/osf.io/a2ydw

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on Web Science* (pp.173–182). New York: ACM. https://doi.org/10.1145/3292522.3326034

McStay, A. (2013). *Creativity and advertising: Affect, events and process*. Routledge.

McStay, A. (2018). *Emotional AI: The rise of empathic media*. Sage.

Miller, J. M. (2020). Psychological, political, and situational factors combine to boost COVID-19 conspiracy theory beliefs. *Canadian Journal of Political Science/Revue canadienne de science politique, 53*(2), 327–334. https://doi.org/10.1017/S000842392000058X

Moonshot. (2020, April). *From #CoronaVirusCoverUp to #NukeChina: An analysis of conspiracy theories, hate speech and incitements to violence across Twitter related to Covid-19*. Retrieved April 13, 2022, https://moonshotteam.com/resource/covid-19-conspiracy-theories-hate-speech-and-incitements-to-violence-on-twitter/

Mouffe, C. (2005). *On the political*. Routledge.

Mouffe, C. (2013). *Agonistics: Thinking the world politically*. Verso.

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication, 67*(4), 586–609. https://doi.org/10.1111/jcom.12312

Naumann, K., Sen, R., & Murali. V. S. (2019). *The impact of digital media on the 2019 Indian general election*. Institute of South Asian Studies. Retrieved April 13, 2022, from https://www.kas.de/documents/288143/4518801/ISAS-Special-Report-Impact-of-Digital-Media-Full.pdf/036704f7-9656-800d-2c7f-71e5c096b657?version=1.0&t=1576720504239

Newman, N., Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Reuters Institute digital news report 2020*. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K. (2022). *Reuters Institute digital news report 2022*. Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Ni, M. Y., Yang, L., Leung, C. M. C., Li, N., Yao, X. I., Wang, Y., Leung, G. M., Cowling, B. J., & Liao, Q. (2020). Mental health, risk factors, and social media use during the COVID-19 epidemic and cordon sanitaire among the community and health professionals in Wuhan, China: Cross-sectional survey. *JMIR Mental Health, 7*(5), e19009. https://doi.org/10.2196/19009

Ogola, G. (2020). Africa and the Covid-19 information framing crisis. *Media and Communication, 8*(2). https://doi.org/10.17645/mac.v8i2.3223

Ott, B. L. (2017). The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication, 34*, 59–68. https://doi.org/10.1080/15295036.2016.1266686

Pajnik, M., & Hrženjak, M. (2022). The intertwining of the Covid-19 pandemic with democracy backlash: Making sense of journalism in crisis. *Journalism Practice*. https://doi.org/10.1080/17512786.2022.2077806

Pantti, M. K., & Wahl-Jorgensen, K. (2011). 'Not an act of God': Anger and citizenship in press coverage of British man-made disasters. *Media, Culture & Society, 33*(1), 105–122. https://doi.org/10.1177/0163443710385503

Pantti, M. K., & Wahl-Jorgensen, K. (2021). Journalism and emotional work. *Journalism Studies, 22*(12), 1567–1573. https://doi.org/10.1080/1461670X.2021.1977168

Papacharissi, Z. (2015). *Affective publics: Sentiment, technology, and politics*. Oxford University Press.

Peters, C. (2011). Emotion aside or emotional side? Crafting an 'experience of involvement' in the news. *Journalism, 12*(3), 297–316. https://doi.org/10.1177/1464884910388224

Petersen, M. B. (2020, March 9). The unpleasant truth is the best protection against coronavirus. *Politiken*. https://pure.au.dk/portal/files/181464339/The_unpleasant_truth_is_the_best_protection_against_coronavirus_Michael_Bang_Petersen.pdf

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.

Quaranto, A., & Stanley, J. (2021). Propaganda. In J. Khoo & R. Sterken (Eds.), *The Routledge handbook of social and political philosophy of language* (pp. 125–146). Routledge.

Radnitz, S., & Underwood, P. (2017). Is belief in conspiracy theories pathological? A survey experiment on the cognitive roots of extreme suspicion. *British Journal of Political Science, 47*(1), 113–129. https://doi.org/10.1017/S000712341400055

Richards, B. (2007). *Emotional governance: Politics, media and terror*. Palgrave

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science, 7*, 201199. https://doi.org/10.1098/rsos.201199

Sawaoka, T., & Monin, B. (2018). The paradox of viral outrage. *Psychological Science, 29*(10), 1665–1678. https://doi.org/10.1177/0956797618780658

Scarantino, A. (2018). The philosophy of emotions and its impact on affective science. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (4th ed., pp. 3–48). Guildford Press.

Schechner, S., Horwitz, J., & Glazer, E. (2021, September 17). How Facebook hobbled Mark Zuckerberg's bid to get America vaccinated. *Wall Street Journal*. https://www.wsj.com/articles/facebook-mark-zuckerberg-vaccinated-11631880296

Sellars, A. F. (2016). *Defining hate speech*. Berkman Klein Center Research Publication No. 2016-20. Retrieved April 13, 2022, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244

Siegel, A. A. (2020). Online hate speech. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 56–88). Cambridge University Press.

Sigurvinsdottir, R., Thorisdottir, I. E., & Gylfason, H. F. (2020). The impact of COVID-19 on mental health: The role of locus on control and internet Use. *International Journal of Environmental Research and Public Health, 17*, 6985. https://doi.org/10.3390/ijerph17196985

Song, Y., Dai, X.-Y., & Wang, J. (2016). Not all emotions are created equal: Expressive behavior of the networked public on China's social media site. *Computers in Human Behavior, 60*, 525–533. https://doi.org/10.1016/j.chb.2016.02.086

Spezio, M. L., & Adolphs, R. (2013). Emotional processing and political judgment: Toward integrating political psychology and decision neuroscience. In G. E. Marcus, W. R. Neuman, & M. MacKuen (Eds.), *The affect effect: Dynamics of emotion in political thinking and behavior* (pp. 71–95). Chicago Scholarship Online. (Original work published 2007). https://doi.org/10.7208/chicago/9780226574431.003.0004

SPI-B. (2021, March 9). *Behavioural considerations for vaccine uptake in Phase 2 and beyond*. Scientific Advisory Group for Emergencies. Retrieved April 13, 2022, from https://www.gov.uk/government/publications/spi-b-behavioural-considerations-for-vaccine-uptake-in-phase-2-and-beyond-9-march-2021

Spinoza, B. (1996). *Ethics*. Penguin. (Original work published 1677).

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the*

National Academy of Sciences, 115(49), 12435–12440. www.pnas.org/cgi/doi/10.1073/pnas.1803470115

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems, 29*(4), 217–248. https://doi.org/10.2753/MIS0742-1222290408

Su, Y., Lee, D. K. L., Xiao, X., Li, W., & Shu, W. (2021, July). Who endorses conspiracy theories? A moderated mediation model of Chinese and international social media use, media skepticism, need for cognition, and COVID-19 conspiracy theory endorsement in China. *Computers in Human Behavior, 120*, 106760. https://doi.org/10.1016/j.chb.2021.106760

Suler, J. (2016). *Psychology of the digital age: Humans become electric*. Cambridge University Press.

Sumiala, J., & Korpiola, L. (2017). Mediated Muslim martyrdom: Rethinking digital solidarity in the 'Arab Spring'. *New Media & Society, 19*(1), 52–66. https://doi.org/10.1177/1461444816649918

Sunstein, C. R. (2016). *The Ethics of influence: Government in the age of behavioural science*. Cambridge University Press.

Sutton, R. M., & Douglas, K. M. (2020). Conspiracy theories and the conspiracy mindset: Implications for political ideology. *Current Opinion in Behavioral Sciences, 34*, 118–122. https://doi.org/10.1016/j.cobeha.2020.02.015

Taber, J. M., Klein, W. M. P., Ferrer, R. A., Han, P. K. J., Lewis, K. L., Biesecker, L. G., & Biesecker, B. B. (2015). Perceived ambiguity as a barrier to intentions to learn genome sequencing results. *Journal of Behavioral Medicine, 38*(5), 715–726. https://doi.org/10.1007/s10865-015-9642-5

Taddicken, M., & Wolff, L. (2020). 'Fake news' in science communication: Emotions and strategies of coping with dissonance online. *Media and Communication, 8*(1), 206–217. https://doi.org/10.17645/mac.v8i1.2495

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

The Behavioural Insights Team. (2020). *About us*. Retrieved April 13, 2022, from https://www.bi.team/about-us/

Tong, J. (2015). The formation of an agonistic public sphere: Emotions, the internet and news media in China. *China Information, 29*(3), 333–351. https://doi.org/10.1177/0920203X15602863

Tuters, M., & Hagen, S. (2020). (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society, 22*(12), 2218–2237. https://doi.org/10.1177/1461444819888746

Valentino, N., Hutchings, V., Banks, A., & Davis, A. (2008). Is a worried citizen a good citizen? Emotions, political information seeking, and learning via the internet. *Political Psychology, 29*(2), 247–273. https://www.jstor.org/stable/20447114

van Prooijen, J.-W., & Jostmann, N. B. (2013). Belief in conspiracy theories: The influence of uncertainty and perceived morality. *European Journal of Social Psychology, 43*(1), 109–115. https://doi.org/10.1002/ejsp.1922

Wahl-Jorgensen, K. (2013). The strategic ritual of emotionality: A case study of Pulitzer Prize-winning articles. *Journalism, 14*(1), 129–145. https://doi.org/10.1177/1464884912448918

Wahl-Jorgensen, K. (2019). *Emotions, media and politics*. Polity Press.

Wakefield, J. (2022, January 26). Google sued in US over 'deceptive' location tracking. *BBC News*. https://www.bbc.co.uk/news/technology-60126012

World Health Organisation. (2020a). *Coronavirus disease (COVID-19) pandemic*. Retrieved April 13, 2022, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019

World Health Organisation. (2020b, April 17). *Q&A on coronaviruses (COVID-19)*. Retrieved April 13, 2022, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses

Xiao, X., Borah, P., & Su, Y. (2021). The dangers of blind trust: Examining the interplay among social media news use, misinformation identification, and news trust on conspiracy beliefs. *Public Understanding of Science, 30*(8), 977–992. https://doi.org/10.1177/0963662521998025

Zollo, F., Bessi, A., Del Vicario, M., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., & Quattrociocchi, W. (2017). Debunking in a world of tribes. *PLoS One, 12*(7), e0181821. https://doi.org/10.1371/journal.pone.0181821

CHAPTER 6

Profiling, Targeting and the Increasing Optimisation of Emotional Life

Introduction

Having examined the nature of false information and understood the energising role of emotion and related states in its promulgation, in this chapter we examine profiling and targeting in citizen-political communications. Profiling and targeting are how emotion is understood, harnessed, amplified, dampened, manipulated and optimised (by platforms and would-be influencers). This chapter focuses on profiling and targeting in political campaigning as this is an intensively studied area awash with emotion and deception (as previous chapters demonstrate) and attracts uneven protections across the world (as we will show below). We examine the targeting and profiling technologies and practices in political campaigning in the USA, the UK and India, so highlighting the impact of different data protection regimes as well as uneven digital literacies. In exploring these issues, this chapter also outlines key tools and techniques utilised by digital political campaigners in the big data era to profile and target datafied emotions.

Profiling and Targeting in Citizen-Political Communications

Profiling and targeting have long been apparent in political campaigning. In one of the first detailed analyses of why Americans vote and arrive at their political attachments, Lazarsfeld et al. (1944, p. 15) describe the

V. Bakir, A. McStay, *Optimising Emotions, Incubating Falsehoods*,
https://doi.org/10.1007/978-3-031-13551-4_6

persuasive advantages that personal face-to-face communication has over mass communication (which, at that time, was radio and print in domestic settings).

> But suppose we do meet people who want to influence us and suppose they arouse our resistance. Then personal contact still has one great advantage compared with other media: the face-to-face contact can counter and dislodge such resistance, for it is much more flexible. The clever campaign worker, professional or amateur, can make use of a large number of cues to achieve his end. He can choose the occasion at which to speak to the other fellow. He can adapt his story to what he presumes to be the other's interests and his ability to understand. If he notices the other is bored, he can change the subject. If he sees that he has aroused resistance, he can retreat, giving the other the satisfaction of a victory, and come back to his point later. If in the course of the discussion he discovers some pet convictions, he can try to tie up his argument with them. He can spot the moments when the other is yielding, and so time his best punches. (Lazarsfeld et al., 1944, p. 15)

While writing in the 1940s, the personally tailored and optimised attributes that Lazarsfeld et al. (1944) ascribe to face-to-face communication are all seemingly achievable by today's digital profiling and targeting and at scale. This was the result of a century-long journey by advertisers, public relations experts and political campaigners to understand, and target, audiences with persuasive messages based on scientifically derived insights (Herbst, 2016; Hopkins, 1923; Wells, 1975). Even at the time that Lazarsfeld et al. (1944) were writing, a vast range of consumer feedback procedures had already been developed in the USA including testing of ads (1906), systematic collection of retail statistics (1910s), questionnaire surveys (1911), coded mailings (1912), audits of publishers' circulations (1914), specialised market research departments and house-to-house interviewing (1916), research text books (1919), saturation (1920), dry waste surveys (1926), a census of distribution (1929), sampling theory for large-scale surveys (c. 1930), field manuals (1931), retail sales indices (1933), national opinion surveys and audiometer monitoring of broadcast audiences (1935) (Beniger, 1986, pp. 378–80).

In terms of political marketing, as mass literacy and mass media rapidly expanded across the 1920s and 1930s in the USA, so did polling the public using more scientific methods (Herbst, 2016). Opinion polling allowed political parties to merge broad demographic data (statistically socio-economic in nature such as population, gender, race, age, income,

education and employment) with insights into how to craft messages that resonate with large parts of the population. This led to the development of targeted campaigning and direct mail in the USA in the late 1970s. By the twenty-first century, the rise of 'big data' and associated datamining techniques, tools and analytics enabled discovery of hidden patterns in seemingly unrelated data points and provided real-time, automated insights into massive, unstructured, diverse, unconventional datasets such as social media, transactional data and administrative data (Ceron et al., 2017). One common datamining technique is 'classification' that classifies items or variables in datasets into predefined groups using linear programming, statistics, decision trees and artificial neural networks. Another common datamining technique is 'clustering' that creates meaningful object clusters that share the same characteristics. Unlike classification that puts objects into predefined classes, clustering algorithms dynamically correlate seemingly unrelated data points into unnamed and undecipherable 'clusters'. These are then translated back into a limited number of describable categories that, in turn, are dependent on the values assigned to them by the people who buy and use them. These are unlikely to enable explainable algorithms, where people can understand why a certain insight has been reached about them from the data (Kotliar, 2020). Nonetheless, using such (and other) datamining techniques, political campaigning can now combine public voter files with commercial information from data brokers to develop detailed, comprehensive voter *profiles* (Bartlett et al., 2018, p. 27; Perloff, 2018, pp. 246–247) to enable microtargeting (Dobber et al., 2019).

Profiling is defined in the European Union General Data Protection Regulation (GDPR) as: 'any form of automated processing of personal data … to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements' (European Union General Data Protection Regulation, 2016, Recital 71). Profiling enables people to be targeted with honed, controlled messages to create adaptive ads; to provide location-based services; or to increase efficiency and personalisation of marketing messages in the 'Internet of Everything' (which brings together things, people, processes and data) (Petrescu et al., 2020). While a regular targeted message does not consider matters of audience heterogeneity, a *microtargeted* audience receives a message tailored to one or several specific characteristic(s) that are perceived by the advertiser as

instrumental in making the audience member susceptible to that message (Dobber et al., 2019).

Political marketing with such granular targeting is not inherently bad and could even service democracy. As noted by the UK's data regulator, it can better engage electorates and citizens on issues of particular importance to them (Information Commissioners Office, 2018, November 6, p. 18). Where conducted openly and honestly, it can manifest voters' desires, concerns and policy preferences to politicians thereby helping elected leaders develop programmes that meet voters' demands (Perloff, 2018, p. 250). However, critics point to more nefarious practices of profiling and microtargeting messages designed to bypass thoughtful deliberation in favour of emotionalised engagement and deception (as detailed in previous chapters). These are more difficult to guard against as political microtargeting is a form of political communication: as such, it is an exercise of the right to freedom of expression, which is guaranteed by Article 11 of the European Union Charter of Fundamental Rights and Article 10 of the European Convention on Human Rights (ECHR). Furthermore, such microtargeting practices can be highly innovative, as exemplified in 2018 when Dutch pro-immigrant party DENK microtargeted people who use a special sim card (one used mostly by immigrants to phone abroad), thereby efficiently reaching traditionally difficult to reach people. In order to scare its own base to vote, DENK experimented with fear appeals in the form of a false ad made to look like it came from the anti-immigration Party for Freedom, with the statement that after election day 'we are going to cleanse the Netherlands' (Dobber et al., 2019).

Unsurprisingly, data regulators have expressed concerns about voter profiling and microtargeting (Information Commissioners Office, 2018, November 6, 2020, November). Reflecting on the situation in the European Union, by December 2020, the European Commission warned:

> Existing safeguards to ensure transparency and parity of resources and airtime during election campaigns are not designed for the digital environment. Online campaign tools have added potency by combining personal data and artificial intelligence with psychological profiling and complex micro-targeting techniques. Some of these tools, such as the processing of personal data, are regulated by EU law. But others are currently framed mainly by corporate terms of service, and can also escape national or regional regulation by being deployed from outside the electoral jurisdiction. (European Commission, 2020, December 3, p. 2)

Such developments have generated concepts like the 'automated public sphere' (Andrejevic, 2020) and 'computational politics' (Chester & Montgomery, 2017). Care should be taken not to overstate the impact of these developments on voting behaviour, as the scholarly field examining the impact of political advertising is divided. For instance, there is a long tradition that finds 'minimal effects' of campaign interventions (Berelson et al., 1954; Dobber et al., 2020; Klapper, 1960). Reinforcing these long-standing findings, Kalla and Broockman's (2018) meta-analysis of field experiments shows that the effects of campaign contact and advertising (mainly via mail, phone calls and canvassing) on candidate choices of Americans in general elections are, on average, zero. However, this meta-analysis caveats that there is less evidence regarding online and television advertising, these also being areas of largest spend. It also concedes that issue-based persuasion remains possible when campaigns have resources to identify and target relevant issue cross-pressures. Furthermore, Jacobson's (2015) review of scholarship on US elections concludes that campaigns do influence voters. More recent studies also find that targeted, data-driven campaigns have some influence on American voters. For instance, in the 2012 US presidential campaign, Republicans influenced Democrats' voting behaviour when targeting them with issues where they and the Republican candidate shared common ground (such influence was minimal when targeting Democrats with incongruent issue messages or when targeting Republicans with either incongruent or congruent issue messages) (Endres, 2020). A field experiment study of a municipal election in Dallas, Texas, in 2017 finds that individually targeted banner ads generate a modest statistically significant increase in turnout among Millennial voters in competitive districts (Haenschen & Jennings, 2019).

What cannot yet be ascertained are the direct effects of continuously refined profiling and targeting techniques on unsuspecting populations' voting behaviour. It would be difficult to find a linear relationship between exposure to political microtargeting and political participation outcomes as it is difficult to separate out microtargeting inputs and outputs from other forms of campaign data and communication (Schäwel et al., 2021). Nonetheless, several studies are instructive. Dobber et al.'s (2020) experiment using a microtargeted deepfake on Dutch respondents finds that political microtargeting can amplify the effects of the deepfake, but for a much smaller portion of their sample than expected. Also of interest is a

study of the campaigning tactics of Jair Bolsonaro, a far-right, legislative backbencher to successfully become Brazil's president in 2018. This big data study of Twitter during the 2016 Rio de Janeiro municipal election concludes that Bolsonaro used that election to prepare his communications strategy for his successful, subsequent presidential campaign by testing potential targets and narratives, experimentally disseminating divisive narratives and microtargeting potential voters who shared a common range of diffused values, capturing anti-systemic tendencies and criticising corruption in financial, moral and religious terms (Santini et al., 2021).

Even in parts of the world lacking infrastructure for fixed-line Internet connections, much higher mobile penetration exposes connected populations to datamining, targeting and profiling during election campaigns. For instance, in the continent of Africa, fixed-line Internet connections are primarily an urban phenomenon and, in many African countries, lag the rest of the world. Yet Africa leads the world in daily time spent on social media (on average, 3 hours 10 minutes compared to the global average of 2 hours 27 minutes in 2022), largely driven by users in Nigeria, Ghana, South Africa, Egypt, Kenya and Morocco (Kemp, 2022). Profiling and targeting for electoral gain in Africa are concerning given that it also suffers from low digital literacy, extensive false information online and poor data privacy regimes. Many countries on the continent are weak democracies with largely unregulated political funding or are governed by autocracies with associated governmental digital surveillance of the political opposition, journalists and activists (Dzisah, 2020; Mare & Matsilele, 2020; Ndlela, 2020; Nothias, 2020). Indeed, an overview of electoral cybersecurity in Commonwealth countries (funded by the UK's Foreign and Commonwealth Office) concludes that the increase in highly targeted digital advertising, often using data obtained via insecure transmission and brokerage, could disrupt electoral campaigning (Brown et al., 2020, p. 28).

For better or for worse, global adoption of datamining, profiling and targeting technologies in political campaigning is accelerating worldwide. The following sections examine these developments in the USA (where many of the globally dominant social media platforms are headquartered), followed by the UK and India (democracies with different data protection regimes and occupying different places in the digital literacy spectrum). In doing so, we outline some key tools and techniques utilised by digital political campaigners in the big data era.

Profiling and Targeting in US Political Campaigning

Although the Fourth Amendment to the US Constitution [1791] upholds the right to privacy, the USA currently lacks any comprehensive privacy framework (unlike Europe). Only the states of California, Virginia and Colorado have comprehensive consumer privacy laws. Instead, privacy protections are embedded in sector-specific laws and regulations, such as the Health Insurance Portability and Accountability Act 1996 (for health-related data) and the Fair Credit Reporting Act 1970 (for credit-related data) (Fukuyama & Grotto, 2020, p. 200). This means that beyond several state and sectoral limitations, the government has largely left it to online companies to set their own privacy policies, which evolved into increasingly broad authorisations for the companies to extract data. The government can take action against the companies if they violate their own privacy policies and deceive consumers, but this does not guarantee institutional change (Starr, 2020). This absence of a comprehensive privacy framework helped spawn the profiling technologies and practices of the globally dominant US technology platforms, these then exploited in each election cycle to microtarget and mobilise voters. This takes place in a wider media context of low levels of trust in mainstream media, polarisation of mainstream media and a weakening of journalism (including local news deserts) created by digital platforms, leading the USA to be ranked only 42nd (out of 180 countries) on press freedom in 2022 (Reporters without Borders, 2022b).

Compared to traditional advertising companies that only track user browsing behaviours via opaque cookies, social media platforms access much richer data sources. For instance, they know users' personally identifiable information and often allow advertisers to target users based on this (Andreou et al., 2018). All Facebook users have some 200 'traits' attached to their profile. These include dimensions submitted by users or estimated by machine learning models, such as race, political and religious leaning, socio-economic class and education level (Hao, 2021). To reconcile conflicting goals of protecting the privacy of users' personal information but also profiting from microtargeted advertising, in 2007 Facebook implemented a targeted online advertising system that provides a layer between individual user data and advertisers. The advertising system collects from advertisers the ads they want to display and their targeting criteria and then delivers the ads to people fitting those criteria. Rather than 'selling' information about their users, the business model is to sell space

to advertisers, giving them access to people based on their demographics and interests (Facebook, 2007, November 6; Korolova, 2010). Why a user received a particular ad is therefore the result of a complex process depending upon many inputs including: what the platform thinks the user is interested in; characteristics of users the advertiser wants to reach; the set of advertisers and parameters of their campaigns; the bid prices of all advertisers; active users on the platform at a particular time; and the algorithm used to match ads to users (Andreou et al., 2018).

Given these legal and platform affordances, it is unsurprising that intensive datamining in political elections is well documented in the USA, with each election cycle adopting technological innovations to microtarget and mobilise voters (Stromer-Galley, 2014). Political parties and aligned political consultancies maintain political technologies (such as canvassing applications) and databases that candidates use, and electoral campaigns have many potential data sources (Kreiss, 2016). As well as their lists of donors and voter rolls (provided by local or state offices, typically containing each voter's party registration and electoral voting history), campaigns can rent lists from other candidates (Edelson et al., 2019). From the 1960s to 2004, campaigns targeted broad demographic groups (such as gender-based) by purchasing television spots (e.g. daytime spots for female voters) (Fowler et al., 2016). Big data and digital targeting in political campaigns was first utilised in a large way for the 2008 US presidential election (Barack Obama v. John McCain) to work out voter sentiments, target key market segments and design messages to mobilise voters in core electoral areas (Kreiss, 2016; Owen, 2014; Tufekci, 2014). Since 2012, digital platforms have advertised their wares to politicians to teach candidates how to use their platforms during elections to reach new voters using data such as demographics, behaviour, interest and attention measures that represent the public in new ways, and to facilitate digital advertising buys (Kreiss, 2016; Kreiss & McGregor, 2018). The amount spent on US digital political advertising increased significantly from $159 million in 2012 to $2847 million in 2020 (Statista, 2021). Edelson et al.'s (2019) analysis of over 1.3 million ads with political content from over 24,000 sponsors archived by Facebook, Twitter and Google in the USA (coinciding with the 2018 US midterm elections) finds that most political ads cost less than $100, confirming the prevalence of small, likely highly targeted, ads that can contain custom political messaging. They also find a significant amount of advertising by quasi for-profit media companies that appear to exist

solely to create deceptive, online astroturf communities to target different demographics and interests via paid and organic political messaging. These arise because regulations that require disclosure of the business that paid for the ad on broadcast stations or via direct mail do not apply to online advertising, largely because laws mandating such disclosures were drafted before these platforms were ubiquitous.

Across the past decade, then, a complex, opaque digital marketing ecosystem has emerged encompassing data brokers and data analytics companies alongside the usual professional persuaders. This enables the rise of influence activities in digital political campaigning. Targeting tools discussed below comprise those offered by social media platforms; those using social media platforms' affordances; and bespoke campaign mobile phone apps that bypass social media platforms. These are far from exhaustive.

Social Media Platforms: Targeting Tools

Social media platforms offer many forms of targeting, and these are utilised by political campaigns. For instance, 'A/B' testing is used by social media companies to rapidly model users' attention and behaviour to interactively nudge it. It compares two versions of a single variable, typically by testing a subject's response to variant A against variant B and determining which is more effective. An old technique, across the past decade, there has been an exponential increase in deployment of rapid A/B testing using AI. In the 2012 presidential election (Barack Obama v. Mitt Romney), Obama's digital team ran 500 A/B tests on their web pages (Formisimo, 2016). By the 2016 US presidential election (Donald Trump v. Hillary Clinton), Trump's digital team tested around 50,000–60,000 ad variations a day (Beckett, 2017, October 9). According to a report by Demos (a British cross-party, independent think tank), this utilised Facebook's tool, Dynamic Creative, to use predefined design features to construct thousands of ad variations, present them to users and find optimal combinations based on engagement metrics (Bartlett et al., 2018, p. 33).

A second important digital marketing tool is targeted advertising. Launched in 2012, Facebook's 'Custom Audiences' product enables marketers to upload their own data files (using personally identifiable information that they hold about their own customers, such as email addresses and

names) which can be matched to specific Facebook users (Andreou et al., 2018; Chester & Montgomery, 2017; Martínez, 2018, February 23). In January 2016, Facebook introduced the audience optimisation tool which allows marketers and advertisers to set preferences to target specific audiences based on a broad range of demographic data, but also interests, languages spoken, relationship status, work status, place of employment, 'ethnic affinity', life events, Facebook connections, tracked behaviours online, politics, likelihood to engage with political content and ideology (Kreiss & McGregor, 2019). Facebook has also allowed advertisers to use provocative targeting criterion, such as 'interested in "pseudoscience"', thereby grouping users by their vulnerabilities (Angwin, 2020, April 25). As Chap. 8 documents, it was not until 2022 that Facebook's parent company, Meta, took steps to prevent advertisers targeting people based on how interested Facebook thinks they are in 'sensitive' topics including political affiliation (Bond, 2021, November 9). Such targeted advertising has been used to try to dissuade target groups from voting. For instance, to dissuade people from voting for Hillary Clinton, the 2016 Trump campaign targeted families of immigrants from Haiti living in South Florida to remind them that her husband, former US president Bill Clinton, had failed to sufficiently aid Haiti as president and as head of a relief effort after a major earthquake in 2010 (Vaidhyanathan, 2018, p. 171).

A third important digital marketing tool is lookalike modelling. This uses big data analytics to acquire information about individuals without directly observing their behaviour or obtaining consent (Chester & Montgomery, 2017). Facebook offers various lookalike modelling tools through its 'Lookalike Audiences' ad platform which allows advertisers to reach new people on Facebook who are likely to be interested in their business, or political candidate, because they are similar to existing audiences (Bartlett et al., 2018, p. 10). Antonio García Martínez, the original product manager for Facebook's Custom Audiences, describes 'Lookalike Audiences' as 'the most unknown, poorly understood, and yet powerful weapon in the Facebook ads arsenal' (Martínez, 2018, February 23). Up until 2020, both Google and Twitter offered political or cause-based advertisers similar targeting criteria to Facebook, including custom audiences and lookalike audiences (Edelson et al., 2019; Hotham, 2021). More broadly, political digital marketing firms offer lookalike modelling to identify potential supporters and voters, by matching millions of voters to hundreds of data points to create detailed voter profiles.

Psychographic and Neuromarketing Tools

Political campaigners also use the affordances of social media platforms to deploy automated psychographic and neuromarketing tools. Psychographics, emotional testing and mood measurement have long been central to political campaigns (Jamieson, 1996) to understand voter values, attitudes, motivations, interests, opinions and lifestyles, but the rise of big data analysis and modelling enables access to psychological characteristics and political inferences beyond the reach of traditional databases (Bakir, 2020; Tufekci, 2014).

For instance, research by controversial psychologist and business academic, Michal Kosinski, finds that Facebook 'Likes' (a fraction of data available to data brokers) may accurately predict personal attributes, including political party affiliation and other highly sensitive personal attributes including religious and political views, sexual orientation, ethnicity, intelligence, happiness, use of addictive substances, parental separation, age and gender (Kosinski et al., 2013). It also claims to predict the 'Big Five' (also called OCEAN) personality traits (these traits have widespread acceptance among personality researchers): namely, Openness to experiences, Conscientiousness, Extroversion, Agreeableness and Neuroticism (Gosling et al., 2003). A meta-analysis of this young field also finds that digital footprints may be used to predict these 'Big Five' personality traits of social media users and that prediction accuracy for each trait is stronger when more than one type of digital footprint is analysed (Azucar et al., 2018, p. 157). Scholars disagree about the effectiveness of psychological targeting on Facebook. Some argue that it is so effective that its use should be regulated (Matz et al., 2017, 2018a, b), while others remain unconvinced (Eckles et al., 2018; Sharp et al., 2018).

Although scholarship disagrees about its effectiveness, some political marketing companies have been quick to deploy this tool. Indeed, it was research such as that by Kosinski et al. (2013) on use of Facebook 'Likes' to predict psychological characteristics and political inferences that attracted the attention of political data analytics and behaviour change company, Cambridge Analytica (Federal Trade Commission, 2019b, p. 3). Cambridge Analytica has since sent out mixed messages on whether it used this data for its psychographic profiling in the 2016 Trump presidential campaign (Bakir, 2020). Furthermore, the UK's data regulator, the Information Commissioner's Office (ICO), observes (after investigating Cambridge Analytica and parent company SCL) that the real-world

accuracy of its algorithmic predictions 'was likely much lower' than the company claimed (Denham, 2020, October 2, p. 17).

Whether or not psychographics was used, or was effective, privacy violations led to the collapse of Cambridge Analytica and its parent companies, SCL Elections and SCL Group. They went into administration in May 2018, after public allegations made by whistleblower Christopher Wylie that Cambridge Analytica had exploited the personal data of Facebook users (Wylie, 2018, p. 14). Following its collapse, in July 2019, as well as levying a record US$5 billion civil penalty against Facebook for failing to protect users' privacy, the US Federal Trade Commission filed an administrative complaint against Cambridge Analytica LLC (the US arm of the company) for deceptive harvesting of personal information from tens of millions of Facebook users for voter profiling and targeting. This personal information had been collected in 2014 from users of a Facebook app (the 'GSRApp' developed by Aleksandr Kogan). It had exploited Facebook's now notorious (and since 2015, ended) data portal ('Friends API') that enabled app developers to share not only users' data but that of users' friends. The information comprised users' Facebook User ID, which connects individuals to their Facebook profiles, as well as other personal information such as gender, birthdate, location and Facebook friends list (Federal Trade Commission, 2019a, July 24; Wylie, 2019, pp. 112–132). In April 2018, Facebook revealed that the maximum number of unique accounts that directly installed the GSRApp, as well as those whose data may have been shared with the app by their friends, comprised 70,632,350 in the USA, 1,175,870 in the Philippines, 1,096,666 in Indonesia, 1,079,031 in the UK, 789,880 in Mexico, 622,161 in Canada, 562,455 in India, 443,117 in Brazil, 427,446 in Vietnam and 311,127 in Australia (Schroepfer, 2018, April 4).

Describing how such data is put to work in political campaigns for deceptive and emotional manipulation, whistleblower Wylie (2019, p. 121) observes that in the USA, across summer 2014, Cambridge Analytica began developing fake pages on Facebook that looked like real forums, groups and news sources. When users joined these fake groups, Cambridge Analytica would post videos and articles to further provoke them. Cambridge Analytica now had users who self-identified as part of an extreme group and could be manipulated with data. The company did not target that many people as most elections are zero-sum games and it needed 'to infect only a narrow sliver of the population, and then it could watch the narrative spread' (Wylie, 2019, p. 122). Once a group reached

a certain number of members, Cambridge Analytica would set up physical events across the USA, where people could find a fellowship of anger and paranoia, allowing them to feel part of a broader movement and reinforce each other's conspiracies. Invitees were selected because of their traits, so Cambridge Analytica knew, generally, how they would react to one another. Once a county-based group started self-organising, they were introduced to a similar group in the next county, creating 'a statewide movement of neurotic, conspiratorial citizens. The alt-right' (Wylie, 2019, p. 123). Those targeted online with test ads had their social profiles matched to their voting records, so Cambridge Analytica knew their names and real-world identities. It then used numbers on the engagement rates of these ads to explore potential impact on voter turnout.

Campaign Mobile Phone Apps

Alongside social media platform targeting tools and psychographic and neuromarketing tools, a third and more recent type of targeting tool are bespoke campaign mobile phone apps. As a digital marketing tool, they assumed increased importance in the 2020 US presidential election (Joe Biden v. Donald Trump). By 2019, 81% of people in the USA were equipped with a smartphone, almost double the global average of 45% (Taylor & Silver, 2019, February 5). By the 2020 US presidential race, each campaign had a bespoke mobile phone app to target likely voters and to collect massive amounts of user data without needing to rely on social media platforms or expose themselves to fact-checker oversight of deceptive messaging.

Trump's app ('The Official Trump 2020 App'), developed by Phunware, offers carefully selected tweets and articles that reinforce the campaign's talking points, often propagating deceptive information without a named author and rarely citing sources beyond government press releases and tweets from Trump's supporters and White House staff. Like the Trump campaign app, Biden's 'Team Joe App' sends users notifications of upcoming campaign events or training sessions for digital activists. Unlike the Trump app, the Team Joe App was bult in-house (to protect users' privacy) and is largely built for a single purpose: relational organising where volunteers leverage their existing networks and relationships to support Biden. If app users share their contact list, this is cross-referenced with the Democratic Party's voter files; the system identifies people the app user may have a personal connection with who might be persuaded to vote for

Biden; and it prompts the app user to send these potentially undecided voters personalised messages (Gursky & Woolley, 2020).

Both apps ask users to provide the campaigns access to their phone contacts. The campaigns do not ask those contacts for permission for that information, and in the USA, they are not legally required to. Beyond users' friend's contacts, the Trump campaign app also seeks permission to access a far more extensive list of data to enable profiling and targeting, drawing comparisons to Cambridge Analytica (Gursky & Woolley, 2020). According to a former executive for Phunware, the data collected from Trump's app can be poured into an information ecosystem designed to replace the Facebook features that made the 2016 Cambridge Analytica scandal possible (Kates, 2020, July 18).

The USA, then, with its weak privacy laws and long history of electoral datamining, is a global leader in developing profiling and targeting tools and applying them to political campaigns. Most of its population do not like this situation. A poll (conducted by Knight Foundation-Gallup across 3–15 December 2019) finds that 72% of Americans say that Internet companies should make *no* information about its users available to political campaigns for targeting voters with online ads. Only 20% of US adults favour allowing campaigns access to limited, broad details about Internet users, such as their gender, age or postal code. This is in line with Google's policy, which, in 2019, reined in the scope of information that political campaigns could use for targeting. Only 7% of Americans say that any information should be made available for a campaign's use. This is in line with Facebook's targeting policies, which up until January 2022 did not put any such limits in place on ad targeting (although Facebook does give users some control over how many ads they see) (Bond, 2021, November 9; McCarthy, 2020, March 2).

Profiling and Targeting in UK Political Campaigning

Unlike the USA, the UK ranks fairly highly on press freedom, coming 24th out of 180 countries in 2022 (Reporters without Borders, 2022c) with a well-funded and regulated broadcasting sector and over 50% of the population trusting broadcast news, local news and regional news in 2022 (Newman, 2022). Furthermore, unlike the USA, the UK (as part of the European Union) was protected by comprehensive privacy legislation (the

European Union General Data Protection Regulation (GDPR) 2016)) and had much stronger data protection laws. Post-'Brexit', the UK GDPR came into effect on 1 January 2021, based on the European Union GDPR, with some changes to make it work more effectively in a British context. The GDPR offers data protections on consent (personal data cannot be processed without freely given, specific, informed and unambiguous consent, unless allowed by law); time limits on how long personal data can be kept; and profiling (the data subject has the right to not be subject to a decision based on automated processing, while profiling to analyse or predict behaviours or preferences is legally regulated) (European Union General Data Protection Regulation, 2016/679, Recital 71).

Consequently, compared to the USA, British political parties have far less access to types of data required to target voters. For instance, many American states have an electoral register that identifies voters by partisan preference, but the UK does not. Nonetheless, digital campaigning has sharply risen in the UK across the second decade of the twenty-first century. The proportion of money that British political campaigners reported spending on digital advertising as a percentage of their total advertising spend rose from 2% in 2014 to 42% in 2017 (The Electoral Commission, 2019a), not least because while paid political advertising in broadcasting is prohibited under the Communications Act 2003, the ban does not apply online (Dobber et al., 2019). Indeed, online political advertising in the UK has been characterised as a 'Wild West' due to its lack of transparency, deficiencies in monitoring by regulators and civil society, and lack of deterrence for election offences (All Party Parliamentary Group on Electoral Campaigning Transparency, 2020, January).

British digital campaigning has also seen increasing use of data analytics and data management approaches to profile and thereby identify target audiences, including 'persuadables' and swing voters. The extent of targeting appears to differ by party. For instance, in the 2015 UK General Election (won by the Conservatives), it was only the Conservatives who seem to have adopted the US model of individual-level targeting (labour used broader segment-based targeting). The Conservative Party targeted seats based on what the party knew about types of voters living there, their propensity to swing their vote, their reactions to certain messages and other seat-specific factors (Anstead, 2017). While partisans are unlikely to change their views based on ads, it only takes a small number of 'persuadables' to swing close elections. According to Dominic Cummings (campaign strategist for 'Vote Leave', the official campaign to leave the

European Union in the 'Brexit' Referendum), the Referendum result of 52% for Leave and 48% for Remain came down to only 'about 600,000 people' (Cummings, 2017, January 30). According to a report by the UK's data regulator, and a report funded by the UK Foreign and Commonwealth Office, Vote Leave heavily relied on data scientists, using data management services of Aggregate IQ (a Canadian digital advertising web and software development company). One of Aggregate IQ's roles was to accumulate data on individuals to build and apply predictive models, and to serve the most easily influenced individuals' heavily targeted messages (Brown et al., 2020, p. 46; Denham, 2020, October 2, p. 10). Cummings states that Vote Leave spent 98% of its budget on digital advertising (rather than mainstream media advertising), with most spent on ads that experiments had demonstrated were effective (Cummings, 2017, January 30). The core messages were highly emotive and deceptive, conveying that staying in the European Union would lead to swarms of Middle Eastern immigrants; that we could only 'take back control' by leaving; and that strained, cherished national resources like the National Health Service would be better financed if Britain left. Cummings estimates that Vote Leave ran around one billion targeted ads before the vote, mostly via Facebook, sending out multiple different versions of messages, testing them in interactive feedback loops (Cummings, 2016, October 29). Additionally, having identified from focus groups that crucial swing voters were confused, and liable to change their voting decision based on whether they had last seen a message from either side of the campaign, Vote Leave ensured that their ads were delivered to swing voters as late as possible in the campaign (Cummings, 2017, January 30, Howard, 2018, November 30).

This growing importance of data brokers (who collect and aggregate data) is noted with concern by the UK's data regulator, the Information Commissioners Office (2018, November 6). In 2019 the regulator conducted its first data protection audit of seven British political parties to assess compliance with data protection law. It finds that all parties typically obtained data from the full electoral register; the marked register (a copy of the electoral register that has a mark by the name of each elector who has voted); directly from individuals, usually by asking them, but also by collecting information electors place in the public domain about their political views; and publicly available data and other datasets such as census, election results, Land Registry and polling (Information Commissioners Office, 2020, November, p. 10). Additionally, the three main political

parties (Labour, Conservative and Liberal Democrats) obtained lifestyle-type information on individuals from data brokers under commercial agreements (Information Commissioners Office, 2020, November, p. 10). The audit finds that political parties analysed and profiled this data to derive further data, such as likelihood of individuals voting a certain way and their likelihood of turning out to vote. Parties then used their datasets and analysis to inform the purchase of ads on social media to target individual social media users; send out targeted emails or telephone canvassing voters to encourage individuals to vote or change their voting behaviour; and decide who to canvass on doorsteps. The Information Commissioner's Office concludes that 'there are systemic vulnerabilities in our democratic systems' (Denham, 2020, October 2, p. 1) and finds only a limited level of assurance that procedures are delivering necessary data protection compliance (Information Commissioners Office, 2020, November).

Despite this accelerated move towards profiling and microtargeting voters, there are few empirical studies on their practices in the UK, and findings are mixed regarding accuracy and prevalence. One study by Open Rights Group (a UK-based digital campaigning organisation working to protect people's rights to privacy and free speech online) suggests that the current state of political profiling does not seem particularly accurate (Crowe et al., 2020, June 23, p. 9). A study on Facebook ads during the 2017 General Election campaign (11,421 participants exposed to 783 unique Facebook political ads) finds that rather than evidence of segmentation, messages adhere closely to national campaign narratives (Anstead et al., 2018). Targeted advertising in British general elections tends to draw on well-honed national messages deployed to reach voters who are likely to be most receptive to them and deemed electorally significant (Anstead, 2017). However, a case study of Leave.EU's campaign (one of the unofficial 'Leave' campaign groups) in the 'Brexit' referendum points to evidence that Leave.EU's founder (Arron Banks) used actuaries from his insurance company to copy Cambridge Analytica's modelling (provided in Cambridge Analytica's pitch for business to Leave.EU, and initial scoping work) to identify 12 areas in the UK most concerned about the European Union, in order to target them with in-person visits from Nigel Farage. Farage (then leader of UK Independence Party (UKIP), a party that had long campaigned to leave the European Union) was regarded as vital to turning out voters who had never voted before but were passionate about leaving the European Union because of immigration concerns (Bakir, 2020).

More field studies on the practices of profiling and microtargeting are needed, but the growing prominence of analytics companies is concerning, especially regarding transparency of their activities to the data regulator, the electoral regulator and citizens. The UK's 2016 'Brexit' referendum saw 'dark ads' (online ads only seen by the recipient) being discussed in public for the first time, but three years later, by the time of the 2019 General Election, many were still unaware of these techniques. YouGov survey research commissioned by Open Rights Group showed that although 54% of the British population were aware of how political parties target or tailor ads based on analysis of their personal data (political microtargeting), almost a third (31%) were not aware at all or not very aware. Only 44% of the national sample were very or fairly aware of 'dark ads' with a similar fig. (41%) not very or at all aware. That there is still relatively low awareness after several years of public discourse on this issue is alarming: it shows that a significant proportion of the electorate are unaware of how parties may try to manipulate them. The survey finds that a majority (58%) said they were against targeting or tailoring ads, based on analysis of people's personal data to segment them into groups during elections (Open Rights Group, 2020, January 10). Furthermore, research into campaigning during the 2019 UK General Election finds that three quarters of people said that it was important for them to know who produced the political information they see online, but less than a third knew how to find out who produced it. Almost half (46%) were concerned about why and how political advertising was targeted at them (The Electoral Commission, 2019b).

Profiling and Targeting in India's Political Campaigning

India, the world's biggest democracy (with a population of 1.393 billion in 2022), provides a context of rapidly expanding access to digital services, but an inadequate data protection regime. It also ranks poorly in the world press freedom index (150th out of 180 countries in 2022) given its politically partisan media, its concentration of print news and television media ownership, and its violence against, and harassment of, journalists who are critical of the government (Reporters without Borders, 2022a). Political parties can exploit these features when campaigning.

With Internet penetration at 54% in 2022 (Krishnan, 2022), compared to the USA and UK, India suffers from a 'digital divide', but this is rapidly changing. India's 2011 census report reveals that only 19% of Dalits (one of India's most marginalised castes) had access to water, but 52% from the community owned a phone (61% in urban areas and 42% in rural areas). From December 2016 to July 2017, the number of mobile phone Internet users in India rose rapidly from 389 million to 420 million, fuelled by a decrease in data rates after a price war between Reliance-owned Jio network (a new entrant in India's telecom market) and other telecom companies (Gowhar, 2018). In terms of daily time spent using the Internet on mobiles, by 2022, India (at 4 hours 5 minutes) was ahead of the world average (of 3 hours 43 minutes) (Kemp, 2022). By the 2019 General Election, nearly half of India's 900 million eligible voters had access to the Internet and social media, and there were 300 million Facebook users (Naumann et al., 2019). By 2021, India had 410 million Facebook users, 440 million YouTube users and 530 million WhatsApp users (Ministry of Electronics & IT, 2021). Over half of India's English-speaking, online news users use Google-owned YouTube (53%) and Meta-owned WhatsApp (51%) for accessing news in 2022 (Krishnan, 2022). The changes in the top ten free apps in Play Store across 2017–2018 also reflect the growing influence of regional language social media applications that are more effective at targeting local populations. For instance, Facebook and Messenger were replaced in 2018 with more vernacular language apps such as ShareChat and Helo that operate in up to 15 different languages (mainly Hindi, Tamil and Telugu) and which target the 100–150 million mobile Internet users in rural India and tier 2 and 3 cities populated by Indian language speakers (Naumann et al., 2019).

India's data protection regime is inadequate to deal with this rapidly expanding access to digital services. India's Personal Data Protection Bill was not introduced until 2019 and, at the time of writing (Spring 2022), is still not enshrined in legislation; neither does India have a national regulatory authority for personal data protection. In the meantime, India's Information Technology Act (2000) gives a right to compensation for improper disclosure of personal information. Furthermore, the Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules 2011 imposed extra requirements on commercial entities in India relating to collection and disclosure of sensitive personal data which has some similarities with the *GDPR*. For instance, a body corporate collecting sensitive personal data should keep the data

provider informed about the fact that data is being collected; for what purposes; intended recipients; and contact details of the agency collecting and retaining the data (Linklaters, 2020). In terms of protecting elections, pre-certification of social media content was mandated by India's Electoral Commission in the 2014 General Election: ads had to be certified within the boundaries of permissibility for an electoral speech, as well as not appealing to caste or religious identity and not promoting hate speech or bribery. The Electoral Commission also has a Model Code of Conduct to promote good conduct, but this lacks enforceability (Naumann et al., 2019). In 2019, the Electoral Commission issued social media guidelines for campaigning, and there is a voluntary adoption of a code of ethics for online campaigning by Internet companies (Rao, 2019).

Given India's growing access to smartphones alongside absence of a robust data protection regime, it is unsurprising that across the past decade, India's successful politicians have turned to data-driven campaigning techniques to target electorates. Hindu nationalist and populist, Narendra Modi (leader of the Bharatiya Janata Party (BJP), the son of a tea-seller and one of a handful of lower-caste politicians to reach the upper echelons of power, was the second most popular politician on Facebook with over 18 million fans (after then US President Obama with over 41 million fans) (Barclay et al., 2015; Shackle, 2018, July 16). According to a report from Tactical Tech (an international non-governmental organisation that engages with citizens and civil society to explore and mitigate the impacts of technology on society), in the 2014 national elections, the BJP was among the first of India's political parties to employ data-driven campaigning techniques, winning a landslide victory. Its techniques included sending global positioning system-enabled video trucks to villages in the most populous and politically weighty state of Uttar Pradesh to ensure digital outreach in remote areas; using 3D hologram technology to hold 1350 3-D rallies across India at the state and constituency level; and leveraging social media to ensure outreach and engagement prior to rallies, electoral data and on-the-ground reports to inform each rally speech with local context (Hickok, 2018).

India's political campaigners claim that they can microtarget India's citizens, although given the uneven nature of digital penetration in India, this requires much fieldwork to generate reliable data streams. For instance, in the 2017 Uttar Pradesh state election, through 45,000 telephone calls a day and multiple field visits to all 403 seats in the state, voters' details including caste, voting pattern and preferred chief minister were fed into

a database for the ruling Samajwadi party. This enabled its candidates to download an app showing voter preferences down to the level of individual booths and along caste, gender and literacy lines. The field visits were necessary as telephone numbers are not always active for long in poorer areas as it is cheaper to buy a new sim card pre-loaded with call credits than buy extra credit. The field visits returned not just detailed voter lists but also relationships with local influencers, such as village chiefs, postal workers and teachers, to help report popular sentiment or convey new telephone numbers. According to the Samajwadi party campaign in Uttar Pradesh, candidates can microtarget messages they know appeal to young, college-graduate, Muslim women, for example, in booths that skew towards those demographics, and can know to call a particular influential member of a village whose support was wavering (Safi, 2017, February 16).

The need for a robust data protection regime in India has been repeatedly highlighted by data exploitation in political campaigning. For instance, in early 2017, ahead of state elections for Uttar Pradesh, the ruling party (BJP) used WhatsApp massively for mobilisation, coordination and voter outreach, forming 10,344 WhatsApp groups to coordinate and circulate media among party workers (Gowhar, 2018). However, as elsewhere, social media not only mobilises but spreads disinformation, stokes communal tension and silences dissent. For instance, on 22 April 2018, a fake tweet began circulating under the name of Rana Ayyub, an Indian political and investigative journalist, critical of Modi and the role he allegedly played in anti-Muslim riots while governor of Gujarat. The fake tweet expressed support for child rapists and was shared tens of thousands of times, including by BJP legislators. According to the Centre for International Governance Innovation (a Canadian-based, independent, non-partisan think tank on global governance), on 23 April, another false tweet appeared under Ayyub's name, saying 'I hate India and Indians' (Shackle, 2018, July 16). That evening, a deepfake pornographic video with Ayyub's face morphed onto another woman's body circulated on WhatsApp groups of the BJP and the Rashtriya Swayamsevak Sangh (an Indian right-wing, Hindu nationalist, paramilitary volunteer organisation) and was made public (European Science Data Hub, 2019, December 4; Shackle, 2018, July 16). Across 2021, there were social media campaigns from far-right Hindu nationalist activists fomenting hatred and calling for the murder of Ayyub, with her personal data posted online (Reporters without Borders, 2022a). As well as this gendered disinformation, more broadly, the 2019 General Election saw a spike in online rumours, fake

news and polarising content on social media, including on vernacular language apps such as ShareChat and Helo as well as Facebook (Naumann et al., 2019; Krishnan, 2022). As many social media app users in India are first-time Internet users, they may lack digital literacy skills to spot disinformation, especially as content shared comes from someone known, producing a tendency to trust the source (Gowhar, 2018).

As well as exploiting WhatsApp, Modi launched his NaMo mobile phone app in 2015 to engage supporters (Kazmin, 2018, March 28). The app has no visible content moderation and propagates polarising posts based on fictitious data about the religion of criminals and voter turnout. The app's news feed also promotes posts from accounts that share regular political updates on the prime minister's app and whose Facebook pages openly circulate fake news. The promotion of such accounts on the NaMo app makes its millions of users vulnerable to disinformation (Bansal, 2019, January 27). Pushed via official government channels, and pre-installed on low-cost Jio mobile phones, it has become one of the most widely used politician's apps in the world, with over ten million downloads in the Google Play Store. In late 2019 the NaMo app received a makeover that included live events, Instagram-like 'Stories' about Modi, gamified engagement strategies, means of accepting micro-donations and promises of a direct line to the prime minister. Also of note is that the transfer of digital campaigning techniques and practices does not always flow from the USA outwards but also in the opposite direction. For instance, Gursky and Woolley (2020) suggest that The Official Trump 2020 App copied Modi's tactics.

Modi's NaMo app also collected large amounts of data for years through opaque phone access requests (Gursky & Woolley, 2020). In 2018, journalists reported that the NaMo app asked users to provide access to 22 personal features on their devices, many more than the 14 data points requested by the official app of the Prime Minister's Office, 'PMO India App'. In March 2018, a day after an anonymous French cybersecurity researcher exposed on Twitter that the app was transferring user details to a third party (a US-based behavioural data analytics company, CleverTap, which helps clients to 'influence' app users' 'behaviour' by uncovering insights), the privacy setting was 'quietly' changed, drawing accusations of parallels with Cambridge Analytica's practices (Kazmin, 2018, March 28; NH Political Bureau, 2018, March 25). In India there is no legislative

restriction regarding transborder dataflows of information that is not sensitive personal data (Linklaters, 2020). The NaMo app's default permission settings gave it nearly full access to the data stored on users' phones, including photos and videos, contacts, location services and ability-to-record audio, although savvy users could opt out by disabling permissions (Kazmin, 2018, March 28).

Consider also that the ruling party, BJP, was the first recorded political party in the world to use a deepfake video in an electoral campaign (the legislative assembly elections in Delhi in 2020) for targeting rather than for spreading disinformation (MIT Technology Review, 2021). The party hired political communications company, Ideaz Factory, to create deepfakes to reach voters in the 22 different languages and 1600 dialects used in India. One that went viral was of BJP president, Manoj Tiwari, criticising the incumbent Delhi government (it reached approximately 15 million people in 5800 WhatsApp groups in the Delhi and National Capital Region). Originally speaking in English, the deepfake simulates convincing mouth movements that Tiwari is speaking in Haryanvi, the Hindi dialect spoken by target voters for the party (to try to persuade the large Haryanvi-speaking migrant worker population in Delhi from voting for the rival party) (Christopher, 2020, February 18).

Despite India's relentless campaigning 'firsts' driven by technology, and rapid changes in mobile phone and Internet penetration, critical digital literacy programmes and public awareness campaigns are minimal. Populations with little or no access to new technologies or limited skills to use them effectively are particularly susceptible to falsehoods peddled online: this includes the poor, rural populations, women, the disabled, migrants, internally displaced populations and the elderly (Rao, 2019). Survey experiments on a highly educated online sample in India on the effectiveness of a media literacy campaign (Facebook's 'Tips to Spot False News' promoted at the top of users' News Feeds in 14 countries including India in April 2017 and printed in full-page newspaper ads in India) find that the intervention improved discernment between mainstream and false news headlines by 17.5%. However, this increase in discernment did not last several weeks later; and there were no measurable effects among a representative face-to-face sample of respondents in a largely rural area of northern India, where rates of social media use are far lower (Guess et al., 2020).

Conclusion

Across the past century, professional persuaders (advertisers, public relations experts and political campaigners) have sought to understand and target audiences with tailored persuasive messages based on scientifically derived insights. This has accelerated globally across the past decade, as data management companies and data brokers joined forces with professional persuaders to exploit the affordances of 'big data', profiling technologies and microtargeting. This is evident even in regions lacking infrastructure for fixed-line Internet connections, compensated for by much higher mobile penetration. Although privacy has long been a universal human right, there are different privacy protections and levels of implementation across countries, and the technology continuously advances, facilitating privacy-invasive levels of profiling and targeting. This chapter reviewed key developments in the USA (where the globally dominant social media platforms are headquartered), the UK and India (democracies with different data protection regimes and with different digital literacies).

The USA, with its weak privacy protections, has led the way in developing profiling and targeting tools and applying them to political campaigns. US-headquartered social media platforms offer many forms of targeting utilised by political campaigns. Political campaigners also use the affordances of social media platforms to deploy psychographic and neuromarketing tools. As the case of Cambridge Analytica and Facebook shows, technological loopholes were exploited for attempted short-term influence (during a specific election campaign). Once the candidate has won the election, this cannot be undone by fines issued several years later. More recently, the development of bespoke campaign mobile phone apps, some with invasive tactics for gathering data and reaching voters, allows political campaigns to collect massive amounts of user data without needing to rely on social media platforms or expose themselves to fact-checker oversight of deceptive messaging. Most Americans do not think that information about Internet users should be made available to political campaigns for targeting voters with online ads.

At the time of writing (in Spring 2022), the UK has retained the European Union GDPR in domestic law as the UK GDPR, although keeping the framework under review. Regarded as one of the strongest and most influential privacy regulations in the world, the GDPR offers data protections on collection and processing of personal data. Yet, even in the UK, digital campaigning, profiling and targeting have sharply risen

across the past decade. With the ban on paid political advertising in broadcasting failing to apply online, and digital political campaigning characterised as a 'Wild West', almost half the population are concerned about why and how political ads are targeted at them online; and the national data regulator concluded in 2020 that there was only limited assurance that procedures were in place to protect data in digital campaigns.

Although India has been on the wrong side of the digital divide, this is rapidly changing, while digital literacy remains low and India's data protection regime and culture is still being constructed. Exploiting this situation, political parties have successfully embraced digital political campaigning and continue to push the boundaries of what is permissible and recognisable. Meanwhile, practices developed in India (on the NaMo app) have been copied in the USA. Such apps cater for minimal literacy levels; have no visible content moderation; are shared on personal networks (and hence are arguably more trusted); and greatly enable delivery of inflammatory and deceptive messages, targeted at profiled users.

Despite this accelerated move towards profiling and microtargeting voters, there are few empirical studies on their practices and impacts. Where they exist, studies find modest impacts on specific types of audience, and mixed findings regarding accuracy and prevalence of microtargeting voters. More studies are needed on the effects of continuously refined profiling and targeting techniques on voting behaviour, especially as it may only take the mobilisation of a small sliver of the population (the persuadables) to generate decisive results. Digital literacy, and awareness of profiling and microtargeting technologies for political purposes, is uneven across the world, but where people are aware, most do not want it (also see Schäwel et al., 2021).

Across the world, different types of government operating under different privacy regimes may be more or less inclined and enabled to allow and deploy such emotional AI on their citizens. Given what has been found in Part I of this book, the outlook could be bleak (for instance, widespread use of far richer, microtargeted disinformation and exploitation of divisive, conspiratorial, post-truth narratives that are highly contextually relevant). However, this is not inevitable: for instance, greater mobilisation and political engagement on issues that diverse voters care about is also possible; and regulators and civil society are increasingly alert to the perils of profiling. Part II of this book now turns to the issue of the social and democratic harms arising from false information online and how best to protect us in an era of increasingly optimised emotions.

References

All Party Parliamentary Group on Electoral Campaigning Transparency. (2020, January). *Defending our democracy in the digital age: Reforming rules, strengthening institutions, restoring trust*. Retrieved April 13, 2022, from https://fairvote.uk/wp-content/uploads/2020/01/Defending-our-Democracy-in-the-Digital-Age-APPG-ECT-Report-Jan-2020-1.pdf

Andreou, A., Venkatadri, G., Goga, O., Gummadi, K. P., Loiseau, P., & Mislove, A. (2018). Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations. In *NDSS 2018 – Network and distributed system security symposium*, February, (pp. 1–15), San Diego, United States. Retrieved April 13, 2022, from https://doi.org/10.14722/ndss.2018.23204ff.ffhal-01955309f

Andrejevic, M. (2020). *Automated media*. Routledge.

Angwin, J. (2020, April 25). Probing Facebook's misinformation machine. *The Markup*. https://www.getrevue.co/profile/themarkup/issues/probing-facebook-s-misinformation-machine-241739

Anstead, N. (2017). Data-driven campaigning in the 2015 United Kingdom General Election. *The International Journal of Press/Politics, 22*(3), 294–313. https://doi.org/10.1177/1940161217706163

Anstead, N., Magalhães, J. C., Stupart, R., & Tambini, D. (2018). Political advertising on Facebook: The case of the 2017 United Kingdom general election. In *Annual meeting of the American political science association*, August 30–September 2, Boston, MA. Retrieved April 13, 2022, from https://ecpr.eu/Filestore/PaperProposal/71b9e776-0ea8-4bf3-943e-d25fa26898b8.pdf

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences, 124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Bakir, V. (2020). Psychological operations in digital political campaigns: Assessing Cambridge Analytica's psychographic profiling and targeting. *Frontiers in Political Communication, 5*, 67. https://doi.org/10.3389/fcomm.2020.00067

Bansal, S. (2019, January 27). Narendra Modi app has a fake news problem. *HuffPost*. https://www.huffingtonpost.in/entry/narendra-modi-app-has-a-fake-news-problemin5c4d5c86e4b0287e5b8b6d52?guccounter=1

Barclay, F. P., Pichandy, C., Venkat, A., & Sudhakaran, S. (2015). India 2014: Facebook "like" as a predictor of election outcomes. *Asian Journal of Political Science, 23*(2), 134–160. https://doi.org/10.1080/02185377.2015.1020319

Bartlett, J., Smith, J., & Acton, R. (2018). *The future of political campaigning*. Demos. Retrieved April 13, 2022, from https://ico.org.uk/media/2259365/the-future-of-political-campaigning.pdf

Beckett, L. (2017, October 9). Trump digital director says Facebook helped win the White House. *The Guardian*. www.theguardian.com/technology/2017/oct/08/trump-digital-director-brad-parscale-facebook-advertising

Beniger, J. R. (1986). *The control revolution: Technological and economic origins of the information society*. Harvard University Press.

Berelson, B. R., Lazarsfeld, P. F., & McPhee, W. N. (1954). *Voting: A study of opinion formation in a presidential campaign*. University of Chicago Press.

Bond, S. (2021, November 9). Facebook scraps ad targeting based on politics, race and other 'sensitive' topics. *NPR*. https://www.npr.org/2021/11/09/1054021911/facebook-scraps-ad-targeting-politics-race-sensitive-topics#:~:text=More%20Podcasts%20%26%20Shows-,Facebook%20scraps%20ad%20targeting%20based%20on%20politics%2C%20race%20and%20other,new%20rules%20begin%20in%20January

Brown, I., Marsden, C. T, Lee, J., & Veale, M. (2020). *Cybersecurity for elections. A Commonwealth guide on best practice*. Retrieved April 13, 2022, from https://thecommonwealth.org/sites/default/files/inline/Commonwealth%20cybersecurity%20for%20elections%20guide.pdf

Ceron, A., Curini, L., & Iacus, S. M. (2017). *Politics and big data: Nowcasting and forecasting elections with social media*. Routledge, Taylor and Francis.

Chester, J., & Montgomery, K. C. (2017). The role of digital marketing in political campaigns. *Internet Policy Review, 6*(4), 1–20. https://doi.org/10.14763/2017.4.773

Christopher, N. (2020, February 18). We've just seen the first use of deepfakes in an Indian election campaign. *VICE*. https://www.vice.com/en/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp

Crowe, P., Rice, M., & Santi, M. D. (2020, June 23). *Who do they think we are? Political parties, political profiling, and the law*. Open Rights Group. Retrieved April 13, 2022, from https://www.openrightsgroup.org/publications/who-do-they-think-we-are-report/

Cummings, D. (2016, October 29). *On the referendum #20: The campaign, physics and data science – Vote Leave's 'Voter Intention Collection System' (VICS) now available for all*. Retrieved April 13, 2022, from https://dominiccummings.com/2016/10/29/on-the-referendum-20-the-campaign-physics-and-data-science-vote-leaves-voter-intention-collection-system-vics-now-available-for-all/

Cummings, D. (2017, January 30). *On the referendum #22: Some basic numbers for the Vote Leave campaign*. Retrieved April 13, 2022, from https://dominiccummings.com/2017/01/30/on-the-referendum-22-some-numbers-for-the-vote-leave-campaign/

Denham, E. (2020, October 2). *Letter to digital, culture and media and sport select committee*. Retrieved April 13, 2022, from https://committees.parliament.uk/publications/2847/documents/27859/default/

Dobber, T., Fathaigh, R., & Zuiderveen Borgesius, F. J. (2019). The regulation of online political microtargeting in Europe. *Internet Policy Review, 8*(4). https://doi.org/10.14763/2019.4.1440

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2020). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics, 26*(1), 69–91. https://doi.org/10.1177/1940161220944364

Dzisah, W. S. (2020). Social media and participation in Ghana's 2016 elections. In M. Ndlela & W. Mano (Eds.), *Social media and elections in Africa, Volume 1: Theoretical perspectives and election campaigns* (pp. 97–118). Palgrave Macmillan, Springer Nature.

Eckles, D., Gordon, B. R., & Johnson, G. A. (2018). Field studies of psychologically targeted ads face threats to internal validity. *PNAS, 115*(23), E5254–E5255. https://doi.org/10.1073/pnas.1805363115

Edelson, L., Sakhuja, S., Dey, R., & McCoy, D. (2019). An analysis of United States online political advertising transparency. Preprint retrieved from *arXiv:1902.04385*.

Endres, K. (2020). Targeted issue messages and voting behavior. *American Politics Research, 48*(2), 317–328. https://doi.org/10.1177/1532673X1987569

European Commission. (2020, December 3). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. On the European democracy action plan*. Retrieved April 13, 2022, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=COM:2020:790:FIN&from=EN

European Science Data Hub. (2019, December 4). *Deepfakes, shallowfakes and speech synthesis: Tackling audiovisual manipulation*. European Parliamentary Research Service. Retrieved April 13, 2022, from https://sciencemediahub.eu/2019/12/04/deepfakes-shallowfakes-and-speech-synthesis-tackling-audiovisual-manipulation/

European Union General Data Protection Regulation. (2016/679). *Recital 71*. Retrieved April 13, 2022, from https://www.privacy-regulation.eu/en/recital-71-GDPR.htm

Facebook. (2007, November 6). *Facebook unveils Facebook ads*. Retrieved April 13, 2022, from https://about.fb.com/news/2007/11/facebook-unveils-facebook-ads/#:~:text=NEW%20YORK%20%E2%80%94%20Facebook%20Social%20Advertising,the%20exact%20audiences%20they%20want

Federal Trade Commission. (2019a, July 24). *FTC sues Cambridge Analytica, settles with former CEO and app developer*. Retrieved April 13, 2022, from https://www.ftc.gov/news-events/press-releases/2019/07/ftc-sues-cambridge-analytica-settles-former-ceo-app-developer

Federal Trade Commission. (2019b). *United States of America before the Federal Trade Commission. In the matter of Cambridge Analytica, LLC, a corporation. DOCKET NO. 9383*. Retrieved April 13, 2022, from https://www.ftc.gov/system/files/documents/cases/1823107cambridgeanalyticaadministrativecomplaint7-24-19.pdf

Formisimo. (2016). *Digital marketing and CRO in political campaigns*. Retrieved April 13, 2022, from www.formisimo.com/blog/digital-marketing-and-cro-in-political-campaigns/

Fowler, E., Franklin, F., Michael, M., & Ridout, T. N. (2016). *Political advertising in the United States*. Routledge, Taylor and Francis.

Fukuyama, F., & Grotto, A. (2020). Comparative media regulation in the United States and Europe. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 99–219). Cambridge University Press.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality, 37*, 504–529. https://doi.org/10.1016/S0092-6566(03)00046-1

Gowhar, F. (2018). Politics of fake news: How WhatsApp became a potent propaganda tool in India. *Media Watch, 9*(1), 106–117. https://doi.org/10.15655/mw/2018/v9i1/49279

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *PNAS, 117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Gursky, J., & Woolley, H. (2020). The Trump 2020 app is a voter surveillance tool of extraordinary power. *MIT Technology Review*, 21 June. Retrieved April 13, 2022, from https://www.technologyreview.com/2020/06/21/1004228/trumps-data-hungry-invasive-app-is-a-voter-surveillance-tool-of-extraordinary-scope/

Haenschen, K., & Jennings, J. (2019). Mobilizing millennial voters with targeted internet advertisements: A field experiment. *Political Communication, 36*(3), 357–375. https://doi.org/10.1080/10584609.2018.1548530

Hao, K. (2021). How Facebook got addicted to spreading misinformation. *MIT Technology Review*, 11 March. Retrieved April 13, 2022, from https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/

Herbst, S. (2016). The history and meaning of public opinion. In A. J. Berinsky (Ed.), *New directions in public opinion* (2nd ed.). Routledge.

Hickok, E. (2018). *Digital platforms, technologies, and data use in the general elections in India*. Tactical Technology Collective. Retrieved April 13, 2022, from https://cdn.ttc.io/s/ourdataourselves.tacticaltech.org/ttc-influence-industry-india.pdf

Hopkins, C. (1923). *Scientific advertising*. Moore.

Hotham, T. (2021). *A breakdown of the 2019 General Election targeted advertising campaign across all platforms*. Retrieved April 13, 2022, from https://tristanhotham.com/2021/07/03/a-breakdown-of-the-2019-general-election-targeted-advertising-campaign-across-all-platforms/

Howard, P. N. (2018, November 30). *Claim No: CO/3214/2018. The Queen on the application of Susan Wilson & others -and-the Prime Minister, report of Dr Philip N. Howard Professor, Oxford University to the High Court of Justice, Queen's Bench Division, Administrative Court*. Retrieved April 13, 2022, from https://www.ukineuchallenge.com/wp-content/uploads/2018/12/257136-Expert-report-of-Prof-Howard-FINAL-Signed.pdf

Information Commissioners Office. (2018, November 6). *Investigation into the use of data analytics in political campaigns: A report to Parliament*. Retrieved April 13, 2022, from https://ico.org.uk/media/action-weve-taken/2260271/investigation-into-the-use-of-data-analytics-in-political-campaigns-final-20181105.pdf

Information Commissioners Office. (2020, November). *Audits of data protection compliance by UK political parties: Summary report*. Retrieved April 13, 2022, from https://ico.org.uk/media/action-weve-taken/2618567/audits-of-data-protection-compliance-by-uk-political-parties-summary-report.pdf

Jacobson, G. C. (2015). How do campaigns matter? *Annual Review of Political Science, 18*, 31–47. https://doi.org/10.1146/annurev-polisci-072012-113556

Jamieson, K. H. (1996). *Packaging the presidency: A history and criticism of presidential campaign advertising*. Oxford University Press.

Kalla, J. L., & Broockman, D. E. (2018). The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *American Political Science Review, 112*(1), 148–166. https://doi.org/10.1017/S0003055417000363

Kates, G. (2020, July 18). The Trump campaign app is tapping a 'gold mine' of data about Americans. *CBS*. https://www.cbsnews.com/news/trump-campaign-app-data-americans-gold-mine-phunware/

Kazmin, A. (2018, March 28). Narendra Modi's personal app sparks India data privacy row. *Financial Times*. https://www.ft.com/content/896cf574-31c0-11e8-b5bf-23cb17fd1498

Kemp, S. (2022). *Digital 2022: Global overview report*. Datareportal. Retrieved April 13, 2022, from https://datareportal.com/reports/digital-2022-global-overview-report

Klapper, J. T. (1960). *The effects of mass communication*. Free Press.

Korolova, A. (2010). Privacy violations using microtargeted ads: A case study. In *IEEE international conference on data mining workshops*, pp. 474–482. Retrieved April 13, 2022, from https://theory.stanford.edu/~korolova/Privacyviolationsusingmicrotargetedads.pdf

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*, 15. https://doi.org/10.1073/pnas.1218772110

Kotliar, D. M. (2020). The return of the social: Algorithmic identity in an age of symbolic demise. *New Media & Society, 22*(7), 1152–1167. https://doi.org/10.1177/1461444820912535

Kreiss, D. (2016). *Prototype politics: Technology-intensive campaigning and the data of democracy*. Oxford University Press.

Kreiss, D., & McGregor, S. C. (2018). Technology firms shape political communication: The work of Microsoft, Facebook, Twitter, and Google with campaigns during the 2016 US presidential cycle. *Political Communication, 35*(2), 155–177. https://doi.org/10.1080/10584609.2017.1364814

Kreiss, D., & McGregor, S. (2019). The "arbiters of what our voters see": Facebook and Google's struggle with policy, process, and enforcement around political advertising. *Political Communication, 36*(2), 499–522. https://doi.org/10.1080/10584609.2019.1619639

Krishnan, A. (2022). India. In N. Newman, R. Fletcher, C. T. Robertson, K. Eddy, & R. K. Nielsen (Eds.), *Reuters Institute digital news report 2022* (pp. 134–135). Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The people's choice: How a voter makes up his mind in a presidential campaign*. Columbia University Press.

Linklaters. (2020). *Data protected – India*. Retrieved April 13, 2022, from https://www.linklaters.com/en/insights/data-protected/data-protected%2D%2D-india#:~:text=General%20data%20protection%20laws&text=India%20has%20also%20not%20yet,improper%20disclosure%20of%20personal%20information

Mare, A., & Matsilele, T. (2020). Hybrid media system and the July 2018 Elections in "Post-Mugabe" Zimbabwe. In M. Ndlela & W. Mano (Eds.), *Social media and elections in Africa, Volume 1: Theoretical perspectives and election campaigns* (pp. 147–176). Palgrave Macmillan, Springer Nature.

Martínez, A. G. (2018, February 23). How Trump conquered Facebook without Russian ads. *Wired*. https://www.wired.com/story/how-trump-conquered-facebookwithout-russian-ads/

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *PNAS, 114*(48), 12714–12719. https://doi.org/10.1073/pnas.1710966114

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2018a). Reply to Sharp et al.: Psychological targeting produces robust effects. *PNAS, 115*(34), E7891. https://www.pnas.org/content/115/34/E7891

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2018b). Reply to Eckles et al.: Facebook's optimization algorithms are highly unlikely to explain the effects of psychological targeting. *PNAS, 115*(23), E5256–E5257. https://doi.org/10.1073/pnas.1806854115

McCarthy, J. (2020, March 2). In U.S., most oppose micro-targeting in online political ads. *Gallup blog*. Retrieved April 13, 2022, from https://news.gallup.com/opinion/gallup/286490/oppose-micro-targeting-online-political-ads.aspx

Ministry of Electronics & IT. (2021). *Government notifies Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021*. Press release. Retrieved April 13, 2022, from https://www.pib.gov.in/PressReleasePage.aspx?PRID=1700749

MIT Technology Review. (2021). *An Indian politician is using deepfake technology to win new voters*, 19 February. Retrieved April 13, 2022, from https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/

Naumann, K., Sen, R., & Murali, V. S. (2019). *The impact of digital media on the 2019 Indian general election*. Institute of South Asian Studies. Retrieved April 13, 2022, from https://www.kas.de/documents/288143/4518801/ISAS-Special-Report-Impact-of-Digital-Media-Full.pdf/036704f7-9656-800d-2c7f-71e5c096b657?version=1.0&t=1576720504239

Ndlela, M. N. (2020). Social media algorithms, bots and elections in Africa. In M. Ndlela & W. Mano (Eds.), *Social media and elections in Africa, Volume 1: Theoretical perspectives and election campaigns* (pp. 13–37). Palgrave Macmillan, Springer Nature.

Newman, N. (2022). United Kingdom. In N. Newman, R. Fletcher, C. T. Robertson, K. Eddy, & R. K. Nielsen (Eds.), *Reuters Institute digital news report 2022* (pp. 62–63). Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

NH Political Bureau. (2018, March 25). BJP changes privacy setting on NAMO App. *NH Political Bureau*. https://www.nationalheraldindia.com/news/bjp-changes-privacy-setting-on-namo-app

Nothias, T. (2020). Access granted: Facebook's free basics in Africa. *Media, Culture & Society, 42*(3), 329–348. https://doi.org/10.1177/0163443719890530

Open Rights Group. (2020, January 10). *Public are kept in the dark over data driven political campaigning, poll finds*. Retrieved April 13, 2022, from https://www.openrightsgroup.org/press/releases/2020/public-are-kept-in-the-dark-over-data-driven-political-campaigning,-poll-finds

Owen, D. (2014). New media and political campaigns. In K. Kenski & K. H. Jamieson (Eds.), *The Oxford handbook of political communication*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793471.013.016_update_001

Perloff, R. M. (2018). *The dynamics of political communication: Media and politics in a digital age*. Routledge.

Petrescu, M., Krishen, A., & Bui, M. (2020). The internet of everything: Implications of marketing analytics from a consumer policy perspective. *Journal of Consumer Marketing, 37*(6), 675–686. https://doi.org/10.1108/JCM-02-2019-3080

Rao, A. (2019). *Social media and critical digital literacy in India's general elections*. Institute of South Asian Studies. Retrieved April 13, 2022, from https://prca.seasia.global/wp-content/uploads/2019/06/Social-Media-and-Critical-Digital-Literacy-in-India%E2%80%99s-General-Elections.pdf

Reporters without Borders. (2022a). *India*. Retrieved June 22, 2022, from https://rsf.org/en/country/india

Reporters without Borders. (2022b). *United States*. Retrieved June 22, 2022, from https://rsf.org/en/country/united-states

Reporters without Borders. (2022c). *United Kingdom*. Retrieved June 22, 2022, from https://rsf.org/en/country/united-kingdom

Safi, M. (2017, February 16). India's 'big data' election: 45,000 calls a day as pollsters target age, caste and religion. *The Guardian*. https://www.theguardian.com/world/2017/feb/16/india-big-data-election-pollsters-target-age-caste-religion-uttar-pradesh

Santini, R. M., Salles, D., & Tucci, G. (2021). Comparative approaches to mis/disinformation. When machine behavior targets future voters: The use of social bots to test narratives for political campaigns in Brazil. *International Journal of Communication, 15*, 1220–1243. https://ijoc.org/index.php/ijoc/article/view/14803

Schäwel, J., Frener, R., & Trepte, S. (2021). Political microtargeting and online privacy: A theoretical approach to understanding users' privacy behaviors. *Media and Communication, 9*(4), 158–169. https://doi.org/10.17645/mac.v9i4.4085

Schroepfer, M. (2018, April 4). An update on our plans to restrict data access on Facebook. *Facebook*. https://newsroom.fb.com/news/2018/04/restricting-data-access/. Accessed 26 Mar 2020.

Shackle, S. (2018, July 16). *How social media can silence dissent*. Centre for International Governance Innovation. Retrieved April 13, 2022, from https://www.cigionline.org/articles/how-social-media-can-silence-dissent/

Sharp, B., Danenberg, N., & Bellman, S. (2018). Psychological targeting. *Proceedings of the National Academy of Sciences, 115*(34), E7890–E7890. https://doi.org/10.1073/pnas.1810436115

Starr, P. (2020). The flooded zone: How we became more vulnerable to disinformation in the digital era. In W. L. Bennett & S. Livingston (Eds.), *The disinformation age: Politics, technology and disruptive communication in the information age* (pp. 67–91). Cambridge University Press. https://doi.org/10.1017/9781108914628

Statista. (2021). *Digital political advertising spending in the United States from 2008 to 2020*. Retrieved April 13, 2022, from https://www.statista.com/statistics/309592/online-political-ad-spend-usa/

Stromer-Galley, J. (2014). *Presidential campaigning in the internet age*. Oxford University Press.

Taylor, K., & Silver, L. (2019, February 5). Smartphone ownership is growing rapidly around the world, but not always equally. *Pew Research Centre*. Retrieved April 13, 2022, from https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/

The Electoral Commission. (2019a). *Report: Digital campaigning – Increasing transparency for voters*. Retrieved April 13, 2022, from https://www.electoral-

commission.org.uk/who-we-are-and-what-we-do/changing-electoral-law/transparent-digital-campaigning/report-digital-campaigning-increasing-transparency-voters#spending

The Electoral Commission. (2019b). *In depth: Campaigning at the 2019 UK Parliamentary general election*. Retrieved April 13, 2022, from http://www.electoralcommission.org.uk/who-we-are-and-what-we-do/elections-and-referendums/past-elections-and-referendums/uk-general-elections/report-2019-uk-parliamentary-general-election-was-well-run/depth-campaigning-2019-uk-parliamentary-general-election.

Tufekci, Z. (2014). Engineering the public: Big data, surveillance and computational politics. *First Monday*, 19(7). Retrieved April 13, 2022, from https://firstmonday.org/ojs/index.php/fm/article/view/4901/4097

Vaidhyanathan, S. (2018). *Antisocial media: How Facebook disconnects us and undermines democracy*. Oxford University Press.

Wells, W. D. (1975). Psychographics: A critical review. *Journal of Marketing Research, 12*(2), 196–213. Retrieved April 13, 2022, from https://www.jstor.org/stable/3150443.

Wylie, C. (2018). *Oral evidence: Fake News*, HC 363, 27 March. Retrieved April 13, 2022, from http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/digital-culture-media-and-sport-committee/disinformation-and-fake-news/oral/81022.pdf

Wylie, C. (2019). *Mindf*ck: Inside Cambridge Analytica's plot to break the world*. Profile Books.

PART II

Strengthening the Civic Body

CHAPTER 7

Harms to the Civic Body from False Information Online

Introduction

In Part I, we deconstructed core features of contemporary false information online, exploring its dynamics across the world and synthesising interdisciplinary scholarship on disinformation, misinformation, affect, emotion, profiling, targeting and the increasing datafication and optimisation of emotional life. Starting with the metaphor of the *civic body*, we highlighted the interconnectedness of bodies (individual and societal) and data about emotions. We identified core incubators of false information online to be the *economics of emotion* and the *politics of emotion*—namely, optimising content for economic or political gain. We discussed how different affective contexts worldwide fuel false information, thereby highlighting the need to understand local specificities of affective contexts, as well as their intersections with international information flows (for instance, regarding information warfare, ideological struggles and platform resources for content moderation). We clarified the nature of false information and its occurrence online, drawing out implications for citizen-political communications. We investigated the role of affect, emotion and moods as an energising force in opinion formation and decision-making, which drives false information online. Finally, we delved into profiling and targeting as the core means of delivering emotively charged, false information throughout the *civic body*, exploring this dynamic in political campaigning in democracies with different data protection

V. Bakir, A. McStay, *Optimising Emotions, Incubating Falsehoods*,
https://doi.org/10.1007/978-3-031-13551-4_7

regimes and digital literacies. Building on this knowledge, Part II explores how we can strengthen the *civic body* across dominant and emergent uses of emotional AI.

Opening this discussion, this chapter examines six core social and democratic harms arising from false information on digital platforms. (1) It produces wrongly informed citizens that (2) in certain circumstances, for certain communities, are likely to stay wrongly informed in digital echo chambers and (3), more widely, be emotionally provoked (given the affective nature of much false information), thereby fuelling polarisation, partisan misperceptions, incivility and hatred. Added to this is a fourth problem: (4) contagion, where false, emotive information incubated in digital echo chambers and highly partisan enclaves influences wider social media and mainstream news, thereby spreading its pollutants far and wide. Meanwhile, (5) profiling and microtargeting raise core democratic harms comprising fragmentation of important national conversations; targeted suppression of voters; and undue influence over susceptible citizens, although this is hard to directly prove. Also related (6) is the impact of false information in seeding distrust in important civic processes and institutions.

Harm 1: Wrongly Informed Citizens

Making decisions on the basis of false information cannot be good for the individual or society. Unfortunately, studies indicate that most people have poor information hygiene; most older people are not at all confident that they can recognise false information (see Chap. 1); and people are poor at recognising deepfakes (see Chap. 4). Furthermore, misperceptions, once formed, are difficult to correct (Flynn et al., 2017, p. 130). US experiments show that repeated exposure to fake news headlines increases their perceived accuracy: this 'illusory truth effect' for fake news headlines occurs despite low levels of overall believability and even when stories are labelled as contested by fact-checkers or are inconsistent with the reader's political ideology (Pennycook et al., 2018). US experiments also show that exposure to elite discourse about fake news leads to lower levels of trust in media and less accurate identification of real news (Guess et al., 2017; Van Duyn & Collier, 2019).

Such negative effects may be unequally distributed across the *civic body*, for instance, if citizens are differentially targeted with, or exposed to, poor-quality, false information. This may be particularly problematic in

elections and more broadly among historically marginalised communities (Gandy, 2009). For instance, according to reports from American non-profit organisations, research into disinformation targeted at Spanish-language-speaking communities in the US 2018 mid-term elections and 2020 presidential elections identifies the problem of 'data voids' in search engines (Golebiewski & Boyd, 2018, May). With little high-quality Spanish-language content online on political candidates or on the voting rights of those of Latin American cultural or ethnic identity, disinformation actors fill this gap (Thakur & Hankerson, 2021).

Of course, questions of facts, reason and evidence are not the only pertinent factors in ensuring a well-functioning democracy. Farkas and Schou (2020) argue that democracy should aspire to popular sovereignty and rule by the people if it is to be true to its Greek roots: *demos* (people) and *kratos* (rule). This requires interlocking exchanges between the individual and the people, with resulting competing political ideas about how society should be structured. However, Farkas and Schou (2020, p. 8) also acknowledge that facts, reason and evidence should not be decoupled from exchange of competing political ideas. Unfortunately, this is precisely what has happened for some communities, as shown in the next harm that we discuss.

Harm 2: Remaining Wrongly Informed in Digital Echo Chambers

A second harm from contemporary false information online is that if it goes uncorrected, it could lead citizens to remain wrongly informed in echo chambers. *Echo chambers* exist where information, ideas or beliefs are amplified and reinforced by communication and repetition inside a defined system where competing views are under-represented. Sunstein (2002, p. 176) describes this group polarisation phenomenon, where 'members of a deliberating group predictably move toward a more extreme point in the direction indicated by the members' predeliberation tendencies'. He ascribes this group polarisation to two forces. Firstly, 'social comparison': namely, people's desire to maintain their reputation within the group and self-conception. The second force is 'persuasive arguments': there are limited 'argument pools' within a group whose members are already inclined in a certain direction, with a disproportionate number of arguments supporting that same direction, so the result of discussion will be to move

individuals further in the direction of their initial inclinations. Echo chambers would be problematic for democracy, because, to make informed decisions, citizens need access to, and engagement with, a sufficiently diverse body of information about public life (Sunstein, 2002, 2017). Sunstein (2002, p. 195) concludes that for deliberation to be valuable as a social phenomenon, we should 'create spaces for enclave deliberation without insulating enclave members from those with opposing views, and without insulating those outside of the enclave from the views of those within it'. Certainly, across the world, people state that they do not desire echo chambers in their news diet. In 2021, Reuters Institute surveyed the digital news consumption of people in 46 countries, finding that most (74%) prefer news that reflects a range of views (Newman et al., 2021).

It is much debated whether echo chambers are natural psycho-social phenomena, the product of the digital media ecology, or even exist at all. On the side of nature, a long line of research highlights the role of people's natural biases and cognition processes. *Selective exposure*, where people prefer and tune into information that supports their existing beliefs, is an old and consistent finding in communication research, but operates mainly among a small minority of highly partisan individuals (Arguedas et al., 2022; Lazarsfeld et al., 1944). A closely related psychological phenomenon is *confirmation bias*, or people's tendency to search for, interpret, notice, recall and believe information that confirms their pre-existing beliefs (Wason, 1960). Another related phenomenon is *motivated reasoning*—an information processing theory that holds that citizens are more accepting of false information that matches their pre-existing worldview (Kunda, 1990; Walter et al., 2020).

Fears have been expressed that when selective exposure, confirmation bias and motivated reasoning are combined with false information fed into self-reinforcing algorithmic systems (namely, filter bubbles), there is little chance of citizens correcting the false information and hence they will remain within their digital echo chamber. Pariser (2011) posits that 'filter bubbles' arise when algorithms applied to online content selectively gauge what information users want to see based on information about the users, their connections, browsing history, purchases and what they post and search. This results in users becoming separated from exposure to wider information that disagrees with their views.

Whether digital echo chambers and filter bubbles exist and whether they are a democratic problem has been vigorously debated. Synthesising these studies, the following sections present empirical evidence that

indicate that digital echo chambers exist on some social media platforms for some communities and are damaging and, conversely, that digital echo chambers are minimal and do not pose a threat.

Digital Echo Chambers Exist for Some and Are Damaging

A number of studies suggest that digital echo chambers exist on some social media platforms for some countries and communities. For instance, a field experiment conducted in 2018 of over 17,000 American participants randomly offered participants subscriptions to conservative or liberal news outlets on Facebook. It then examined the causal chain of media effects (subscriptions to outlets, exposure to news on Facebook, visits to online news sites, sharing posts, and changes in political opinions and attitudes). It finds that news sites visited through Facebook are associated with more segregated, pro-attitudinal and extreme news, compared to other news sites visited, and that Facebook's content-ranking algorithm may limit users' exposure to news outlets offering viewpoints contrary to their own (Levy, 2021).

Big data studies also find evidence of digital echo chambers. Analysis of the USA's press and social media landscape across 18 months leading up to the 2016 presidential election shows that the right-wing media ecosystem (dominated by Breitbart and Fox News) was more insulated than the left-wing media ecosystem and so was susceptible to disinformation (Faris et al., 2017, August 16). Computational approaches from other countries also empirically demonstrate that digital echo chambers exist on specific platforms for some communities and result in limited exposure to, and lack of engagement with, different ideas and other people's viewpoints (Bessi et al., 2016; Cinelli et al., 2021; Cossard et al., 2020; del Vicario et al., 2016; Milani et al., 2020). For instance, Milani et al.'s (2020) social network analysis of how vaccination-related images are shared on Twitter (over 9000 English-language tweets from 2016) finds pro- and anti-vaccination users formed two polarised networks that hardly interacted with each other and disseminated images among their members differently. Bessi et al.'s (2016) examination of information consumption patterns of 1.2 million Italian Facebook users shows that their engagement with verified content (science news) or unverified content (conspiracy news) correlates with the number of friends having similar consumption patterns (homophily). While there is a scarcity of comparative studies across platforms on digital echo chambers, one such analysis of over

100 million pieces of content on controversial topics (including gun control, vaccination and abortion) from Facebook, Twitter, Reddit and Gab (sampling different time periods across 2010–2017) finds differences between the platforms. Digital echo chambers dominate online interactions on Facebook and Twitter (platforms that did not have a feed algorithm tweakable by users), but not in Reddit and Gab (platforms whose feed algorithm was tweakable by users). The study's comparison of news consumption on Facebook and Reddit also finds higher segregation on Facebook (Cinelli et al., 2021).

Evidence from computational approaches shows that users accept confirmatory information on Facebook even if containing deliberately false claims (Bessi et al., 2014, 2016). For instance, Bessi et al.'s (2016) Italian Facebook study finds that users who are polarised towards conspiracy are most inclined to spread unverified rumours. Other studies show that dissenting information is mainly ignored or might even increase group polarisation. For instance, Zollo et al. (2017) examine the effectiveness of debunking through a quantitative analysis of 54 million US Facebook users across five years (2010–2014), comparing how users interact with proven (scientific) and unsubstantiated (conspiracy-like) information. They find that attempts at debunking are largely ineffective because only a small fraction of consumers of unsubstantiated information interact with the posts; those few are often the most committed conspiracy users; and rather than internalising debunking information, they often react to it negatively by retaining, or even increasing, engagement with the unsubstantiated information.

Digital Echo Chambers Are Minimal and Not a Threat

Yet, a sizeable body of research suggests that the extent and threat of digital echo chambers and filter bubbles has been overblown. While search engines were the anecdotal evidence for filter bubbles used by the originator of the concept, Eli Pariser (2011), studies of personalisation in Google News in the USA and Germany find only small differences between news stories suggested to different 'profiles' (Haim et al., 2018; Nechushtai & Lewis, 2019). Research by Facebook into how 10.1 million American Facebook users interacted with socially shared news across 2014–2015 finds that it is users' clicking behaviour on its platform that plays a larger role than algorithmic ranking in limiting exposure to contrary content

(Bakshy et al., 2015). Studies on Twitter that look beyond inherently ideological or polarised communities also find less homophily and polarisation in non-political contexts, observing considerable cross-connections between political groups (Bruns, 2019). For instance, a network analysis mapping Australian Twitter's follower connections from 2015 to 2016 finds many interconnections around topics from politics to sports, although also finding that for some topics (hard-right politics, education and porn) followers have very few interconnections with others (Bruns et al., 2017). A network analysis mapping Norwegian Twitter's follower connections in 2016 also suggests that digital echo chambers did not exist there at that time (Bruns & Enli, 2018). A big data study conducted in 2020 to evaluate the effectiveness of rumour rebuttal about COVID-19 on China's Weibo concludes that there might not be a significant digital echo chamber effect on community interactions (Wang & Qian, 2021).

Some studies further show that social networks lead to greater exposure to diverse ideas (Flaxman et al., 2016; Messing & Westwood, 2012). For instance, a study of a three-month period in 2013 of web browsing histories for 50,000 American users who regularly read online news finds that this both increases ideological segregation (namely, echo chambers) and (counterintuitively) exposure to diverse perspectives (Flaxman et al., 2016). An online survey of incidental exposure to news on social media in Australia, Italy, the UK and USA in 2015 finds that incidentally exposed users use significantly more online news sources than people who never use social media (Fletcher & Nielsen, 2018).

Surveys that ask users about their overall media diet (rather than their activity on a single social media platform) also find that echo chambers are very small. Surveys on samples representative of Internet users in Denmark, France, Germany, Greece, Italy, Poland, Spain, UK and USA between 2015 and 2018 find that social media mostly do not constitute digital echo chambers or filter bubbles, as most users see a mixture of political content with which they agree and disagree (Vaccari & Valeriani, 2021). Fletcher et al.'s (2021) study of online survey data in 2020 from seven countries (Austria, Denmark, Germany, Norway, Spain, UK, USA) finds that while politically partisan online news echo chambers exist, in most countries, only a minority (about 5% of Internet users) inhabit them. The figure for the USA is slightly higher: on average, 10% are in a left-wing online news echo chamber, and 3% in a right-wing online news echo chamber.

On balance, then, the research indicates that digital echo chambers and filter bubbles exist on some social media platforms for some communities, but do not exist for search engine results, other social media communities, or for most people.

Harm 3: Affective Content, Polarisation, Partisan Misperceptions, Incivility and Hate

A third harm from false information online is that it is often deliberately affective, as explained in Chaps. 2, 3 and 5. This promotion of content with high emotional appeal can generate various harms including encouraging affective polarisation and extreme views, fuelling partisan misperceptions, promoting incivility and increasing hate crimes.

In March 2021, Facebook executives circulated a memo to employees to discredit the idea that its social media platforms contribute to political polarisation. In testimony before a US House of Representatives subcommittee that month, Mark Zuckerberg instead blamed the USA's media and political environment. Indeed, Chap. 3 highlights the long-standing affectively polarised media and politics of the USA. Yet, this does not absolve social media platforms, as studies show the importance of polarising discourse on affective polarisation and ideological polarisation. A recent review of empirical studies on social media and polarisation (most of them US-based) concludes that social media shapes affective and ideological polarisation through partisan selection, message content, platform design and algorithms (Van Bavel et al., 2021). Bail's (2021) study of thousands of US-based social media users concludes that although the source of political tribalism on social media lies deep inside Americans, tapping their fears and resentments, social media distorts and amplifies these already strong emotions, fuelling status-seeking extremists and muting moderates who see little point discussing politics on social media.

Indeed, leaked Facebook documents confirm that extreme positions on social media are encouraged algorithmically. Facebook's internal research from 2016 found extremist content thriving in over a third of large German political groups on the platform. Swamped with racist, conspiracy-minded and pro-Russian content, 64% of new members in extremist groups joined because of Facebook's recommendation tools (Horwitz & Seetharaman, 2020, May 26). Leaked Facebook documents from 2019 include a report titled 'Carol's Journey to QAnon' (a cult that holds that

a cabal of Satanic cannibals operates a global child sex trafficking ring and conspired against Donald Trump while he was US president). The documents examine how Facebook's recommendation algorithms affected the feeds to an experimental account representing a conservative mother in North Carolina. It finds that rapid polarisation was an entrenched feature in the platform's operation: the first QAnon page landed in the conservative user's feed in just five days, even though the account set out to follow conservative political news and humour content and began by following high-quality conservative pages (Timberg et al., 2021, October 22).

Such social media polarisation, in turn, can skew the actual political offer. A study of US Twitter politicians and their followers from 2010 finds that politicians with more extreme ideological views had more followers than those with less extreme views. If politicians use social media feedback to inform their political stance, and if social media represents polarised views back to politicians, this can escalate polarised political offerings (Hong & Kim, 2016). Indeed, Chap. 2 points to leaked Facebook documents that show that political parties in Poland, Spain, India and Taiwan objected to Facebook's change to its algorithm in 2018 (that rewarded more emotionalised engagement and reshares) on the grounds that it forced them into more negative, extreme policy positions in their communications on Facebook to reach wider audiences (Hagey & Horwitz, 2021, September 15; Pelley, 2021, October 4).

Such affective content may also fuel partisan misperceptions. Politically motivated reasoning is thought to be driven by automatic affective processes that establish the direction and strength of biases (Taber & Lodge, 2006, p. 756), with people updating their beliefs towards political objects using their existing affective evaluations (Flynn et al., 2017). Indeed, Chap. 5 discusses several American studies that show that there are notable increases in belief in fake news as audience emotionality increases and that people are more likely to believe fake news political headlines that align with their existing beliefs (also see Weeks, 2015).

A further problem arising from the affective nature of false information online is the relationship between affective content and incivility. If civility constitutes political argumentation characterised by speakers who present themselves as reasonable, courteous and respectful of those with whom they disagree (Berry & Sobieraj, 2014), incivility involves 'speech that is impolite, insulting, or otherwise offensive' (Ott, 2017, p. 62). Online incivility levels differ greatly worldwide, according to Microsoft (2021), and worsened during the first year of COVID-19, especially public (rather

than private) interactions, and for women. While passionate politics is lauded by some, for others incivility is the antithesis of the norms of a well-functioning democracy which requires citizens and politicians to engage respectfully, even on controversial topics. As with media effects research in general, studies on the extent and impacts of mediated incivility on politics are contradictory and mixed (for overviews, see Otto et al., 2019). For instance, American studies show that exposure to mediated political incivility (namely, violation of social norms in the media) erodes political trust and decreases perceived legitimacy of political figures (Fridkin & Kenney, 2008; Mutz, 2007). A study in the Netherlands, UK and Spain shows that mediated political incivility reduces political participation intention and policy support (Otto et al., 2019). Yet, more positively, incivility and negative political speech can enable social engagement and information diffusion, leading to higher participation and voter turnout (Geer & Lau, 2006; Lu & Myrick, 2016).

While incivility can be democratically beneficial as well as harmful, scholarship on hate crimes is less equivocal. Several studies show that social media usage has measurable causal effects on hate crimes. One study isolates the causal effect of anti-refugee social media posts (on the Facebook page of Germany's far-right AfD party) on hate crimes against refugees by examining associations with local Internet and Facebook outages. The association between Facebook posts and attacks disappears in localities where Internet outages prevented access to Facebook (Müller & Schwarz, 2020). Similar results are found in a longitudinal study (2007–2018) on the causal effects of Russia's most popular social media platform, VKontakte (VK), on ethnic hate crimes and xenophobic attitudes in Russia. According to the study conducted by the US-based, non-partisan, National Bureau of Economic Research, the presence of this platform (measured by its extent of penetration across Russian cities) significantly increases hate crime in areas where there is pre-existing support for nationalist and xenophobic political party, Rodina (Bursztyn et al., 2019).

Harm 4: Contagion

In 2012, Facebook demonstrated that emotional expression is contagious on its platform (although it should be noted that expressions and the emotion that a person may be undergoing can be quite different (McStay, 2018)). Studies have since confirmed similar contagion of emotional expression on social media platforms from the USA (Facebook and

Twitter) and beyond (China's Weibo) (see Chap. 5). Although there is little consensus on what type of emotion expressions lead to stronger contagion, especially considering different languages and cultures (Goldenberg & Gross, 2020), joy, moral-emotional words and especially anger appear to be front runners. Indeed, according to leaked Facebook documents, when Facebook tweaked its News Feed algorithm in 2018 in search of increased user engagement, it made Facebook an angrier place (Hagey & Horwitz, 2021, September 15).

It has also been shown that false information is contagious online, influencing mainstream news and wider social media, thereby spreading its pollutants far and wide. Chapter 4 documents big data studies on Twitter that find that falsehood diffuses significantly farther, faster, deeper and more broadly than the truth, inspiring fear, disgust and surprise; that misinformation spreads faster and more widely than fact-checking content; and that low-credibility content is equally or more likely to spread virally as fact-checked articles.

Emotional and deceptive contagion online and offline also works through careful organisation by architects of disinformation. For instance, following Donald Trump's loss of the 2020 presidential election to Joe Biden in November 2020, in the run-up to the 6 January 2021 congressional certification of electoral votes, Trump riled up his supporters via Twitter and Facebook. He repeatedly summoned them to Washington to protest, leading to pro-Trump supporters storming the Capitol on 6 January 2021. In Facebook's internal analysis several months later, titled 'Stop the Steal and Patriot Party: The Growth and Mitigation of an Adversarial Harmful Movement', Facebook notes that 67% of Stop the Steal joins came through group invites: and 30% of invites came from just 0.3% of inviters. Such activity also helps avoid enforcement of Facebook's content moderation as backup groups replace disabled groups. The Facebook report highlights how it was unable to cope with this level of growth:

> In response [to the rapid growth of anti-quarantine Groups], a cap of 100 invites/person/day was implemented. We released an additional new invite rate limit of 30 adds/hour (now deprecated) during the growth of Stop the Steal Groups for users adding new friends (<3 days) to new groups (<7 days) to Groups with some certain ACDC properties. However, all of the rate limits were effective only to a certain extent and the groups were regardless able to grow substantially. (Mac et al., 2021, April 26)

In terms of false information on social media infecting the press, big data studies of the American press and social media landscape in the 18 months prior to the 2016 presidential election conclude that while highly partisan and clickbait news sites existed on both sides of the partisan divide, especially on Facebook, on the right-wing these sites were amplified and legitimated through an 'attention backbone' that tied the most extreme conspiracy to bridging sites such as Breitbart (Faris et al., 2017, August 16; also see Benkler et al., 2017). Another computational study investigating the role of fake news in the online media landscape from 2014 to 2016 finds that not only is fake news particularly responsive to the agendas of partisan media across many issues but also it has a relatively stable ability to influence the entire mediascape. Across all three years, fake news set the agenda for the key issue of international relations, and for two years, it set the agenda on the economy and religion (Vargo et al., 2018). Mainstream news media pay attention to fake news because exposing and correcting lies is a basic imperative of the journalistic profession. Less charitably, covering fake news stories is made much easier by the growth of independent fact-checkers whose fact-checking provides information subsidies for news organisations (Tsfati et al., 2020). Arguably, according to a study from Data and Society (an independent, non-profit, US research organisation that seeks evidence-based public debate about emerging technology), mainstream news also amplify false information in media environments where there is low public trust in media; a proclivity for sensationalism; lack of resources for fact-checking and investigative reporting; and lack of media pluralism at the hands of corporate consolidation (Marwick & Lewis, 2017). More worryingly, in countries where the state or political parties have undue political or commercial influence over legacy media, disinformation narratives developed online have ready outlets for widespread contagion.

Harm 5: Microtargeting

Profiling and microtargeting practices have been empirically demonstrated in elections in the USA, UK and India (see Chap. 6). They raise three key democratic harms: fragmentation of important national conversations; targeted suppression of voters; and undue influence.

Fragmentation of National Conversations

Data-driven politics is about communicating efficiently, talking to voters who are most useful to a campaign. Microtargeting has potential democratic benefits such as reaching social groups that are hard to contact, increasing knowledge among voters about individually relevant issues, and increasing the efficiency of political parties' campaigns. However, as Anstead (2017, p. 309) argues, 'inefficient targeting' might lead to better democratic outcomes as it could include more people in the electoral conversation. The UK's data regulator agrees that it is essential that 'voters have access to the full spectrum of political messaging and information and understand who the authors of the messages are' (Information Commissioners Office, 2018, November 6).

The opacity of online profiling and targeting provides capacity for 'dog whistle' campaigns that emphasise a provocative position only to sympathetic audiences while remaining invisible to others. It also enables targeted, secretive delivery of 'wedge' issues (namely, issues that are highly important to specific segments of a voting population) to mobilise small, but crucial, segments (Tufekci, 2014). According to a report for the Electoral Reform Society (a UK-based independent campaigning organisation which promotes electoral reform), such activities could lead to campaigners focusing on voters in marginal seats while ignoring voters considered less politically valuable, such as those in traditionally safe electorates (Dommett & Power, 2020). Indeed, during the 2019 UK General Election, ads tended to be targeted at marginal constituencies and certain demographics. For example, early in the campaign, the Conservatives pitched ads about the National Health Service, schools and police to women, while men received a 'Get Brexit Done' message. As observed by First Draft (a now ceased, non-profit coalition that provided practical guidance on how to find, verify and publish content sourced from the social web), such microtargeting matching of content with demographics had not been done in previous British elections (First Draft, 2019).

The importance of, and threat to, shared national conversations must be recognised where microtargeted ads deprive recipients of wider, diverse collective scrutiny of the messages therein. For instance, a study by First Draft during the 2019 UK General Election finds that a significant number of ads from all political parties contained statements flagged as at least partially incorrect by independent fact-checkers (Newman et al., 2020). If such false information disseminates through microtargeting and if this is

not scrutinised by mass media (or if citizens are no longer paying attention to such sources), then there is little chance of those elected on such platforms being held to public account. While the UK has some protection from fragmentation of important national conversations in that it has a well-funded and regulated broadcasting sector, and over 50% of its population trust broadcast, local and regional news (Newman, 2022), this is not the case in all parts of the world. Furthermore, such microtargeting makes it difficult for regulators to enforce advertising rules because, by the very nature of the microtargeting, a regulator is unlikely to see those ads. This risk will intensify if algorithmic marketing techniques become available to all political parties, as the UK's data regulator observes has already happened in the UK (Information Commissioners Office, 2020, November) (see Chap. 6). This would enable parties to routinely run millions of algorithmically tuned messages, on a scale that could overwhelm regulators, with deleterious consequences for the transparency and political accountability of campaigns.

Targeted Suppression of Voters

There are few academic studies on targeted voter suppression online, not least because of methodological difficulties in studying this area. A big data study of American voters and Twitter in the 2018 mid-term elections failed to find evidence of voter suppression, but this may be due to methodological failings (Deb et al., 2019) and does not mean that voter suppression is not attempted. Certainly, parliamentary inquiries, investigative journalists, civil rights groups and think tanks have unearthed multiple offers and efforts to dissuade certain types of people from voting.

For instance, in the UK, evidence submitted to the UK Inquiry into Disinformation and Fake News describes a pitch during the 'Brexit' Referendum campaign to the Leave.EU group from Cambridge Analytica/SCL Group to choose their company for electoral data analytics. Part of this pitch offered voter suppression, namely, 'groups to dissuade from political engagement or to remove from contact strategy altogether' (Bakir, 2020). Similarly, in the 2016 US presidential campaign, Trump's digital campaign (called 'Project Alamo') involved Cambridge Analytica working with the Republican National Committee. Brad Parscale, the digital director of Trump's campaign in 2016, reportedly used Facebook's Lookalike Audiences ad tool to identify voters who were not Trump supporters, to then target them with psychographic, personalised negative

messages designed to discourage them from voting. Campaign operatives openly referred to such efforts as 'voter suppression' aimed at three targeted groups: idealistic White liberals, young women and African Americans (Green & Issenberg, 2016, October 27). This targeted voter suppression of Black Americans was confirmed in 2020 by investigative journalists, based on leaked data used by Project Alamo on almost 200 million American voters. It found that in 16 key battleground states, millions of Americans were separated by an algorithm into one of eight categories, to then be targeted with tailored ads on social media: one of the categories was named 'Deterrence' and disproportionately held 3.5 million Black Americans. While causality cannot be proven, not least as there are numerous sources of voter suppression in the USA beyond online campaigning efforts (Boyd-Barrett, 2020), the 2016 campaign preceded the first fall in Black turnout in 20 years and allowed Trump to win in key states by thin margins (Channel 4 News Investigations Team, 2020, September 28). According to a report from the Center for Democracy and Technology (a US non-profit organisation whose stated aims include enhancing freedom of expression globally and stronger legal controls on government surveillance), attempted targeted suppression of Spanish-language-dominant voters in the 2020 US presidential elections has also been observed, with disinformation about basic voting details and messaging intended to intimidate such voters (Thakur & Hankerson, 2021).

Undue Influence

If profiled and behaviourally driven messages are being used to try to surreptitiously influence people, then this may contravene the right to Freedom of Thought (Alegre, 2017, 2021, May). This right protects our mental inner space. It formally became international law in 1976 as part of Article 18 of the International Covenant on Civil and Political Rights (McCarthy-Jones, 2019). Alegre notes that 'the concept of "thought" is potentially broad including things such as emotional states, political opinions and trivial thought processes' (Alegre, 2017, p. 224). It includes the right to keep our thoughts private, the right not to have our thoughts manipulated and the right not to be penalised for our thoughts and opinions (Alegre, 2017, 2021, May). McCarthy-Jones (2019, p. 2) adds that thought includes attentional and cognitive agency, as well as external actions that are arguably constitutive of thought (such as reading, writing and many forms of Internet search behaviour). Unsurprisingly, then,

freedom of thought (unlike freedom of expression) is protected as an absolute right in international human rights law: in other words, there are no restrictions allowed (Alegre, 2017). Freedom of thought has been described as 'the foundation of democratic society' and 'the basis and origin of all other rights' (Alegre, 2017, p. 221). Yet, this right has received little attention in the courts, partly because of an assumption that our inner thoughts were beyond reach (Alegre, 2017; McCarthy-Jones, 2019). This lacuna is problematic as recent developments in technology are providing new ways to access, alter and potentially manipulate our thoughts in ways we had not previously conceived.

While challenges to the right to freedom of thought are well worth exploring, it remains the case that studies on the impact of political messages on political behaviour are at best, mixed, with more studies finding minimal effects, but with recent studies also finding that targeted, data-driven campaigns have some influence (as discussed in Chap. 6). With few studies examining actual effects on political behaviour of false information campaigns conducted on social media, Bail et al.'s (2020) US study is instructive in calling into question their effectiveness. Their study uses longitudinal survey data and privileged access to Twitter data to assess the impact in late 2017 of a Russian Twitter false information campaign on political attitudes and behaviours of frequent American-based Twitter users who identified as either strong or weak partisans. They show that it was those users who were already highly polarised that engaged the most with the misinformation content. They also find no evidence that interacting with accounts linked to the false information campaign substantially impacted issue attitudes, partisan stereotypes or political behaviours that they measured. Proving that undue influence has taken place is hard. Yet, with few studies attempting to disentangle the influence of online disinformation in real-world settings, it would be unwise to dismiss concerns about undue influence at this stage, especially regarding carefully crafted and targeted disinformation. What is more apparent, however, from user-based studies in the USA and UK, is that people dislike the premise of being manipulated via their emotions, especially for political ends (Andalibi & Buss, 2020; also see Chap. 9).

It is also worth reflecting on the conditions that would enable undue influence (or manipulation). In their discussion of the online environment, Susser et al. (2019, p. 3, 26) define manipulation as using hidden or covert means to subvert another person's decision-making power, undermining their autonomy. Bakir et al. (2019) argue that persuasive

communications, to avoid being manipulative, should be guided by principles akin to informed consent. In short, to ethically persuade (rather than manipulate) people towards a particular viewpoint, the persuadee's decision should be both informed (with *sufficient information* provided and none of it of a *deceptive* nature) and *freely chosen* (namely, no coercion or incentivisation). Yet, these conditions would be disabled by widespread false information (*deception*) driven by affect and emotion (that prompt gut reactions, thereby raising questions about the extent to which the decision was *freely chosen*), and profiling and targeting (that might exclude people from exposure to *sufficient* information).

Harm 6: Seeding Distrust in the Civic Body

False information seeds distrust in important civic processes and institutions, from health messaging to democratic processes.

Where the knowledge base is uncertain, people are more susceptible to false information, as evidenced by the COVID-19 pandemic (explored in Chap. 5). The impacts of false COVID-19 information on trust in government vary across the globe. Several months into the pandemic, Newman et al.'s (2020) survey of six countries (conducted in April 2020) finds high levels of trust in news and information about COVID-19 from scientists and doctors (83%), national health organisations (76%) and global health organisations (73%). However, only a small majority trusts the national government (59%) and news organisations (59%), raising concerns about the impact of public health messaging where behaviour change is needed across the entire population. In some countries, this is an improving or deteriorating situation depending on how governments have responded. In Vietnam, for instance, initially confusing governmental responses in a chaotic sphere of false information online and incivility greatly heightened public anxiety and fear. However, this then forced Vietnam's one-party state to become unusually transparent in responding to public concerns across 2020, leading to every new COVID-19 case being immediately published on governmental websites, mainstream and social media (Nguyen & Nguyen, 2020). By contrast, in Africa, governmental denial, secrecy and misinformation, together with the fact that mainstream media are state controlled, encouraged alternative narratives of the COVID-19 crisis, especially online (where WhatsApp and Facebook are the two most common platforms). It also encouraged public distrust, apprehension and ambivalence towards public health messaging by governments in many

parts of the continent. This builds on a long-standing cultural practice where rumour is how state narratives are routinely subverted in challenges by the public, civil society and religion (Ogola, 2020).

Disinformation also seeds distrust in wider democratic processes—an information warfare aim of Russia and, to a lesser extent, China, but also engaged in by populist domestic actors (see Chap. 2). While the impact of such efforts is unclear, a study by the Australian Strategic Policy Institute (a government-funded defence and strategic policy think tank) suggests that the very perception of interference could be enough to threaten democratic outcomes if, for instance, people refuse to accept that an election result is legitimate (Hanson et al., 2019). For example, during the 2020 US presidential election, President Trump repeatedly made false statements, attacking the integrity of the USA's voting process, spawning diverse false claims online (Clayton et al., 2020; Lytvynenko & Silverman, 2020, November 3). Surveys indicate that among Trump's supporters, the cumulative impact of such claims erodes trust and confidence in elections and increases belief that the election is rigged (Clayton et al., 2020, also see Pennycook & Rand, 2021). The conviction behind such false beliefs is evident in that it resulted in a violent mob descending on Capitol Hill on 6 January 2021 to overturn the election results. It is telling that in Facebook's internal analysis, it notes that Stop the Steal and Patriot Party were harmful at the network level: 'as a movement, it normalized delegitimization and hate in a way that resulted in offline harm and harm to the norms underpinning democracy' (Mac et al., 2021, April 26). Indeed, some psychological studies find that exposure to anti-government conspiracy theories lowers intention to vote and decreases political trust among American and British citizens (although in other countries such as Germany, it increases intention to engage in political action) (Douglas et al., 2019: 20; Kim & Cao, 2016).

Conclusion

There are numerous social and democratic harms to the *civic body* arising from false information online. Across the six core harms that we have identified in this chapter, false information attacks our shared knowledge base, our togetherness, our democratic institutions and processes, and perhaps even our individual agency (although more studies are needed on this aspect).

In terms of harm 1 (wrongly informed citizens), people have trouble recognising fake news and deepfakes; and misperceptions, once formed, are difficult to correct. Furthermore, there is some evidence from the USA that, due to data voids, marginalised Spanish-language communities are exposed to poor-quality, false information on voting. More research into the extent to which this harm is present in countries beyond the USA as well as the extent to which it is unequally distributed across *civic bodies* is needed.

In terms of harm 2 (remaining wrongly informed in echo chambers), on balance, most scholarship agrees that digital echo chambers and filter bubbles exist for certain communities (right-wing, anti-vax and conspiracy groups, and on controversial topics), in certain countries (the USA and Italy, themselves, polarised societies) and on some platforms (on Twitter and on Facebook, the world's biggest social media platform). This incubates conspiracy theories, rumours and fake news, and makes users resistant to debunking. Overall, however, digital echo chambers and filter bubbles are inhabited by a small proportion of national populations; social networks and recommendation algorithms lead to greater exposure to diverse ideas and news; and the effect of personalisation on news exposure on multiple platforms is smaller than often assumed. Yet, it is concerning that some communities remain wrongly informed in echo chambers and that this helps drive false information online. Furthermore, whether or not this is an improving or deteriorating situation is difficult to determine because how platforms are used and who is on them changes, as do platforms' algorithms (as detailed in Chap. 2), but, to date, platforms largely have not made available to researchers their internal data or algorithms. This is especially so for researchers in small economies such as Guatemala or Honduras, a situation that journalist Luis Assardo (2021, August 27) terms 'the disinformation backyard'. Clearly, more research is needed into digital echo chambers and filter bubbles, ideally with access to the platforms' data and algorithms. The research should be conducted across a wider range of countries than those examined to date (largely Western democracies, especially the USA and Italy, both of which are polarised countries and prefer more partial news). We know very little, for instance, about digital echo chambers in countries that depend on Facebook to access the internet (via Free Basics). It is also vital to consider the wider information ecology, and people's overall consumption of news, rather than focusing on single platforms.

Multiple countries have experienced harm 3, where the deliberately affective nature of false information online encourages affective polarisation and extreme views (found in the USA, Germany, Poland, Spain, India and Taiwan); fuels partisan misperceptions (found in the USA); promotes hate crimes (found in Germany and Russia); and promotes mediated incivility (found in the USA, Netherlands, UK and Spain). While most of these impacts are viewed as unequivocal harms, the rise in mediated incivility has a mixed reception because as well as generating harms (eroding political trust, decreasing the perceived legitimacy of political figures and reducing political participation intention and policy support), it can also lead to increased social engagement, information diffusion, higher participation and voter turnout, all of which are democratically valuable. More research on social media's role in the various aspects of harm 3, and how this harm manifests in countries beyond the USA, would be worthwhile.

Big data studies evidence harm 4, contagion. Emotion expression, especially anger, joy and moral-emotional words, is contagious online on social media platforms based in the USA and China. Deception is also contagious on Twitter, inspiring fear, disgust and surprise, and spreading faster and more widely than fact-checking. Despite their content moderation actions, social media platforms have been unable to prevent the growth of harmful adversarial movements such as Stop the Steal in the USA. Studies, especially from the USA, show that false information on social media infect the wider press. More studies are needed to explore the extent to which deception is contagious on platforms other than Twitter and the extent to which false information online influences wider media in countries other than the USA.

In terms of the various harms stemming from microtargeting (harm 5), studies so far are indicative rather than conclusive. On fragmentation of important national conversations, a study shows that in the 2019 British General Election, for the first time, ads (many containing false information) tended to be targeted at marginal constituencies and certain demographics. While the UK has some protection from fragmentation of important national conversations in that it has a well-funded and regulated broadcasting sector, and over half the population trust mainstream news outlets, this is not the case in all parts of the world. Fragmentation of important national conversations from routinely running millions of algorithmically tuned messages, thereby damaging the transparency and political accountability of campaigns, remains feasible in all countries running digital political campaigns, especially where those countries lack media

outlets with broad reach and trust. On the harm of targeted suppression of voters arising from profiling and microtargeting, this service has already been offered (in the UK) and implemented (in the USA). Lastly, in all countries where social media platforms have a presence, the potential for undue influence of citizens is present. Although studies have yet to prove undue influence, the idea of emotional manipulation is disliked by (American and British) populations. More studies across the world are needed, to examine to what extent targeted groups are subjected to insufficient, deceptive and affective content during important periods of civic activity (such as voting periods, census periods and vaccination drives).

Harm 6, seeding distrust in the *civic body*, is evident in some countries. The impacts of false COVID-19 information on trust in government vary worldwide. In African countries where governmental denial, secrecy and misinformation are common, alternative narratives of the COVID-19 crisis flourish online, thereby encouraging public distrust, apprehension and ambivalence towards public health messaging. Political disinformation about rigged elections and exposure to anti-government conspiracy theories also seeds distrust in wider democratic processes, lowers intention to vote and decreases political trust in the USA and UK. More studies across the world that examine the impact of false information on trust in important civic processes and institutions are needed.

Various stakeholders and countries have put forward solutions to counter false information online. It is to these that the next chapter turns, focusing on globally dominant digital platforms as the prime incubator of optimised emotions prevalent in the world today. We follow this with our final chapter that reflects more broadly on emergent forms of emotional AI, outlining the harms that are visible on the horizon line, but also drawing out how there is scope for stakeholders, countries and larger regions to act.

References

Alegre, S. (2017). Opinion. Rethinking freedom of thought for the 21st century. *European Human Rights Law Review, 3*, 221–233. Retrieved April 13, 2022, from https://www.doughtystreet.co.uk/sites/default/files/media/document/Rethinking%20Freedom%20of%20Thought%20for%20the%2021st.pdf

Alegre, S. (2021, May). *Protecting freedom of thought in the digital age.* Policy Brief No. 165. Centre for International Governance Innovation. Retrieved April 13, 2022, from https://www.cigionline.org/publications/protecting-freedom-of-thought-in-the-digital-age/

Andalibi, N., & Buss, J. (2020). *CHI '20: Proceedings of the 2020 CHI conference on human factors in computing systems*, April, pp. 1–16. https://doi.org/10.1145/3313831.3376680.

Anstead, N. (2017). Data-driven campaigning in the 2015 United Kingdom General Election. *The International Journal of Press/Politics, 22*(3), 294–313. https://doi.org/10.1177/1940161217706163

Arguedas, A. R., Robertson, C. T., Fletcher, R., & Nielsen, R. K. (2022). *Echo chambers, filter bubbles, and polarisation: A literature review*. Reuters Institute and the Royal Society. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review

Assardo, L. (2021, August 27). The disinformation backyard. *Medium*. Retrieved April 13, 2022, from https://luisassardo.medium.com/?p=5643ad671bd5

Bail, C. (2021). *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.

Bail, C. A., Guay, B., Maloney, E., Combs, A., Sunshine Hillygus, D., Merhout, F., Feelon, D., & Volfovsky, A. (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences, 117*(1), 243–250. https://doi.org/10.1073/pnas.1906420116

Bakir, V. (2020). Psychological operations in digital political campaigns: Assessing Cambridge Analytica's psychographic profiling and targeting. *Frontiers in Political Communication., 5*, 67. https://doi.org/10.3389/fcomm.2020.00067

Bakir, V., Herring, E., Miller, D., & Robinson, P. (2019). Organized persuasive communication: A conceptual framework. *Critical Sociology, 45*(3), 311–328. https://doi.org/10.1177/0896920518764586

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Benkler, Y., Faris, R., Roberts, H., & Zuckerman, E. (2017). Study: Breitbart-led right-wing media ecosystem altered broader media agenda. *Columbia Journalism Review*. Retrieved April 13, 2022, from https://www.cjr.org/analysis/breitbart-media-trump-harvard-study.php

Berry, J. M., & Sobieraj, S. (2014). *The outrage industry: Political opinion media and the new incivility*. Oxford University Press.

Bessi, A., Scala, A., Rossi, L., Zhang, Q., & Quattrociocchi, W. (2014). The economy of attention in the age of (mis) information. *Journal of Trust Management, 1*(1), 1–13. https://doi.org/10.1186/s40493-014-0012-y

Bessi, A., Petroni, F., Del Vicario, M., Zollo, F., Anagnostopoulos, A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Homophily and polarization in the age of misinformation. *The European Physical Journal Special Topics, 225*, 2047–2059. https://doi.org/10.1140/epjst/e2015-50319-0

Boyd-Barrett, O. (2020). *Russiagate. Disinformation in the age of social media*. Routledge.

Bruns, A. (2019). Filter bubble. *Internet Policy Review, 8*(4), 1–14. https://doi.org/10.14763/2019.4.1426

Bruns, A., & Enli, G. (2018). The Norwegian Twittersphere: Structure and dynamics. *Nordicom Review, 39*(1), 129–148. https://doi.org/10.2478/nor-2018-0006

Bruns, A., Moon, B., Münch, F., & Sadkowsky, T. (2017). The Australian Twittersphere in 2016: Mapping the follower/followee network. *Social Media + Society, 3*(4), 1–15. https://doi.org/10.1177/2056305117748162

Bursztyn, L., Egorov, G. Enikolopov, R., & Petrova, M. (2019). *Social media and xenophobia: Evidence from Russia*. Technical report, National Bureau of Economic Research. Retrieved April 13, 2022, from https://home.uchicago.edu/bursztyn/SocialMediaXenophobia_December2019.pdf

Channel 4 News Investigations Team. (2020, September 28). Revealed: Trump campaign strategy to deter millions of Black Americans from voting in 2016. *Channel 4 News*. https://www.channel4.com/news/revealed-trump-campaign-strategy-to-deter-millions-of-black-americans-from-voting-in-2016

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *PNAS, 118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Clayton, K., Davis, N. T., Nyhan, B., Porter, E., Ryan, T. J., & Wood, T.J. (2020). *Does elite rhetoric undermine democratic norms?* Retrieved April 13, 2022, from https://www.dartmouth.edu/~nyhan/democratic-norms.pdf

Cossard, A., De Francisci Morales, G., Kalimeri, K., Mejova, Y., Paolotti, D., & Starnini, M. (2020). Falling into the echo chamber: The Italian vaccination debate on Twitter. *Proceedings of the international AAAI conference on web and social media, 14*, 130–140. Retrieved April 13, 2022, from https://ojs.aaai.org/index.php/ICWSM/article/view/7285

Deb, A., Luceri, L., Badaway, A., Ferrara, E. (2019). Perils and challenges of social media and election manipulation analysis: The 2018 US midterms. *Companion of the web conference 2019*, pp. 237–247. https://doi.org/10.1145/3308560.3316486

del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarellia, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *PNAS, 113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

Dommett, K., & Power, S. (2020). *Democracy in the dark: Digital campaigning in the 2019 general election and beyond*. Electoral Reform Society. Retrieved April 13, 2022, from https://www.electoral-reform.org.uk/latest-news-and-research/publications/democracy-in-the-dark-digital-campaigning-in-the-2019-general-election-and-beyond/

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *Advances in Political Psychology, 40*(1), 3–35. https://doi.org/10.1111/pops.12568

First Draft. (2019, November 14). UK Election: How political parties are targeting voters on Facebook, Google and Snapchat ads. *First Draft*. Retrieved April 13, 2022, from https://firstdraftnews.org/articles/uk-election-how-political-parties-are-targeting-voters-on-facebook-google-and-snapchat-ads/

Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., & Benkler, Y. (2017, August 16). *Partisanship, propaganda, and disinformation: Online media and the 2016 U.S. presidential election*. Berkman Klein Center for Internet and Society at Harvard University. Retrieved April 13, 2022, from https://cyber.harvard.edu/publications/2017/08/mediacloud

Farkas, J., & Schou, J. (2020). *Post-truth, fake news and democracy: Mapping the politics of falsehood*. Routledge.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly, 80*(1), 298–320. https://doi.org/10.1093/poq/nfw006

Fletcher, R., & Nielsen, R. K. (2018). Are people incidentally exposed to news on social media? A comparative analysis. *New Media and Society, 20*(7), 2450–2468. https://doi.org/10.1177/1461444817724170

Fletcher, R., Robertson, C. T., & Nielsen, R. K. (2021). How many people live in politically partisan online news echo chambers in different countries? *Journal of Quantitative Description: Digital Media, 1*, 1–56. https://doi.org/10.51685/jqd.2021.020

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology, 38*, 127–150. https://doi.org/10.1111/pops.12394

Fridkin, K. L., & Kenney, P. (2008). The dimensions of negative messages. *American Politics Research, 36*, 694–723. https://doi.org/10.1177/1532673X08316448

Gandy, O. H. (2009). *Coming to terms with chance engaging rational discrimination and cumulative disadvantage*. Ashgate.

Geer, J., & Lau, R. (2006). Filling in the blanks: A new method for estimating campaign effects. *British Journal of Political Science, 36*, 269–290. https://doi.org/10.1017/S0007123406000159

Goldenberg, A., & Gross, J. J. (2020). Digital emotion contagion. *Trends in Cognitive Sciences, 24*(2), 316–328. https://doi.org/10.1016/j.tics.2020.01.009

Golebiewski, M., & Boyd, D. (2018, May). Data voids: Where missing data can easily be exploited (pp. 1–8). *Data & Society*. Retrieved June 22, 2022, from https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf

Green, J., & Issenberg, S. (2016, October 27). Inside the Trump bunker, with days to go. *Bloomberg*. https://www.bloomberg.com/news/articles/2016-10-27/inside-the-trump-bunker-with-12-days-to-go

Guess, A., Nyhan, B., & Reifler, J. (2017). *"You're fake news!" Findings from the Poynter media trust survey*. Retrieved April 13, 2022, from https://poyntercdn.blob.core.windows.net/files/PoynterMediaTrustSurvey2017.pdf

Hagey, K., & Horwitz, J. (2021, September 15). Facebook tried to make its platform a healthier place. It got angrier instead. *Wall Street Journal, 16*. https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215?mod=articleinline

Haim, M., Graefe, A., & Brosius, H.-B. (2018). Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital Journalism, 6*(3), 330–343. https://doi.org/10.1080/21670811.2017.1338145

Hanson, F., O'Connor, S., Walker, M., & Courtois, L. (2019). *Hacking democracies: Cataloguing cyber-enabled attacks on elections*, Policy Brief 16. Australian Strategic Policy Institute. Retrieved April 13, 2022, from https://www.aspi.org.au/report/hacking-democracies

Hong, S., & Kim, S. H. (2016). Political polarisation on Twitter: Implications for the use of social media in digital governments. *Government Information Quarterly, 33*(4), 777–782. https://doi.org/10.1016/j.giq.2016.04.007

Horwitz, J., & Seetharaman, D. (2020, May 26). Facebook executives shut down efforts to make the site less divisive. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

Information Commissioners Office. (2018, November 6). *Investigation into the use of data analytics in political campaigns: A report to Parliament*. Retrieved April 13, 2022, from https://ico.org.uk/media/action-weve-taken/2260271/investigation-into-the-use-of-data-analytics-in-political-campaigns-final-20181105.pdf

Information Commissioners Office. (2020, November). *Audits of data protection compliance by UK political parties: Summary report*. Retrieved April 13, 2022, from https://ico.org.uk/media/action-weve-taken/2618567/audits-of-data-protection-compliance-by-uk-political-parties-summary-report.pdf

Kim, M., & Cao, X. (2016). The impact of exposure to media messages promoting government conspiracy theories on distrust in the government: Evidence from a two-stage randomized experiment. *International Journal of Communication, 10*, 38083827. https://ijoc.org/index.php/ijoc/article/view/5127/1740

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The people's choice: How a voter makes up his mind in a presidential campaign*. Columbia University Press.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review, 111*(3), 831–870. https://doi.org/10.1257/aer.20191777

Lu, Y., & Myrick, J. G. (2016). Cross-cutting exposure on Facebook and political participation: Unravelling the effects of emotional responses and online incivility. *Journal of Media Psychology: Theories, Methods, and Applications, 28*(3), 100–110. https://doi.org/10.1027/1864-1105/a000203

Lytvynenko, J., & Silverman, C. (2020, November 3). Here's a running list of false and misleading information about the election. *Buzzfeed News*. https://www.buzzfeednews.com/article/janelytvynenko/election-rumors-debunked?bfsource=relatedmanual

Mac, R., Silverman, C., & Lytvynenko, J. (2021, April 26). Facebook stopped employees from reading an internal report about its role in the insurrection. You can read it here. *Buzzfeed News*. https://www.buzzfeednews.com/article/ryanmac/full-facebook-stop-the-steal-internal-report

Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online*. Data & Society Research Institute. http://www.chinhnghia.com/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

McCarthy-Jones, S. (2019). The autonomous mind: The right to freedom of thought in the twenty-first century. *Frontiers in Artificial Intelligence, 2*(19), 1–17. https://doi.org/10.3389/frai.2019.00019

McStay, A. (2018). *Emotional AI: The rise of empathic media*. Sage.

Messing, S., & Westwood, S. J. (2012). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research, 41*, 1042–1063. https://doi.org/10.1177/0093650212466406

Microsoft. (2021). *Microsoft digital civility index*. Retrieved April 13, 2022, from https://www.microsoft.com/en-us/online-safety/digital-civility

Milani, E., Weitkamp, E., & Webb, P. (2020). The visual vaccine debate on twitter: A social network analysis. *Media and Communication, 8*(2), 364–375. https://doi.org/10.17645/mac.v8i2.2847

Müller, K., & Schwarz, C. (2020). *Fanning the flames of hate: Social media and hate crime*. Retrieved April 13, 2022, from https://ssrn.com/abstract=3082972.

Mutz, D. C. (2007). Effects of 'In-Your-Face' television discourse on perceptions of a legitimate opposition. *American Political Science Review, 101*(4), 621–635. https://doi.org/10.1017/S000305540707044X

Nechushtai, E., & Lewis, S. C. (2019). What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior, 90*, 298–307. https://doi.org/10.1016/j.chb.2018.07.043

Newman, N. (2022). United Kingdom. In N. Newman, R. Fletcher, C. T. Robertson, K. Eddy, & R. K. Nielsen (Eds.), *Reuters Institute digital*

news report 2022 (pp. 62–63). Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Newman, N., Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Reuters Institute digital news report 2020*. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Newman, N., Fletcher, R., Schulz, A., Andı, S., Robertson, C. T., & Nielsen, R. K. (2021). *Reuters Institute digital news report 2021*. Retrieved April 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf

Nguyen, H., & Nguyen, A. (2020). Covid-19 misinformation and the social (media) amplification of risk: A Vietnamese perspective. *Media and Communication, 8, 2*, 444–447. https://doi.org/10.17645/mac.v8i2.3227

Ogola, G. (2020). Africa and the Covid-19 information framing crisis. *Media and Communication, 8*(2), 440–443. https://doi.org/10.17645/mac.v8i2.3223

Ott, B. L. (2017). The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication, 34*, 59–68. https://doi.org/10.1080/15295036.2016.1266686

Otto, L. P., Lecheler, S., & Schuck, A. R. T. (2019). Is context the key? The (non-) differential effects of mediated incivility in three European countries. *Political Communication, 37*(1), 88–107. https://doi.org/10.1080/10584609.2019.1663324

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

Pelley, S. (2021, October 4). Whistleblower: Facebook is misleading the public on progress against hate speech, violence, misinformation. *60 Minutes*. https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-misinformation-public-60-minutes-2021-10-03/

Pennycook, G., & Rand, D. G. (2021). Research note: Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. *The Harvard Kennedy School Misinformation Review, 2*(1). https://doi.org/10.37016/mr-2020-51

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General, 147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Sunstein, C. R. (2002). The law of group polarization. *The Journal of Political Philosophy, 10*(2), 175–195. https://doi.org/10.1111/1467-9760.00148

Sunstein, C. R. (2017). *# Republic: Divided democracy in the age of social media*. Princeton University Press.

Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online manipulation: Hidden influences in a digital world. *Georgetown Law Technology Review, 1*, 1–45. https://doi.org/10.2139/ssrn.3306006

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50*(3), 755–769. https://www.jstor.org/stable/3694247.

Thakur, D., & Hankerson, D. L. (2021). *Facts and their discontents: A research agenda for online disinformation, race, and gender*. Center for Democracy & Technology. https://osf.io/3e8s5/

Timberg, C., Dwoskin, E., & Albergotti, R. (2021, October 22). Inside Facebook. Jan. 6 violence fueled anger, regret over missed warning signs. *Washington Post*. https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/

Tsfati, Y., Boomgaarden, H. G., Strömbäck, J., Vliegenthart, R., Damstra, A., & Lindgren, E. (2020). Causes and consequences of mainstream media dissemination of fake news: Literature review and synthesis. *Annals of the International Communication Association, 44*(2), 157–173. https://doi.org/10.1080/23808985.2020.1759443

Tufekci, Z. (2014). Engineering the public: Big data, surveillance and computational politics. *First Monday*, 19(7). Retrieved April 13, 2022, from https://firstmonday.org/ojs/index.php/fm/article/view/4901/4097

Vaccari, C., & Valeriani, A. (2021). *Outside the bubble: Social media and political participation in western democracies*. Oxford University Press.

Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C., & Sternisko, A. (2021). How social media shapes polarization. *Science & Society, 25*(11), 913–916. https://doi.org/10.1016/j.tics.2021.07.013

Van Duyn, E., & Collier, J. (2019). Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication and Society, 22*(1), 29–48. https://doi.org/10.1080/15205436.2018.1511807

Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society, 20*(5), 2028–2049. https://doi.org/10.1177/1461444817712086

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication, 37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Wang, D., & Qian, Y. (2021). Echo chamber effect in rumor rebuttal discussions about COVID-19 in China: Social media content and network analysis study. *Journal of Medical Internet Research, 23*(3), e27009. https://doi.org/10.2196/27009

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*(3), 129–140. https://doi.org/10.1080/17470216008416717

Weeks, B. E. (2015). Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of Communication, 65*, 699–719. https://doi.org/10.1111/jcom.12164

Zollo, F., Bessi, A., Del Vicario, M., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., & Quattrociocchi, W. (2017). Debunking in a world of tribes. *PLoS One, 12*(7), e0181821. https://doi.org/10.1371/journal.pone.0181821

CHAPTER 8

Defending the Civic Body from False Information Online

Introduction

We have established that false information online harms the *civic body*, driven by the *economics of emotion* and the *politics of emotion*. What should be done about this? Global and regional surveys conducted in 2018 indicate public appetite for interventions to stop 'fake news' but are unclear where primary responsibility lies (Eurobarometer, 2018, February; Newman et al., 2018). Accordingly, multi-stakeholder solutions have been proffered by various countries' governmental inquiries into disinformation and fake news, and by supranational bodies including the United Nations (UN), European Union and Commonwealth. This chapter assesses seven solution areas: namely, (1) government action, (2) cybersecurity, (3) digital intermediaries/platforms, (4) advertisers, (5) professional political persuaders and public relations, (6) media organisations and (7) education. These are intrinsically difficult areas to solve individually, let alone in concert, and in every country. We conclude that such solutions merely tinker at the edges as they do not address a fundamental incubator for false information online: namely, the business model for social media platforms built on the *economics of emotion*.

V. Bakir, A. McStay, *Optimising Emotions, Incubating Falsehoods*,
https://doi.org/10.1007/978-3-031-13551-4_8

Solution Area 1: Governmental Action

Across the world, governmental approaches to tackling false information online range from those that respect freedom of expression (non-coercive responses) to those that do not (coercive responses).

Non-coercive Responses

In seeking to prevent negative democratic impacts of false information online, supranational reports recognise the importance of balancing measures to combat false information with the right to freedom of expression. For instance, on 3 March 2017, a Joint Declaration on 'Freedom of Expression and 'Fake News', Disinformation and Propaganda' was adopted by the United Nations Special Rapporteur on Freedom of Opinion and Expression, alongside other organisations. While noting the growing prevalence of disinformation, the Joint Declaration reaffirms the right to freedom of expression (Principle 1) and stipulates standards on disinformation and propaganda (Principle 2), including that states should not ban 'dissemination of information based on vague and ambiguous ideas, including 'false news' or 'non-objective information' (United Nations Special Rapporteur on Freedom of Opinion and Expression et al., 2017).

Instead, as will be developed in later sections, non-coercive governmental responses include media monitoring and development of early warning detection systems as part of cybersecurity operations against disinformation and as part of digital literacy programmes. Control over diffusion of false information is difficult in today's media ecology, but there is some evidence that rumours on social media can be stopped by early, strong corrections by officials, as seen in studies in Japan (Takayasu et al., 2015) and Germany (Jung et al., 2020).

Non-coercive governmental responses also include a preference for self-regulation rather than regulation of digital platforms and intermediaries. Significantly, the European Commission created the European Union Code of Practice on Disinformation in 2018, where for the first time worldwide industry agreed on a voluntary basis to self-regulatory standards to fight disinformation (European Commission, 2018b). Similar Codes of Practice have since been developed in other jurisdictions, such as Australia (Digital Industry Group Inc., 2021). The European Union Code of Practice was signed by Facebook, Google, Twitter, Mozilla and

parts of the advertising industry in 2018, with Microsoft and TikTok signing in 2020, and further signatories in 2021 as the Code of Practice was strengthened (European Commission, 2021c). While the Code of Practice ensured greater transparency and accountability of signatories' disinformation policies, the European Commission concluded that more needed to be done in consistently applying and monitoring the Code across platforms and Member States, in providing access to platforms' data for disinformation research and in increasing participation from the advertising sector (European Commission, 2021b, c). As we shall see below, this has since resulted in more coercive legislation in the European Union.

Coercive Responses

Given the failure of self-regulation to prevent widespread circulation of false information, many nations have resorted to more coercive responses. Arrests and shutting down the entire Internet are arguably the most coercive actions taken by governments. In the year across June 2020 to May 2021, Freedom House (2021) (a non-profit, majority US government-funded organisation that conducts research and advocacy on democracy, political freedom and human rights) observed that of the 70 states covered by its annual study of human rights in the digital sphere, officials arrested or convicted people for their online speech in 56 countries; and governments suspended Internet access in at least 20 countries, usually during political turmoil across elections and protests.

Other actions which do not respect freedom of expression include enacting legislation on false information online. Authoritarian China, keen to maintain control over its diverse population of 1.4 billion, was among the first to legislate in this area. In 2016, China criminalised creating or spreading online 'rumours' that 'undermine economic and social order'. A 2017 law requires Internet news providers to reprint information published by government-acknowledged news organisations without 'distorting or falsifying news information' (Repnikova, 2018, September 6). In 2018, Chinese authorities required microblogging sites to highlight and refute rumours on their platforms. Across 2020–2021, China's Internet regulator introduced new rules to restrict independently operated social media accounts that publish current affairs, leading to many accounts being removed (Freedom House, 2021). China's State Council published guidelines for building a 'civilised' Internet in September 2021, stating that the web should be used to promote education about the ruling

Communist Party and its achievements. Beyond China, by 2019, over 40 national laws to combat disinformation had been chronicled worldwide (Marsden et al., 2020). Some are intended to eliminate critical reporting. For instance, as reported by international non-governmental human rights organisation Amnesty International (2022, March 10), during Russia's invasion of Ukraine in 2022, Russia's parliament criminalised spreading 'false information' about Russian Armed Forces or 'discrediting' Russian troops (punishable by 15 years in prison). In other countries, such laws are intended to pressurise social media platforms to take action. For instance, in Germany, the 'Netzwerkdurchsetzungsgesetz' (Network Enforcement Act) (NetzDG) was introduced in 2018 to reduce the spread of hate speech and false information (Netzwerkdurchsetzungsgesetz, 2017). Online platforms must remove 'obviously illegal' posts within 24 hours or risk fines of up to €50 million. While well-intentioned, as observed by Human Rights Watch (an international non-governmental organisation that conducts research and advocacy on human rights), this law damages free speech by tasking companies that host third-party content to make difficult determinations of when user speech violates law. Even courts can find these determinations challenging, as they require nuanced understanding of context, culture and law. Faced with short review periods and steep fines, companies have little incentive to err in favour of free expression (Human Rights Watch, 2018, February 18).

Elsewhere, rather than legislating on false information online generally, legislation seeks to protect elections while trying to respect freedom of expression. For instance, in 2018, France passed a law that establishes an expedited judicial procedure for adjudicating complaints about fake news preceding elections, imposing heightened transparency obligations on platforms during these periods (Fukuyama & Grotto, 2020, p. 204). The USA, in 2019, approved proposals for online paid political ads to be required to be appropriately labelled and to clearly display or link to key information (McNeice, 2019, November 5).

Appreciating that self-regulation has not sufficiently tackled false information online, in April 2022 the European Union agreed the broad terms of the Digital Services Act to make technology companies take greater responsibility for content appearing on their platforms. Expected to come fully into force by 2024 at the latest, new obligations on platforms include new strategies for dealing with misinformation during crises (such as a pandemic or war); explaining clearly why they have removed illegal content; giving users the ability to appeal takedowns; explaining how their

recommender algorithms work; offering a recommender system not based on profiling (for instance, chronological listing); prohibiting 'dark patterns' (namely, confusing or deceptive user interfaces designed to steer people into decisions they may not otherwise have made); banning targeted ads based on an individual's religion, sexual orientation, ethnicity, health information or political beliefs or targeted at children; allowing European Union governments to request removal of illegal content; and dissuasive sanctions of up to 6% of global turnover. The online platforms will also have to identify and tackle 'systemic risks' stemming from the design and use of their services including those that adversely impact fundamental rights or seriously harm users' physical or mental health, and manipulation of services that impact democratic processes and public security (Council of the EU, 2022, April 23; Goujard, 2022, April 23; Vincent, 2022a, April 23).

Also noteworthy is that the nurturing policy environment evident in many countries across recent decades that encouraged big technology platforms to innovate and grow now appears to be shifting against monopoly power. For instance, alongside the Digital Services Act, the European Union is advancing the Digital Markets Act. Its broad details, agreed in March 2022, aim to curb the dominant big technology platforms and enable future anti-trust actions. Its proposed penalties for infringement include fines of up to 10% of total worldwide turnover in the preceding financial year and 20% for repeated infringements, and a time-limited ban on acquiring other companies in the case of systematic infringements (Vincent, 2022b, March 24). Whether this policy shift against monopolistic big technology platforms will endure remains to be seen, not least as the technology sector intensively lobbies parliaments to water down proposed legislation. Indeed, a study from Corporate Europe Observatory and Lobby Control (an independent research and campaign group working to challenge the privileged influence of corporations and their lobby groups in European Union policy-making) finds that the technology sector is the biggest lobby sector in Europe (Bank et al., 2021). Across 2021, the biggest spenders lobbying the European Union were Apple (€6.5 million), Google (€6 million) and Facebook (€6 million). A major target was to protect their surveillance advertising business model from an outright ban (Lomas, 2022, April 22). Beyond the European Union, in 2021 China, too, ended its stance of minimal regulation of its own big technology companies (Au, 2021, September 27). For instance, wishing to curb capitalist excess, and increase national security, China's Central Commission

for Cybersecurity and Informatization (2021, *December 28*) issued its 14th Five-Year Plan for National Informatisation. Its plans to build technology norms and digital governance systems include reducing its technology industry's 'disorderly expansion' and monopolistic business practices; launching 'technical algorithm regulation'; and clarifying the responsibility that Internet platforms bear over the content they publish.

Solution Area 2: Cybersecurity

The spread of false information online, especially through information warfare conducted via social media platforms, is a significant cybersecurity issue. Information warfare includes coordinated, deceptive efforts to manipulate public debate, often spreading hate speech or populism; trying to undermine faith in democracy; or trying to manipulate electorates through negative campaigning, fear and divisions.

In December 2020, the European Commission recognised that more effort was needed on cybersecurity to strengthen European democracies. It notes that only by pooling existing knowledge on hybrid threats across different sectors (such as disinformation, cyber operations and election interference) can the European Union respond effectively to disinformation and influence operations (European Commission, 2020, December 3, p. 20). While the European Union and selected countries are addressing cybersecurity and social media platforms, the response is more uneven worldwide (Brown et al., 2020). For democratic governments, responding to foreign interference can be difficult as methods used by adversaries typically exploit democratic principles such as free speech, trust and openness. Detection can be hard both because the methods are difficult to identify and because democracies pertain to avoid surveillance of their own domestic populations and debates, with most intelligence resources directed towards external collection to actively monitor foreign disinformation campaigns (Bakir, 2019 [2018]; Hanson et al., 2019).

The digital platforms have also adopted cybersecurity measures. For instance, since Russian attempts to influence the 2016 US presidential election were exposed, Facebook has since built a team of over 200 people globally (experts in cybersecurity, disinformation, digital forensics, law enforcement, national security and investigative journalism) focused on combating such operations. Its approach detects and removes violating content, known bad actors and coordinated deceptive behaviour. It is designed to have flexibility, understanding that tactics evolve as bad actors

take evasive actions (Facebook, 2020a, April). Indeed, rapid technological change and adaptive tactics by disinformation purveyors have spurred platforms, news outlets and researchers to find automated ways of detecting deceptive forms such as fake news online and deepfakes.

In terms of research on automatically detecting fake news online to help fact-checkers, existing approaches mainly rely on training classifiers, for which past events or claims are gathered and labelled as real or fake, and significant features are extracted to generate appropriate data representations (Cha et al., 2020). However, problems abound with using AI for fake news detection. Firstly, unlike detection of hateful, sexist, or hyperpartisan language, linguistic classifiers alone cannot detect fake news and propagandists exploit such weaknesses. For instance, Russia's Facebook ads used to try to disrupt the 2016 US presidential election and typically posted words superimposed on images, which allowed them to evade Facebook's machine-learning algorithms for detecting fake news (Levy, 2020, p. 375). A second problem with using AI for fake news detection is that linguistic classifiers need humans in the loop, such as fact-checkers, to keep the models updated, otherwise accuracy rapidly degrades, even within one week. Thirdly, building blacklists of websites spreading false information is not scalable for content produced every minute and will produce bias towards specific websites in the database. Fourthly, removal of fake accounts is problematic because of the vast scale at which fake accounts are produced. Fifthly, stakeholders require models to combat fake news that provide explainable outcomes that highlight which users and publishers are creating fake news, on which topics and through what types of textual and social manners, but automated AI solutions do not lend themselves to explainability (Cha et al., 2020; Ghulati, 2020, 27 November). Nonetheless, the field continues to advance. For instance, to learn feature representations from multiple aspects, deep neural networks have been successfully applied to tasks such as visual question answering; image captioning; and a deep learning-based fake news detection model which extracts multimodal and social context features and fuses them by attention mechanism (Wang et al., 2018).

Alongside detecting fake news through automated means, various countermeasures against burgeoning deepfake technology have been developed in collaborations between the US military and dominant platforms. These have provided tools and datasets of manipulated and non-manipulated videos to help develop identification techniques (Vizoso et al., 2021). They include the US Defense Advanced Research Projects

Agency establishing the Media Forensics programme (Langguth et al., 2021; Vizoso et al., 2021) and a competitive challenge organised by Facebook's Deepfake Detection Challenge across 2019–2020, boosted by companies like Microsoft and Amazon Web Services and university research (Facebook, 2020b, June 25). Many tools have been created to automatically detect deepfakes, based on intrinsic contradictions in the algorithm synthesis. These include a lack of eye blinking or mismatching lip movement with speech. There are systems that use a convolutional neural network that extracts frame-level features that are then used to train a recurrent neural network that learns to determine if a video has been manipulated. Google also created a tool called Assemble that helps journalists identify manipulated images (Pérez Dasilva et al., 2021). However, Langguth et al. (2021) warn that the success of such approaches depends ultimately on their mode of deployment. Furthermore, recent research using adversarial strategies indicates that even the best detectors can be fooled. Adversarial strategies consist of adding noise (imperceptible to the human eye) to a video or image to confuse a fake news detector. They conclude that it is likely that many of these systems are ultimately flawed in application because they do not offer 100% detection accuracy, and if they are available to the public, they will also be available to disinformation creators.

Solution Area 3: Digital Platforms/Intermediaries

Under pressure from regulators and bad publicity arising from whistleblowers and political inquiries, globally dominant digital platforms have undertaken design reforms and algorithmic tweaks to address some of the harms arising from their existence while preserving their core business model of maximising user engagement (see Chap. 2). As noted earlier in this chapter, the European Union Code of Practice on Disinformation regarded as a landmark document and signed by dominant digital media platforms sets out voluntary commitments including those on better platform transparency, digital literacy and content moderation (European Commission, 2018b, 2021b). Such voluntary commitments were not all successfully fulfilled (European Commission, 2021c) and have since been hardened into the landmark Digital Services Act. Later in this chapter, we address some of these commitments and their shortcomings, but here we focus on the thorny issue of content moderation.

In efforts to promote rapid growth of Internet platforms, US federal legislation passed in 1996 (Section 230 of the Communications Decency Act) freed Internet intermediaries from almost all liability for user-generated content, placing the burden of content curation on the platforms themselves (see Chap. 2). In the USA, where dominant digital platforms are based, freedom of speech (especially political or ideological speech) is a constitutional right, with exceptions for narrow speech categories of obscenity, defamation, fraud, incitement, fighting words, true threats, speech integral to criminal conduct and child pornography (Killion, 2019, January 16). It is only in exceptional cases that platforms censor politicians. For instance, in an unprecedented move, in January 2021 Facebook banned then outgoing US President Trump until at least 2023 for inciting the deadly January 6 insurrection at the US Capitol (Hendrix, 2021, January 7).

While political speech is protected, digital platforms are more forthcoming in banning deliberately misattributed or manipulated content (with exceptions for satire). For instance, YouTube's (2021) misinformation policies prohibit misattributed content, namely, content 'that may pose a serious risk of egregious harm by falsely claiming that old footage from a past event is from a current event'. It also bans manipulated content, namely, content that is 'technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm'. Its examples of manipulated content include videos that are technically manipulated to make it appear that a government official is dead or to fabricate events where there is serious risk of egregious harm. Similarly, TikTok (2021) prohibits 'digital forgeries (Synthetic Media or Manipulated Media) that mislead users by distorting the truth of events and cause harm to the subject of the video, other persons, or society'.

While platforms ban certain types of content to prevent harms to the *civic body*, they prefer to promote authoritative sources and demote borderline content. Most commonly, such content moderation occurs in areas of health messaging and elections. For instance, harms to public health from false COVID-19 information prompted platforms to moderate content via their recommendation algorithms. Google prioritises information from the World Health Organization in its search rankings, even if this information is not optimised for Google (O'Donovan, 2020, November 27). According to a civil rights audit of US Facebook, Facebook shows messages in News Feed to people who interacted with harmful COVID-19

misinformation that was later removed as false, using these messages to connect people to the World Health Organization's COVID-19 myth-buster website (Murphy, 2020, July 8, p. 53). On content moderation during elections, in June 2020, US Facebook announced its planned Voting Information Center, modelled after the COVID-19 Information Center that Facebook uses to connect users to trusted information from health authorities.

Despite a plethora of ever-evolving policies and community guidelines on false media forms and content, digital platforms perform poorly at enforcing intermediary liability laws (which tell platforms what responsibility they have for unlawful content posted by users) consistently at scale. One reason may be divergence in national laws on (a) how neutral platforms must be to qualify for immunity from legal claims arising from users' unlawful speech and (b) the degree of content moderation they can perform without being exposed to liability. Another reason is that most intermediary liability laws oblige platforms to take down illegal content once they 'know' about it, but laws vary in what counts as 'knowledge'. Under some national rules, platforms can only be legally required to take down users' speech if a court adjudicates it unlawful; elsewhere, platforms can decide themselves (Keller & Leerssen, 2020). Because platforms' private rulesets (Community Guidelines) are privately defined and enforced, platforms' decisions are generally not subject to review by courts, there is little transparency in how their policies are applied, and these appear inconsistent internationally (Ajder & Glick, 2021; Keller & Leerssen, 2020; McNamee, 2019). Against further transparency, Facebook states that it fears giving people with bad intentions a playbook to explain its algorithms (Merrill & Oremus, 2021, October 26). If digital platforms were more explicit about their algorithms' workings, this could also give competitors an easy means of duplicating and surpassing their service (Gillespie, 2014). At the time of writing (April 2022), the European Union Digital Services Act (referred to earlier in this chapter), which will make platforms explain their content moderation policies, practices and decisions more clearly including how their recommender algorithms work, looks promising on paper, but time will tell whether sufficient resources have been ringfenced to ensure compliance or whether lobbying will dilute the law and provide platforms with workarounds that mean that they do not have to significantly alter their behaviour. Certainly, in the run up to the agreement of this act, a major lobbying target for Google and Spotify (the world's largest music streaming service provider) was to limit researchers'

access to data on algorithmic content ranking systems (Lomas, 2022, April 22).

A further problem with content moderation of online disinformation campaigns is that this requires action from dominant and minor platforms alike to prevent those censored on one platform from simply moving onto other platforms (Siegel, 2020, p. 73). For instance, according to technology journalist Sarah Emerson, in response to dominant social media platforms' efforts to quell Trump's claims of election fraud across the 2020 US presidential election campaign, movements such as Stop the Steal, QAnon and right-wing militia groups moved to platforms such as MeWe, where they encouraged violent responses to post-election events (Emerson, 2021, January 14). Very unusually, following the US Capitol Hill riot in January 2021, censorship was enforced across the entire platform ecosystem in the following fortnight, including YouTube, Facebook, Instagram, Twitter, Google, TikTok, Amazon, Apple and Airbnb, but also lesser-known platforms such as Gab, Parler, 4chan, Stripe, Twitch, Zello and MeWe.

The US Capitol Hill riot also highlights two intrinsic technical difficulties of content moderation on any platform. Firstly, it requires skilled detective work to understand the nature of posts (Emerson, 2021, January 14). As Zello, which is often encrypted end to end for privacy reasons, points out: 'This makes the task of proactive monitoring for compliance with our terms of service unrealistic: we simply cannot just "search our data" for specific keywords in conversations' (Zello Staff, 2021, January 13). Secondly (and related), content moderation requires resources, and there is a divergence between the capabilities of dominant platforms and smaller digital intermediaries. As of 2021, MeWe employed less than 100 moderators, while Facebook employed 15,000 people reviewing content in over 70 languages. However, even in Facebook, this resource is unevenly distributed worldwide (see Chap. 2), relies heavily on automation geared towards English-language communities and, to date, has often fallen short of what is needed (Simonite, 2021, October 25, Thakur & Hankerson, 2021).

Solution Area 4: Advertising

There are various advertising-driven causes of, and solutions to, false information online. As Chap. 2 observes, the very successful Google-Facebook duopoly in online behavioural advertising means that news

outlets are deprived of advertising funds, with the resulting news desert ultimately, perhaps, the biggest challenge to tackling false information online: we address how to combat this in the section below on Media Organisations. In Chap. 2, we also discussed how datafied emotional content is optimised to generate Facebook shares for Internet traffic and advertising income (clickbait audiences)-generating fake news, hate speech and deceptive, emotive political campaigning: we address how to combat this in the coming section on Professional Political Persuaders and Public Relations. In this section, we focus specifically on the problem of commercial ads online inadvertently funding false information via adtech.

Adtech is used to profile and target people in order to serve behaviourally targeted ads. It funds fake news sites, commercial ads and political ads alike. Here, the prime actor is Google's ad network, *Doubleclick* but there are other behavioural and programmatic ad networks including seemingly countless lesser-known networks such as *OpenX*, *Tribal Fusion* and *33Across.* By March 2018, the European Commission (2018a) reported that online platforms were tackling disinformation by disrupting the business model for its production and amplification. Disruptions included ad networks not placing ads on websites identified as purveyors of disinformation, thereby directly reducing income to disinformation providers, and ad networks not disbursing revenues to sites and partners until they could confirm that they operate within relevant terms and conditions. However, by late 2020 Konrad Shek (Deputy Director, Policy and Regulation, UK Advertising Standards Association) observed that although brands are incentivised to choke funds to fake news websites, the volume and speed of the supply chain makes this difficult: for instance, brands already employ negative lists, but must keep these updated (Shek, 2020, November 27).

The issue of brand safety is an ongoing one within the digital advertising industry, and the issue of false information online adds political and public impetus to resolve it: reputable advertisers are unlikely to want their advertising associated with content that by its very nature cannot be trusted. Various efforts have been made to help advertisers identify (and avoid) false information providers online. For instance, British-based nonprofit organisation, Global Disinformation Index, deploys its assessment framework to rate news domains' risk of disinforming their readers, aiming to generate neutral ratings for advertisers, ad tech companies and platforms to redirect their online ad spending, in line with their brand safety and disinformation risk mitigation strategies (Global Disinformation

Index, 2021). Other initiatives from various ad networks and programmatic companies promise to deliver brand-safe ads. Rubicon, for example, claims it can identify undesirable publishers before the ads are released and can track activity during and after the campaign to see who clicked on which ads and where. However, to be effective, all ad networks need to be involved to prevent undesirable sites (such as fake news sites) that have been ejected from one ad network from simply moving to less discriminating ones. With greater transparency in the system for advertisers, non-fake news publishers and advertisers could be encouraged to stop using the less discriminating ad network. Given that ad networks benefit from economies of scale, the departure of reputable advertisers and publishers would be harmful and possibly terminal to that ad network. Indeed, a study that tracked ads served in a sample of fake, low-quality and traditional news outlets over 12 weeks in in 2019 (1.32 million ads served by 565 unique ad servers on 1600 news sites) finds that fake news publishers were still strongly reliant on credible ad servers: the top ten credible ad servers alone accounted for 67% and 56% of fake and low-quality ad traffic, respectively (Bozarth & Budak, 2021).

As the European Union General Data Protection Regulation (GDPR) has taken effect in media markets, the end of the cookie-based behavioural advertising market is increasingly likely, especially as Google and Apple tighten control of third-party cookies for Chrome and Safari users to prevent unwanted tracking. As of 2022, for example, the default for cookies will be 'off' in Chrome. What will replace the third-party tracking cookie is not yet clear but is likely to consist of new approaches to identifying a user and targeting by much larger cohorts rather than individual profiles. Another approach is Universal IDs, which are based on a person providing their personal details to advertisers, such as logins to sites and details held about their interactions with sites. Similarly, third-party identity management services also exist, which would allow for microtargeting and cross-site tracking (just as third-party cookies do today). With adtech's trade association, the Internet Advertising Bureau, noting that 'universal ID solutions work very similarly to third-party cookies', it remains it be seen how viable this is (Internet Advertising Bureau UK, 2021). Notably, Google will not support Universal IDs, effectively locking out smaller adtech firms. Both Apple and Google prefer topic- and cohort-based approaches, which are built on larger clusters of people. Google's 'Topics' approach, for example, targets by cohorts of people (potentially of thousands of people) (Goel, 2022). This, then, would involve what we see as

meso-targeting, the middle layer between micro- and macro-. Here, input features to the ad network algorithm, including web history, are kept local on the browser and are not uploaded elsewhere—the browser only exposes the generated topics for that week to the ad network. Yet, notably, if a publisher has subscriber details, they will have access to the cohort a person belongs to. Although Google's GitHub pages note restrictions on sensitive categories, Google's policy on political content is based on regional legal compliance, whereas other categories are explicitly restricted (such as targeting by or in relation to personal hardship, systemic discrimination, sexual interests, or societal biases). There is then the contextual approach (targeting based on content rather than who is looking at the content), which is certainly a more privacy-friendly approach, but it remains to be seen whether this addresses the problem of the over-emotionalised *civic body*, especially as the digital version of contextual advertising seeks to profile sentiment of the content on the site itself (such as keywords, website content and other metadata), in turn showing ads in relation to what else is displayed on the site at the time. Conceivably, one could see that publishers would work to clarify the emotional tone of their sites, to ensure brand-emotion uniformity and that programmatic advertising is in line with this. Yet, this could feed further news and audience polarisation, as publishers avoid being caught in a 'balanced' middle ground, which would be of less value to advertisers due to absence of clarity of which audience is being reached.

Solution Area 5: Professional Political Persuaders and Public Relations

As discussed in Chap. 1, an international survey of digital news consumption across 40 countries finds that it is domestic politicians that are regarded as by far the most responsible for false and misleading information online (Newman et al., 2020, p. 18). As such, this section focuses on professional persuaders and public relations in the political domain, addressing two problematic areas in incubating false information: the use of political online ads and broader use of strategic communications.

Political Online Ads

Across the world, electoral laws greatly lag developments in the digital media ecology. Regulating online political ads is challenging due to the borderlessness of online space; the difficulty of recognising seemingly organic, but paid-for, political material and distinguishing it from other political content; and microtargeting and behavioural profiling techniques that can rely on improperly obtained data, which in turn may be misused to direct polarising narratives (European Commission, 2020, December 3, p. 4). Stakeholders looking for solutions are divided on the value of microtargeting but are more united on increasing the transparency of online political ads, so enabling advertisers to be held accountable for what they say and for breaking rules. Of interest, given its global focus, are recommendations from the Kofi Annan Commission on Elections and Democracy in the Digital Age (2020) discussed below.

The Kofi Annan Commission urges countries to adapt their political advertising regulations to the online environment and recommends that relevant public authorities should *define in law what is considered to be a political ad*. Such a move would enable digital intermediaries and platforms to know what to include in their own policies on ads about elections and politics. The European Commission's (2021a) proposed rules on political ads, published in November 2021, took a broad definitional approach to political ads, to include those concerning political actors and issue-based ads liable to influence voting behaviour. Of course, digital political campaigning is far broader than simply paid-for advertising and may include branded content, influencers and other activities that look like ads.

The Kofi Annan Commission recommends that countries should *compel social media platforms to make public all information involved in the purchase of an ad*, including the advertiser's real identity, the amount spent, targeting criteria and actual ad creative. Since the 2016 US presidential elections, some social media platforms have introduced measures to verify the identity of people purchasing political ads. Facebook, for instance, requires those running ads about 'social issues, elections or politics' to have their identity verified using documents issued by the country they want to run ads in (Facebook, n.d.), although this is not active in every country (Facebook, 2021). In 2021, the European Commission (2021a, November 25) proposed transparency rules that would require political ads and electorally relevant issue ads to be clearly labelled,

including information such as who paid for it and how much. In the USA, proposed legislation that would have created an archive maintained by the Federal Election Commission of purchased political ads online prompted Twitter, Google and Facebook to provide publicly accessible, searchable libraries of election ads and spending on their US platforms in 2018, with rollouts in certain other countries since then. Although by October 2019, Twitter stopped accepting most political ads, Google allows users to see election-related ads, showing statistics on audience demographics. On Facebook, users can click on political, electoral or social issue ads to access information about the ad's reach, who was shown the ad and the entity responsible; and Facebook took more measures to increase the transparency of Political Ads and its Public Ad Library following criticisms of its explainability and functionality (Edelson et al., 2018; Murphy, 2020, July, pp. 836–837). While such political ad archives have enabled journalists to call attention to influence networks and monitor ad content for disinformation and hate speech, they remain minimally useful for electoral regulators (Gorwa & Ash, 2020; Leerson et al., 2021). Meaningful political ad archives need to archive ads accurately, rapidly (ideally, in real time), over long time periods (ideally, all), provide granular information about spending and targeting, and provide precise names of organisations that paid for the ads (Dommett & Power, 2020; Leerson et al., 2021). Such archives are needed in every country where digital platforms allow political advertising. Leerson et al. (2021) argue that these should be publicly regulated, otherwise journalists are reliant on voluntary, incomplete access frameworks controlled by the very platforms they aim to scrutinise.

The Kofi Annan Commission recommends that countries should *specify by law the minimum audience segment size for an ad*. Since 2017, digital platforms started to limit the level of detail campaigns could use to target voters. In November 2019, Google said that while it had 'never offered granular microtargeting of election ads', it was further limiting election ad targeting to general categories of age, gender and general location (postal code level), as well as to contextual targeting (Spencer, 2019); that advertisers would no longer be able to target political messages based on users' interests inferred from their browsing or search histories (Glazer, 2019, November 21); and that this approach would be enforced worldwide from 2020. Reviewing Google's policy in 2022, this is not globally uniform as Google has different requirements for political and election ads based on region (Google, 2022). The policy of disclosure requirements (that an ad is political) and targeting restrictions (low granularity) are only applied to

regions where election ad verification is required. (According to Google, disclosure and restrictions apply in Australia, Brazil, European Union, India, Israel, New Zealand, Taiwan, the UK and the USA.)

Facebook continues to microtarget, arguing that advertising is an important part of free speech, especially when it comes to political messaging. However, given increased legislative scrutiny of these practices, Facebook's parent company (Meta) announced that from January 2022, it would no longer let advertisers target people based on how interested the social network thinks they are in 'sensitive' topics including political affiliation, religion, sexual orientation, health, race and ethnicity. This would apply across Meta's apps, including Facebook, Instagram and Messenger, and its audience network, which places ads on other smartphone apps (Bond, 2021, November 9). This is in line with the European Commission's (2021a) proposed rules on political ads and electorally relevant issue ads, published in November 2021, that stipulate that targeting and amplification would be banned when using sensitive personal data (such as ethnic origin, religious beliefs or sexual orientation) without explicit consent of the individual. The proposed rules also stipulate that political targeting and amplification techniques would need to be explained publicly in unprecedented detail including clear information on what basis the person is targeted, which groups of people were targeted, based on which criteria and with what amplification tools or method.

Strategic Communications

As Chap. 6 shows, it is not just paid for political advertising that promotes harmful disinformation. Rather, professional persuaders have been joined by data management companies and data brokers, spawning self-regulated strategic communications consultants whose aim is audience influence and behaviour change, often leveraged through localised influencers, bots and trolls.

As a case in point, an ethnographic study across 2016–2017 in the Philippines problematises the work hierarchies and institutions that professionalise and incentivise 'paid troll' work (see Chap. 3). The study stresses the importance of understanding local contexts of how architects of disinformation evade responsibility and entice young creative professionals in need of paid employment to join them. Similar processes have been documented in Guatemala (Currier & Mackey, 2018, April 7). The Philippines' study recommends greater industry self-regulation and

development and enforcement of stronger codes of ethics to encourage transparency and accountability in digital marketing, political marketing (including a requirement to disclose political consultancies) and the digital influencer industry (where undisclosed paid sponsorships and collaborating with anonymous digital influencers enable people to elide accountability) (Ong & Cabañes, 2018). More prescriptively, the Final Report from the UK Inquiry into Disinformation and Fake News recommends that the government move beyond self-regulation to consider new regulations on transparency in strategic communications companies, with a public record of all campaigns that they work on domestically and abroad (Digital, Culture, Media and Sport Committee, 2019, February 14, pp. 83–84). Globally, however, the strategic communications industry remains largely unregulated and opaque, with self-regulation failing to stymie the architects of disinformation.

Solution Area 6: Media Organisations

Media organisations can raise awareness of disinformation and how it works, propagate true stories that connect with audiences and hold powerholders to account. However, this requires a healthy media ecology. Where the media ecology is unhealthy, steps should be taken to strengthen it. This section considers two macro solutions (namely, restoring competitive balance and rebuilding trust in mainstream news) and one solution that has become globally prominent in recent years (fact-checking).

Restoring Competitive Balance

In most liberal democracies, print media content is not extensively regulated because these markets are usually decentralised and competitive, whereas broadcast media are highly regulated because of their formerly oligopolistic or monopolistic position. Today, the scale and reach of dominant Internet platforms means that they occupy a position similar to that of legacy television networks (Fukuyama & Grotto, 2020). Furthermore, as Chap. 2 details, the impact of digital platforms on the business model of legacy news has been profoundly damaging, siphoning ad revenue and discouraging people from paying for news, generating news deserts where it has become uneconomic to provide news.

Previous democratic crises of media pluralism involving new technologies (from radio onwards) saw parliaments legislating to increase media

pluralism by, for instance, funding new sources of trusted local information (notably, public service broadcasters) and introducing media ownership laws to prevent existing monopolists reaching into new media (Marsden et al., 2020). Competitive balance in the digital platform-dominated media ecology could be restored by breaking up the platforms to diminish their influence (McNamee, 2019) or by demanding that technology platforms divert more of their profits to finance local news, investigative journalism and public service journalism. Since the 2016 fake news furore captured public and political attention, Google and Facebook have voluntarily paid publishers around the world hundreds of millions of dollars to sponsor news-related projects (Benton, 2022). Some criticise such voluntary efforts as too minimal, suggesting instead that platforms redistribute a small percentage of their revenue as part of a new social contract to address the loss of public service journalism (Pickard, 2020). More recently, governments have also started to apply pressure, as evidenced in Australia, which passed the Treasury Laws Amendment (News Media and Digital Platforms Mandatory Bargaining Code) Act, 2021, requiring social media companies to pay media outlets for using their content. However, such regulatory measures risk dominant digital platforms shifting their investment away from news altogether, especially where news is not core to their product; in Facebook, for instance, only *one out of every 250* News Feed content views in the first quarter of 2022 were to external links to a news site (Benton, 2022). As Chap. 2 reminds us, algorithmic tweaks on the dominant platforms have huge impacts on the fortunes of news outlets. Moreover, news outlets struggle to make a profit in the digital environment, so if dominant digital platforms roll back their recent overtures towards financially supporting news outlets, then the onus will fall on others to step in.

Rebuilding Trust in Mainstream News

Trust in the truthfulness of journalism has been low for decades, predating the social media era, as Chap. 4 reminds us, fuelled by long-standing political and commercial processes of manipulation and commodification that shape how news stories are constructed. Trust remains low across the world, as shown in a survey in 2020 of the digital news consumption of over 80,000 people in 40 countries: it finds overall levels of trust in news at their lowest point since they started to track such data, with only 38% saying they trust most news most of the time (Newman et al., 2020). A

similar fig. (42%) was found in a 2022 survey of 46 countries, with only 19% saying all or most news organisations put what is best for society ahead of their own commercial or political interests. Notably, it is public service broadcasting organisations with a strong track record of independence that attract the highest trust ratings (Newman et al., 2022). Ultimately, then, it would seem apposite to invest in public service broadcasting across the world. Beyond such large-scale investment, various solutions addressing false information online have been proposed to rebuild trust in journalism.

One proposed solution involves creating guidelines for journalists for reporting false information. A 2018 survey of 803 American and British journalists finds that such reporting guidelines are not widespread. This is problematic because highlighting insignificant fake stories can draw extra attention to false information, giving those propagating the false story credibility because they can point to mainstream media engagement with it (Persen et al., 2021). A study for the Council of Europe (an organisation that seeks to develop throughout Europe common and democratic principles based on the European Convention on Human Rights and other reference texts) argues that newsrooms need policies on strategic silence on fake news to inform decisions about what stories to debunk and which to ignore (namely, those not gaining traction) (Wardle & Derakshan, 2017, p. 19).

Other proposed solutions involve greater journalistic transparency of online news sources and journalistic processes. For instance, the European Commission (2018a) suggests that platforms should integrate source transparency indicators into their ranking algorithms to better signal trustworthy and identifiable sources in search engine results and social media news feeds. This challenge has been taken up by the Coalition for Content Provenance and Authenticity (led by Adobe, ARM, the British Broadcasting Corporation, Intel, Microsoft, TruePic and Twitter). It is developing technical specifications for content provenance enhancing technologies, to help users decide whether content is manipulated by applying the content's metadata to determine who created it, how and when (The Royal Society, 2022, January). However, this does nothing to reduce the scale of false information online. Matters are not helped by the fact that large influence operations on social media often use established media outlets as camouflage (as discussed in Chap. 4). Conversely, a more radical solution to rebuilding trust would be to discard the unobtainable professional ideal of impartial and objective journalism. Winston and Winston (2021)

propose that a more openly subjective, biased journalism would be better at providing a public forum, analysing context, mobilising citizens and building empathy between communities while being no worse at providing new information and holding power to account. These diverse solutions to improving journalistic transparency could be worth trying, but currently lack empirical evidence on ability to rebuild trust, itself a complex phenomenon that, once lost, is difficult to regain. These solutions also do nothing to address the root cause of distrust in news, namely, the long-standing political and commercial processes of manipulation and commodification.

Fact-Checking

Fact-checking is the practice of systematically publishing assessments of the validity of claims made by public bodies to identify whether a claim is factual (Walter et al., 2020). Political fact-checking emerged in the late 1980s, following deceptive, unchallenged ads in the 1988 US presidential race. In 2015, the Poynter Institute established the International Fact-Checking Network to bring together fact-checkers from around the world. By 2020 there were 290 active fact-checking sites in 83 countries, although most are in Europe, North America and Asia, with fewer in South America or Africa (Stencel & Luther, 2020).

A key challenge to fact-checking is its resource-intensiveness and hence expense: for example, fact-checker, PolitiFact, takes three editors to judge whether a piece of news is false (Oshikawa et al., 2020). Fact-checking therefore tends to be reserved for important moments where the *civic body* requires protection, such as elections (Rodríguez-Pérez et al., 2021). In countries such as Brazil, Argentina and Mexico, journalists have been successfully collaborating with each other during elections to reduce the costs of fact-checking, prevent duplication of newsrooms debunking the same content and ensure that quality information reaches larger audiences. Successful collaborations also counter false pandemic information, such as the 91 verification units from 70 countries that feed the database, The CoronaVirusFacts/DatosCoronaVirus Alliance, supported by the International Fact-Checking Network (Palomo & Sedano, 2021). Yet, given its resource-restricted factual focus, fact-checking will fail to identify gendered disinformation, where the claims being made are couched in value judgements or are about people's character (Judson et al., 2020, October).

Furthermore, fact-checking itself is not immune to the influence of powerful actors. For instance, fact-checkers are *beholden to being recognised by dominant digital platforms.* For instance, Google applies a series of tests, including that the fact-checking organisation must qualify for inclusion in Google News, itself an opaque and controversial process, and that publishers must be algorithmically determined to be an authoritative source of information (Graves & Anderson, 2020).

Ultimately, the efficacy of fact-checking may be minimal as *those who most need to see the fact-check do not* (Moreno-Gil et al., 2021). Guess et al. (2018) estimate that about one in four Americans visited a fake news website around the 2016 US presidential election and that fact-checking failed to counter fake news because consumption of fact-checks was concentrated among non-fake news consumers. Beyond the USA, a study of public attitudes towards fact-checking in Europe finds greater acceptance of fact-checking in Sweden and Germany than in Italy, Spain, France and Poland. Dissatisfaction with democracy and the European Union also predicts negative feelings towards fact-checkers in five of the countries examined (although not France) (Lyons et al., 2020). Furthermore, fact-checking sites do not seem to influence the issue agenda of other media. Vargo et al.'s (2018) computational study of the role of fake news in the online news media landscape from 2014 to 2016 finds that fact-checking websites had half the influence of fake news in 2016. A report by the Atlantic Council (a US non-partisan think tank) observes that fact-checking is also ineffective where telecommunications companies' zero-rating policies incentivise social media users to remain in a closed online space within platforms, making it hard for them to verify claims using external resources (Bandeira et al., 2019). There are also psychological factors that can mitigate effectiveness of fact-checking (explored in the following section).

Solution Area 7: Education

There have been multi-stakeholder efforts to improve citizen's digital literacy around false information. Increasingly promoted by the globally dominant digital platforms, in 2017, Facebook launched its 'Facebook Journalism Project' and announced that news literacy would be a priority. Beyond financially supporting non-profits working in this space, it also

rolled out a Public Service Announcement-type message at the top of the News Feed in 14 countries, linking to a post with tips for spotting 'false news' (Murphy, 2020, July 8, p. 29). As Chap. 6 observes, however, even among highly educated audiences in India, this media literacy campaign had only short-term effects in improving discernment between mainstream and false news headlines (Guess et al., 2020). Reportedly more successful efforts consider how best to reach the digitally illiterate. In India, for instance, to counter digitally illiterate village communities reacting with terrified mob violence towards false information on WhatsApp (Bali & Desai, 2019), the police (in Telangana state) in 2018 used 'Janapadam', namely, folklore that establishes a connection with locals. This involved short skits where primarily lower caste communities share religious tales and important news. They typically feature a man and two women sitting together to narrate a story, ending with a message promoting digital literacy. This audience-targeted approach reportedly generated broad reach and acceptance among local communities (Singh, 2019, January 9).

In much more digitally literate Finland, the government launched an anti-fake news initiative in 2014 to teach citizens, journalists and politicians how to counter false information designed to sow division. Finland was attuned to Russian propaganda having faced this since declaring independence from Russia a century prior. As online trolling increased in 2014, after Moscow annexed Crimea and backed rebels in eastern Ukraine, Finland reformed its education system in 2016 to emphasise critical thinking. However, Finland may have unique features that make media literacy efforts more likely to succeed. As well as a long history of dealing with foreign propaganda, it is a small, homogenous country that consistently tops international indexes on happiness, press freedom, gender equality, social justice, transparency, education and trust in national media, making it hard for external actors to find social fissures to exploit (Mackintosh, 2019, May; Newman et al., 2018).

Such media literacy governance solutions (policies, funding, tools) may be beneficial when conducted under appropriate conditions attuned to local contexts. However, they may have only short-term effects and are unevenly rolled out worldwide. They also run into complex psychological and sociological issues of how and why people spread and remember false information, which we discuss below.

Correcting False Information Does Not Change Beliefs

On whether fact-checking messages influence what we believe, Walter et al. (2020, pp. 17–18) present optimistic and pessimistic conclusions from their meta-analysis of 30 studies. Their optimistic interpretation is that people's beliefs become more accurate and factually consistent, even after a single exposure to a fact-checking message. Their pessimistic interpretation is that fact-checking has weak impacts on beliefs that become negligible the more the study resembles real-world scenarios of exposure to fact-checking. Chan et al.'s (2017) meta-analysis of the psychological efficacy of messages countering misinformation finds that debunking effects were weaker when audiences generate reasons in support of the initial misinformation, supporting what we know about the power of confirmation bias. Correcting misinformation therefore does not necessarily change people's beliefs (Flynn et al., 2017).

By contrast, a near-universal finding is 'the continued influence effect' where, even after its correction, misinformation continues to influence people's attitudes and beliefs (Wittenberg & Berinsky, 2020, p. 174). Experiments show that repeated exposure to fake news headlines increases their perceived accuracy: this occurs despite a low level of overall believability and even when stories are labelled as contested by fact-checkers or are inconsistent with readers' political ideology. These results suggest that platforms help incubate belief in false information and that tagging such stories as 'disputed' is ineffective as any repetition of misinformation, even in the context of refuting it, may be harmful (Pennycook et al., 2018).

Given this state of affairs, psychological research shows that inoculating people with information *before* their minds are made up on an issue may better ensure that false information does not circulate (Cook et al., 2017). Inoculation theory (McGuire, 1964) was pioneered to induce attitudinal resistance against propaganda and persuasion. It holds that activating people's 'mental antibodies' through a weakened dose of the infectious agent can confer resistance against future attempts to persuade them. A decade-old meta-analysis of studies finds that inoculation is effective at conferring resistance (Banas & Rains, 2010). Recent studies find that inoculating people with facts against misinformation works for a highly politicised issue (global warming), regardless of prior attitudes (Cook et al., 2017; van der Linden et al., 2017). Applying inoculation theory to fake news finds that inoculation has some effect in making participants more sceptical and attuning people to deception (Roozenbeek & van der Linden, 2019).

Nudges

Experiments have deployed nudges to make people more careful about what they circulate online. Theories suggest that 'social norm' nudges work by informing people how others behave, so triggering desire to conform; by reminding people what the norms are, thereby changing behaviour to avoid social sanctions from norm-breaking; and by indicating what the 'best' course of action is, so changing behaviour (Legros & Cislaghi, 2020). 'Confront nudges' try to pause unwanted actions by instilling doubt, attempting to break mindless behaviour and prompting reflective choices (Caraban et al., 2019).

Numerous nudging experiments have been conducted on social media to see if they can reduce harms. For instance, Andı and Akesson (2021) designed a social norm-based message that nudges people towards better sharing behaviour. Their study placed the nudge above a thumbnail link to a false news article and provided a reminder that false news is prevalent online and that most responsible people think twice before sharing news. Participants exposed to the nudge were 5% less likely to say that they were willing to share the article. Such nudges could form a firebreak in online emotional contagion. A 'confront nudge' that provides multiple viewpoints to overcome our confirmation bias is NewsCube: it collects different points of view and offers an unbiased clustered overview in evenly distributed sections, while identifying unread sections, to nudge users to read all viewpoints (Park et al., 2009). Levy's (2021) US-based field experiment conducted in 2018 (>17,000 participants) provides the first experimental evidence that exposure to counter-attitudinal news on Facebook decreases affective polarisation, so demonstrating that nudges diversifying social media news exposure could be effective.

However, researchers increasingly note the inability of behaviour change technologies to sustain user engagement. Furthermore, few examine long-term effects of nudging, and most do not examine possible backfires and unexpected effects. Reasons why nudges fail include that techniques tapping into the automatic mind lack any educational effects, and hence their effects may cease when nudges are removed; reminders might cause reactance after repeated exposure; and graphic warnings can lose resonance over time (Caraban et al., 2019).

Reason and Emotion

Scholarship suggests a positive role for reasoning in resisting false information. Ross et al. (2021) conducted two studies asking 1973 Americans to assess true, false and hyperpartisan news headlines from Facebook. It finds that analytical thinking was mostly associated with an increased tendency to distinguish true headlines from false and hyperpartisan headlines and that analytical thinking was not generally associated with increased willingness to share hyperpartisan or false headlines. Pennycook and Rand's (2019) study of 3446 Mechanical Turk workers concludes that analytical thinking is used to assess plausibility of headlines, regardless of whether stories are consistent with one's political ideology. Their findings suggest that susceptibility to fake news is driven more by lazy thinking than partisan bias. As such, training people to think more analytically, or giving (nudging) people time to take a moment for an analytical breath, could be fruitful.

Also important is the need to educate people on the power of emotive content to manipulate, as well as on the power of emotion when deployed in AI-driven behavioural prediction models that can be used for influence (McNamee, 2019, p. 260). Wardle and Derakshan (2017, p. 70) argue that any media literacy curriculum should include techniques for developing emotional scepticism to override our brain's tendency to be less critical of content that provokes our emotions. Of course, this may fail where disinformation is crafted to provoke more subtle emotional responses. Indeed, as Bennett and Livingston (2020) observe, recommendations that focus on educating people about detecting false information avoid the question of why so many people easily exchange facts for deeper emotional truths. Sociologically informed research, for instance, suggests that sharing fake news might be an expression of group identity or dissatisfaction with the current political system. As such, it is important for educators to address the impact of past disinformation campaigns, as well as current inequalities, on people's willingness to believe falsehoods (Nisbet & Kamenchuk, 2019).

Conclusion

In assessing seven solution areas to false information online, we conclude that each has an important role in strengthening the *civic body*, but also faces intrinsic challenges. Some solutions trample on human rights; others

come up against the limits of technological fixes; and others are stymied by commercial imperatives, lack of political will, or the complexity of our interactions with false information online. The unrelenting scale, speed and spread of false information online; the unpreparedness of automation to detect and address all false information online; the lack of transparency and ethics in digital political advertising and wider strategic communications in the political sphere; the unhealthy media ecology dominated by global digital platforms, decreasing trust in news and under-resourced fact-checking and journalism; and the practical, psychological and sociological limits to increasing people's digital literacy truly make this a 'wicked' problem.

The first solution area, governmental action, varies from non-coercive to coercive responses. Supranational, and many national, declarations urge better self-regulation of platforms, but more coercive responses include arrests, Internet shutdowns, legislation on false information online that stifles dissenting views, targeted legislation to protect key moments for the *civic body* such as elections, and broader legislation and actions to make dominant big technology platforms more responsible for the content that they host and to curb their monopoly power. Many of the coercive responses contravene the human right to freedom of speech, are often abused by authoritarian states and require significant resourcing for compliance. However, non-coercive responses have not solved the problem either.

The second solution area, cybersecurity, involves countries and supranational networks (such as the European Union) actively monitoring and combating foreign disinformation campaigns; social media platforms detecting and removing disinformation content and networks; and multistakeholder approaches to develop appropriate technology such as automated recognition of deceptive media forms. However, cybersecurity responses are uneven worldwide; and there are many methodological and practical problems with using AI for fake news and deepfake detection.

The third solution area, digital platforms and intermediaries, has found globally dominant platforms signing up to self-regulatory approaches with multiple commitments, but it is efforts by platforms around content moderation that have attracted sustained criticism. As well as freedom of speech issues (a right unevenly enforced worldwide), content moderation raises the issue of lack of transparency about what content has been removed or promoted, and why; and there is inconsistency as platforms' policies differ in application regarding what content is removed or promoted, their

stance changing over time. A related problem is that platforms perform poorly at enforcing intermediary liability laws consistently at scale. Meanwhile, the media ecology enables those censored on one platform to simply move onto others. There are also intrinsic technical difficulties of content moderation on any platform, but especially in end-to-end encrypted systems, as it requires skill and resources to detect the nature of posts. While the forthcoming European Union Digital Services Act has demonstrated legislative will to make platforms explain their content moderation policies, practices and decisions more clearly, it is too soon to know if sufficient resources are being ringfenced to ensure compliance, whether platform lobbying will dilute the law, or whether similar legislation will be passed outside of the European Union.

The fourth solution area, advertising, has seen dominant digital platforms, ad networks, programmatic companies and non-profit organisations acting to disrupt business models for producing and amplifying disinformation. However, their activities continue to be challenged by the volume and speed of the supply chain for fake news outlets. The likely ending of the cookie-based behavioural advertising market beckons as GDPR takes effect in media markets throughout the European Union, but what it will be replaced by, as well as its likely impact on the *civic body*, is unclear.

The fifth solution area, professional persuaders and public relations in the political domain, finds broad stakeholder agreement and legislative activity in certain regions (such as the European Union) on the need to greatly increase transparency of online political ads in terms of who purchased them, to whom they are targeted, and on what basis, and to enable advertisers to be held accountable. However, such legislation is needed in every country where digital political campaigning occurs. Ad libraries remain under the control of the dominant platforms and, in their current form, are minimally useful for electoral regulators. Finding solutions to broader strategic communications (a self-regulated area) that disseminate disinformation worldwide has proven harder, given lack of transparency, absence of professional ethics and localised conditions that entice creative professionals to engage with paid troll work. Diverse countries recommend greater transparency, self-regulation and regulation of strategic communications companies (including political marketing, digital marketing and the digital influencer industry).

The sixth solution area, media organisations, could play a vital role in combating online disinformation campaigns, but this requires a healthy

media ecology. This in turn requires restoration of competitive balance, with suggestions ranging from breaking up dominant digital platforms to making them redistribute more of their advertising revenue back to media organisations. Such solutions, however, require uncompromising legislative intent and action by governments worldwide and also risk provoking dominant digital platforms to pivot away from news altogether (further damaging the revenue streams of news outlets). There are also proposed solutions to rebuild trust in journalism such as through newsroom policies on when to debunk false information online and greater journalistic transparency regarding news story construction. While such actions may help, empirical studies on efficacy are lacking. Declining trust in news is a long-standing, complex issue and unlikely to be solved any time soon given that news stories are a construct and that journalism remains beholden to long-standing political and commercial processes of manipulation and commodification. Ultimately, it would seem apposite to invest in independent public service broadcasting across the world, as it is such news outlets that currently garner greatest trust, but this would require large-scale investment. A dominant solution globally is promotion of fact-checking, but obstacles include resource-intensiveness and expense; that fact-checking itself is not immune to the influence of powerful actors; and that the efficacy of fact-checking may be minimal as those who most need to see the fact-checks do not.

The seventh solution area, education, has been embraced by many countries which have adopted campaigns to improve their citizen's digital literacy and awareness of online disinformation. Those considered successful have carefully considered how best to reach the digitally illiterate or operate in small, relatively homogenous, progressive countries with a history of dealing with disinformation. However, while media literacy solutions may work when conducted under appropriate conditions attuned to local contexts, they may have only short-term effects and are unevenly rolled out worldwide. They also run into complex psychological and sociological issues of how and why people spread and remember false information. Scholarship shows limited impact on people's beliefs from correcting false information (although inoculation can prove useful); the potential of nudging to make people more careful in what they circulate online (but that nudging may only have short-term effects); and various roles played by reason (training people to think or act analytically) and emotion (developing emotional awareness and scepticism towards content and algorithms). Fundamentally, however, literacy approaches alone cannot address

why so many people easily exchange facts for deeper emotional truths. The task then broadens to educators (especially of history, sociology and communications) to address the impact of past disinformation combined with present-day inequalities on people's current willingness to believe falsehoods.

Where does this leave us? Reducing the overall volume, and impacts, of false information circulating online would seem paramount. However, over six years of intensive governance and multidisciplinary academic interest in tackling false information online has not yet fixed the problem. We conclude that the ultimate solution would be to alter the business models of platforms, so that they do not seek maximal user engagement and so that they do not design algorithms that make emotional and deceptive content go viral (see Section I). In lieu of directly addressing the innate dynamics of informational capitalism and the *economics of emotion*, we are left to tinker at the edges with imperfect solutions. Ultimately, when set against business models that promote emotive, false information, any proposed solution faces an uphill task. As Chap. 2 explains, leaked Facebook documents show that Facebook's News Rank algorithm has prioritised emotional, engaging reactions, with posts sparking 'Angry', 'Wow' and 'Haha' Reaction emoji disproportionately likely to include misinformation, toxicity and low-quality news. The power of the algorithmic promotion undermined efforts by Facebook's content moderators and integrity teams to reduce toxic, harmful content. Yet, Facebook has the power to address matters at source. In 2020, Facebook cut the weight of all Reactions to one and a half times that of a 'Like' and, in September 2020, cut the weight of the 'Angry' Reaction to zero. As a result, Facebook users began to get less misinformation, less 'disturbing' content and less 'graphic violence' (Merrill & Oremus, 2021, October 26). Twitter also has the power to reduce viral false information and occasionally does so to protect the *civic body*, as discussed in Chap. 4. For instance, in preparation for the 2020 US presidential elections, Twitter temporarily introduced friction to slow the spread of misleading information by reducing the overall amount of sharing on the platform (Gadde & Beykpour, 2020, November 12). Whether social media platforms will address their business models to permanently dampen false information online remains to be seen.

Eager to prevent regulation along these lines, globally dominant digital platforms regularly point to their many mitigation efforts and to the good that their platforms enable, including the large amount of money and

creativity that their presence creates in countries. As such, a redesign of algorithms to make platforms *less engaging* is unlikely to happen without either (a) a mass exodus of users (which is unlikely given how strongly imbricated the dominant digital platforms are into people's daily lives) or (b) strong governmental and coordinated intergovernmental intervention to regulate algorithms that promote emotive, false information (care would be needed not to sacrifice the benefits of free speech). At stake is whether it is acceptable for globally dominant digital platforms to be deciding, ultimately, what is optimal, optimisable, or optimised in a public sphere shaped by datafied emotion, given the many harms to the *civic body* that we have identified.

Importantly, false information online has been incubated to date by globally dominant digital and social media platforms. But they are just the currently most prevalent use case of emotional profiling, with many more emergent forms of emotional AI being trialled and rolled out globally. As such, we need to consider near-horizon possible futures and formulate principles to strengthen the *civic body* when faced with the rising tide of emotional AI. It is to this task that we turn in the following, and final, chapter.

References

Ajder, H., & Glick, J. (2021). *Just joking! Deepfakes, satire and the politics of synthetic media*. WITNESS and MIT Open Documentary Lab. Retrieved 13 Apr 2022, from https://cocreationstudio.mit.edu/just-joking/

Amnesty International. (2022, March 10). *Russia: Kremlin's ruthless crackdown stifles independent journalism and anti-war movement*. Retrieved 13 Apr 2022, from https://www.amnesty.org/en/latest/news/2022/03/russia-kremlins-ruthless-crackdown-stifles-independent-journalism-and-anti-war-movement/

Andı, S., & Akesson, J. (2021). Nudging away false news: Evidence from a social norms experiment. *Digital Journalism, 9*(1), 106–125. https://doi.org/10.1080/21670811.2020.1847674

Au, L. (2021, September 27). Why China crushed its tech giants. *Wired*. https://www.wired.co.uk/article/china-tech-giants-policy

Bakir, V. (2019). *Intelligence elites & public accountability: Relationships of influence with civil society*. Routledge. (Original work published 2018).

Bali, A., & Desai, P. (2019). Fake news and social media: Indian perspective. *Media Watch, 10*(3), 737–750. https://doi.org/10.15655/mw/2019/v10i3/49687

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs, 77*(3), 281–311. https://doi.org/10.1080/03637751003758193

Bandeira, L., Barojan, D., Braga, R., Peñarredonda, J. L., & Pérez Argüello, M. F. (2019). *Disinformation in democracies: Strengthening digital resilience in Latin America* (pp. 20–29). Atlantic Council. Retrieved 13 Apr 2022, from https://www.atlanticcouncil.org/in-depth-research-reports/report/disinformation-democracies-strengthening-digital-resilience-latin-america/

Bank, M., Duffy, F., Leyendecker, V., & Silva, M. (2021). *The lobby network: Big tech's web of influence in the EU*. Corporate Europe Observatory and Lobby Control: Brussels and Cologne.EU. Retrieved 13 Apr 2022, from https://corporateeurope.org/sites/default/files/2021-08/The%20lobby%20network%20-%20Big%20Tech%27s%20web%20of%20influence%20in%20the%20EU.pdf

Bennett, W. L., & Livingston, S. (2020). A brief history of the disinformation age: Information wars and the decline of institutional authority. In W. L. Bennett & S. Livingston (Eds.), *The disinformation age: Politics, technology and disruptive communication in the information age* (pp. 3–42). Cambridge University Press. https://doi.org/10.1017/9781108914628

Benton, J. (2022, June 16). Facebook looks ready to divorce the news industry, and I doubt couples counseling will help. *Nieman Lab*. Retrieved 20 June 2022, from https://www.niemanlab.org/2022/06/facebook-looks-ready-to-divorce-the-news-industry-and-i-doubt-couples-counseling-will-help/

Bond, S. (2021, November 9). Facebook scraps ad targeting based on politics, race and other 'sensitive' topics. *NPR*. https://www.npr.org/2021/11/09/1054021911/facebook-scraps-ad-targeting-politics-race-sensitive-topics#:~:text=More%20Podcasts%20%26%20Shows-Facebook%20scraps%20ad%20targeting%20based%20on%20politics%2C%20race%20and%20other,new%20rules%20begin%20in%20January

Bozarth, L., & Budak, C. (2021). Market forces: Quantifying the role of top credible ad servers in the fake news ecosystem. *Proceedings of the International AAAI Conference on Web and Social Media, 15*(1), 83–94. Retrieved 13 Apr 2022, from https://ojs.aaai.org/index.php/ICWSM/article/view/18043

Brown, I., Marsden, C. T, Lee, J., & Veale, M. (2020). *Cybersecurity for elections. A commonwealth guide on best practice*. Retrieved 13 Apr 2022, from https://thecommonwealth.org/sites/default/files/inline/Commonwealth%20cybersecurity%20for%20elections%20guide.pdf

Caraban, A., Karapanos, E., Goncalves, D., & Campos, P. (2019). 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. *CHI 2019*, May 4–9, Glasgow. https://doi.org/10.1145/3290605.3300733.

Central Commission for Cybersecurity and Informatization. (2021, December 28). *14th Five-Year Plan for National Informatization*. Retrieved 27 Apr 2022, from https://digichina.stanford.edu/work/translation-14th-five-year-plan-for-national-informatization-dec-2021/

Cha, M., Gao, W., & Li, C.-T. (2020). Detecting fake news in social media: An Asia-Pacific perspective. *Communications of the ACM*, March. https://doi.org/10.1145/3378422.

Chan, M. S., Jones, C. R., Jamieson, K. H., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science, 28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS One, 12*(5), 1–21. https://doi.org/10.1371/journal.pone.0175799

Council of the EU. (2022, April 23). *Digital Services Act: Council and European Parliament provisional agreement for making the internet a safer space for European citizens*. Retrieved 26 Apr 2022, from https://www.consilium.europa.eu/en/press/press-releases/2022/04/23/digital-services-act-council-and-european-parliament-reach-deal-on-a-safer-online-space/

Currier, C., & Mackey, D. (2018, April 7). The rise of the net centre. *The Intercept*. https://theintercept.com/2018/04/07/guatemala-anti-corruption-trolls-smear-campaign/

Digital, Culture, Media and Sport Committee. (2019, February 14). *Disinformation and 'fake news': Final report*. Digital, Culture, Media and Sport Committee, House of Commons 1791. Retrieved 13 Apr 2022, from https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

Digital Industry Group Inc. (2021). *Australian code of practice on disinformation and misinformation: An industry code of practice developed by the Digital Industry Group Inc*. Retrieved 13 Apr 2022, from https://digi.org.au/wp-content/uploads/2021/02/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL-PDF-Feb-22-2021.pdf

Dommett, K., & Power, S. (2020). *Democracy in the dark: Digital campaigning in the 2019 general election and beyond*. Electoral Reform Society. Retrieved 13 Apr 2022, from https://www.electoral-reform.org.uk/latest-news-and-research/publications/democracy-in-the-dark-digital-campaigning-in-the-2019-general-election-and-beyond/

Edelson, L., Sakhuja, S., Dey, R., & McCoy, D. (2018). *An analysis of United States online political advertising*. Preprint retrieved from https://arxiv.org/abs/1902.04385

Emerson, S. (2021, January 14). MeWe sold itself on privacy. Then the radical right arrived. *Medium*. Retrieved 13 Apr 2022, from https://onezero.medium.com/mewe-sold-itself-on-privacy-then-the-radical-right-arrived-e527b38e4718

Eurobarometer. (2018, February). *Fake news and disinformation online*. Flash Eurobarometer 464. Retrieved 13 Apr 2022, from https://europa.eu/eurobarometer/surveys/detail/2183

European Commission. (2021a, November 25). *European democracy: Commission sets out new laws on political advertising, electoral rights and party funding*. Retrieved 13 Apr 2022, from https://ec.europa.eu/commission/presscorner/detail/en/ip_21_6118

European Commission. (2021b). *Code of practice on disinformation*. Retrieved 13 Apr 2022, from https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

European Commission. (2021c). *Code of practice on disinformation: Commission welcomes new prospective signatories and calls for strong and timely revision*. Retrieved 13 Apr 2022, from https://ec.europa.eu/commission/presscorner/detail/en/IP_21_4945

European Commission. (2020, December 3). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. On the European democracy action plan*. Retrieved 13 Apr 2022, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=COM:2020:790:FIN&from=EN

European Commission. (2018b). *EU code of practice on disinformation*. Europa.eu. Retrieved 13 Apr 2022, from https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

European Commission. (2018a). *A multi- dimensional approach to disinformation*, Directorate-General for Communication Networks, Content & Technology. Retrieved 13 Apr 2022, from https://ec.europa.eu/digitalsingle-market/en/news/final-report-high-level-expert-group-fake-news-and-onlinedisinformation

Facebook. (n.d.). *Become authorised to run ads about social issues, elections or politics*. Retrieved 13 Apr 2022, from https://www.facebook.com/business/help/208949576550051?id=288762101909005

Facebook. (2021). *Availability for ads about social issues, elections or politics*. Retrieved 13 Apr 2022, from https://www.facebook.com/business/help/2150157295276323?id=288762101909005

Facebook. (2020a, April). *Coordinated inauthentic behavior report*. Retrieved 13 Apr 2022, from https://about.fb.com/wp-content/uploads/2020/05/April-2020-CIB-Report.pdf

Facebook AI. (2020b, June 25). *Deepfake detection challenge dataset*. Retrieved 13 Apr 2022, from https://ai.facebook.com/datasets/dfdc/

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology, 38*(51), 127–150. https://doi.org/10.1111/pops.12394

Freedom House. (2021). *Freedom on the net 2021: The global drive to control big tech*. Retrieved 13 Apr 2022, from https://freedomhouse.org/report/freedom-net/2021/global-drive-control-big-tech

Fukuyama, F., & Grotto, A. (2020). Comparative media regulation in the United States and Europe. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 199–219). Cambridge University Press.

Gadde, V., & Beykpour, K. (2020, November 12). *An update on our work around the 2020 US Elections*. Retrieved 13 Apr 2022, from https://blog.twitter.com/en_us/topics/company/2020/2020-election-update.html

Goel, V. (2022). Get to know the new topics API for privacy sandbox, *Google*. Retrieved 13 Apr 2022, from https://blog.google/products/chrome/get-know-new-topics-api-privacy-sandbox/

Ghulati, D. (2020, November 27). Factmata. Tackling fake news and online misinformation. *Westminster media forum policy conference*, 27 November.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies* (pp. 167–194). MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Glazer, E. (2019, November 21). Facebook weighs steps to curb narrowly targeted political ads. *The Wall Street Journal*. https://www.wsj.com/articles/facebook-discussing-potential-changes-to-political-ad-policy-11574352887?mod=

Global Disinformation Index. (2021). *Disinformation risk assessment: The online news market in Kenya*. Retrieved 13 Apr 2022, from www.disinformationindex.org

Google. (2022). *Political content*. Retrieved 13 Apr 2022, from https://support.google.com/adspolicy/answer/6014595?hl=en-GB

Gorwa, R., & Ash, T. G. (2020). Democratic transparency in the platform society. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 286–312). Cambridge University Press.

Graves, L., & Anderson, C. W. (2020). Discipline and promote: Building infrastructure and managing algorithms in a "structured journalism" project by professional fact-checking groups. *New Media and Society, 22*(2), 342–360. https://doi.org/10.1177/1461444819856916

Goujard, C. (2022, April 23). Big Tech firms set to face tough EU content rules. *Politico*. https://www.politico.eu/article/eu-strikes-deal-on-law-to-fight-illegal-content-online-digital-services-act/

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment

between mainstream and false news in the United States and India. *PNAS, 117*(27), 15536–15545. www.pnas.org/cgi/doi/10.1073/pnas.1920498117

Guess, A., Nyhan, B., & Reifler, J. (2018). *Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign*. Retrieved from https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf

Hanson, F., O'Connor, S., Walker, M. & Courtois, L. (2019). *Hacking democracies: Cataloguing cyber-enabled attacks on elections*, Policy Brief 16. Australian Strategic Policy Institute. Retrieved 13 Apr 2022, from https://www.aspi.org.au/report/hacking-democracies

Hendrix, J. (2021, January 7). *Deplatforming Donald Trump*. Tech Policy Press. https://techpolicy.press/deplatforming-donald-trump/

Human Rights Watch. (2018, February 18). *Germany: flawed social media law: NetzDG is wrong response to online abuse*. Retrieved 13 Apr 2022, from https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

Internet Advertising Bureau UK. (2021). *User ID solutions from IAB UK members*. Retrieved 13 Apr 2022, from https://www.iabuk.com/user-identity/understanding-user-enabled-id-solutions

Judson, E., Atay, A., Krasodomski-Jones, A., Lasko-Skinner, R., & Smith, J. (2020, October). The contours of state-aligned gendered disinformation online. Demos, London. Retrieved 23 June 2022, from https://demos.co.uk/project/engendering-hate-the-contours-of-state-aligned-gendered-disinformation-online/

Jung, A.-K., Ross, B., & Stieglitz, S. (2020). Caution: Rumors ahead – a case study on the debunking of false information on Twitter. *Big Data and Society, 7*(2). https://doi.org/10.1177/2053951720980127

Keller, D., & Leerssen, P. (2020). Facts and where to find them: Empirical research on internet platforms and content moderation. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 220–251). Cambridge University Press.

Killion, V. L. (2019, January 16). The First Amendment: Categories of speech. *Congressional Research Service*. IF11072. Retrieved 13 Apr 2022, from https://crsreports.congress.gov/product/pdf/IF/IF11072#:~:text=The%20Court%20generally%20identifies%20these.criminal%20conduct%2C%20and%20child%20pornography

Kofi Annan Commission on Elections and Democracy in the Digital Age. (2020). *Protecting electoral integrity in the digital age*. Retrieved 13 Apr 2022, from www.kofiannanfoundation.org/app/uploads/2020/01/f035dd8e-kafkaceddareport2019web.pdf

Langguth, J., Pogorelov, K., Brenner, S., Filkukova, P., & Schroeder, D. (2021). Don't trust your eyes: Manipulation of visual media in the age of deepfakes.

Frontiers in Political Communication. https://doi.org/10.3389/fcomm.2021.632317

Leerson, P., Dobber, T., Helberger, N., & de Vreese, C. (2021). News from the ad archive: How journalists use the Facebook Ad Library to hold online advertising accountable. *Information, Communication & Society*. Advance online publication. https://www.tandfonline.com/doi/full/10.1080/1369118X.2021.2009002

Legros, S., & Cislaghi, B. (2020). Mapping the social-norms literature: An overview of reviews. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 15*(1), 62–80. https://doi.org/10.1177/1745691619866455

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review, 111*(3), 831–870. https://doi.org/10.1257/aer.20191777

Levy, S. (2020). *Facebook: The inside story*. Penguin.

Lomas, N. (2022, April 22). Europe seals a deal on tighter rules for digital services. *TechCrunch*, https://techcrunch.com/2022/04/22/google-facebook-apple-eu-lobbying-report/

Lyons, B., Mérola, V., Reifler, J., & Stoeckel, F. (2020). How politics shape views toward fact-checking: Evidence from six European countries. *The International Journal of Press/Politics, 25*(3), 469–492. https://doi.org/10.1177/1940161220921732

Mackintosh, E. (2019, May). Finland is winning the war on fake news. What it's learned may be crucial to Western democracy. *CNN*. https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/

Marsden, C., Meyer, T., & Brown, I. (2020). Platform values and democratic elections: How can the law regulate digital disinformation? *Computer Law & Security Review, 36*(April), Article 105373. https://doi.org/10.1016/j.clsr.2019.105373

McGuire, W. J. (1964). Some contemporary approaches. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 1, pp. 191–229). Academic Press.

McNamee, R. (2019). *Zucked: Waking up to the Facebook catastrophe*. Harper Collins.

McNeice, S. (2019, November 5). Proposals aimed at regulating online political ads approved by Cabinet. *Newstalk*. https://www.newstalk.com/news/online-political-ads-regulation-921802

Merrill, J. B. & Oremus, W. (2021, October 26). Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation. *Washington Post*. https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/

Moreno-Gil, V., Ramon, X., & Rodríguez-Martínez, R. (2021). Fact-checking interventions as counteroffensives to disinformation growth: Standards, values,

and practices in Latin America and Spain. *Media and Communication, 9*(1), 251–263. https://doi.org/10.17645/mac.v9i1.3443

Murphy, L. W. (2020, July 8). *Facebook's civil rights audit – Final report.* Retrieved 13 Apr 2022, from https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf

Netzwerkdurchsetzungsgesetz vom 1. September 2017 (BGBl. I S. 3352). (Network Enforcement Law or "NetzDG"). Retrieved 13 Apr 2022, from www.gesetze-im-internet.de/netzdg/BJNR335210017.html

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K. (2022). *Reuters Institute digital news report 2022.* Retrieved June 20, 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Newman, N., Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Reuters Institute digital news report 2020.* Retrieved 13 Apr 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2018). *Reuters Institute digital news report 2018.* Retrieved 13 Apr 2022, from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/digital-news-report-2018.pdf

Nisbet, E. C., & Kamenchuk, O. (2019). The psychology of state-sponsored disinformation campaigns and implications for public diplomacy. *The Hague Journal of Diplomacy, 14*, 65–82. https://doi.org/10.1163/1871191X-11411019

O'Donovan, K. (2020, November 27). Google UK. Tackling fake news and online misinformation. *Westminster media forum policy conference,* 27 November.

Ong, J. C. & Cabañes, J. V. A. (2018). *Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines.* Retrieved 13 Apr 2022, from https://doi.org/10.7275/2cq4-5396

Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A survey on natural language processing for fake news detection. *Proceedings of the 12th language resources and evaluation conference (LREC 2020)* pp. 6086–6093. Preprint retrieved from https://arxiv.org/pdf/1811.00770.pdf

Palomo, B., & Sedano, J. (2021). Cross-media alliances to stop disinformation: A real solution? *Media and Communication, 9*(1), 239–250. https://doi.org/10.17645/mac.v9i1.3535

Park, S., Kang, S., Chung, S., & Song, J. (2009). NewsCube: Delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 443–452). ACM. https://doi.org/10.1145/1518701.1518772

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning.

Cognition, 188(July), 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General, 147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Pérez Dasilva, J., Meso Ayerdi, K., & Mendiguren Galdospin, T. (2021). Deepfakes on twitter: Which actors control their spread? *Media and Communication, 9*(1), 301–312. https://doi.org/10.17645/mac.v9i1.3433

Persen, K., Carter, A., & Woolley, S. C. (2021). Computational propaganda and the news: Journalists' perceptions of the effects of digital manipulation on reporting. In M. Boler & E. Davis (Eds.), *Affective politics of digital media* (pp. 245–260). Routledge.

Pickard, V. (2020). The public media option: Confronting policy failure in an age of misinformation. In W. L Bennett & S. Livingston (Eds.), *The disinformation age: Politics, technology and disruptive communication in the information age* (pp. 238–258). Cambridge University Press https://doi.org/10.1017/9781108914628

Repnikova, M. (2018, September 6). China's lessons for fighting fake news. *Foreign Policy*. https://foreignpolicy.com/2018/09/06/chinas-lessons-for-fighting-fake-news/

Rodríguez-Pérez, C., Paniagua-Rojano, F. J., & Magallón-Rosa, R. (2021). Debunking political disinformation through journalists' perceptions: An analysis of Colombia's fact-checking news practices. *Media and Communication, 9*(1), 264–275. https://doi.org/10.17645/mac.v9i1.3374

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications, 5*(65). https://doi.org/10.1057/s41599-019-0279-9

Ross, R. M., Rand, D. G., & Pennycook, G. (2021). Beyond "fake news": Analytic thinking and the detection of false and hyperpartisan news headlines. *Judgment and Decision making, 16*(2), 484–504. Retrieved 13 Apr 2022, from http://journal.sjdm.org/20/200616b/jdm200616b.html

Shek, K. (2020, November 27). Advertising Standards Association. Tackling fake news and online misinformation, *Westminster Media Forum policy conference*, November 27.

Siegel, A. A. (2020). Online hate speech. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 56–88). Cambridge University Press.

Simonite, T. (2021, October 25). Facebook is everywhere; its moderation is nowhere close. *Wired*. https://www.wired.com/story/facebooks-global-reach-exceeds-linguistic-grasp/

Singh, V. (2019, January 9). An ancient story-telling technique is helping cops fight fake news in India. *Quartz India*. https://qz.com/india/1518770/indian-cops-fightwhatsapp-fake-news-with-ancient-story-telling/

Spencer, S. (2019). *An update on our political ads policy*. Retrieved 13 Apr 2022, from https://www.blog.google/technology/ads/update-our-political-ads-policy

Stencel, M., & Luther, J. (2020). *Annual census finds nearly 300 fact-checking projects around the world*. Duke Reporters' Lab. Retrieved 13 Apr 2022, from https://reporterslab.org/annual-census-finds-nearly-300-fact-checking-projects-around-the-world

Takayasu, M., Sato, K., Sano, Y., Yamada, K., Miura, W., & Takayasu, H. (2015). Rumor diffusion and convergence during the 3.11 earthquake: A twitter case study. *PLoS ONE, 10*(4), Article e0121443. https://doi.org/10.1371/journal.pone.0121443

Thakur, D., & Hankerson, D. L. (2021). *Facts and their discontents: A research agenda for online disinformation, race, and gender*. Center for Democracy & Technology. Retrieved 23 June 2022, from https://osf.io/3e8s5/

The Royal Society. (2022, January). *The online information environment: Understanding how the internet shapes people's engagement with scientific information*. DES7656. Retrieved 13 Apr 2022, from https://royalsociety.org/-/media/policy/projects/online-information-environment/the-online-information-environment.pdf

TikTok. (2021). *Community guidelines*. Retrieved 13 Apr 2022, from https://www.tiktok.com/community-guidelines?lang=en

Treasury Laws Amendment (News Media and Digital Platforms Mandatory Bargaining Code) Act. (2021). Parliament of Australia. Retrieved 13 Apr 2022, from https://parlinfo.aph.gov.au/parlInfo/download/legislation/bills/r6652aspassed/tocpdf/20177b01.pdf;fileType=application%2Fpdf

United Nations Special Rapporteur on Freedom of Opinion and Expression, Organization for Security and Co-operation in Europe Representative on Freedom of the Media, the Organization of American States Special Rapporteur on Freedom of Expression, and African Commission on Human and Peoples' Rights Special Rapporteur on Freedom of Expression and Access to Information. (2017). *Joint declaration on freedom of expression and "fake news," disinformation and propaganda*, U.N. Doc. FOM.GAL/3/17 (Mar. 3, 2017). Retrieved 13 Apr 2022, from https://www.osce.org/fom/302796?download=true [https://perma.cc/5TZC-TPM8].

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges, 1*(2), Article 1600008. https://doi.org/10.1002/gch2.201600008

Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society, 20*(5), 2028–2049. https://doi.org/10.1177/1461444817712086.

Vincent, J. (2022a, April 23). Google, meta, and others will have to explain their algorithms under new EU legislation. *The Verge*. https://www.theverge.com/2022/4/23/23036976/eu-digital-services-act-finalized-algorithms-targeted-advertising

Vincent, J. (2022b, March 24). EU targets Big Tech with sweeping new antitrust legislation. *The Verge*. https://www.theverge.com/2022/3/24/22994234/eu-antitrust-legislation-dma-digital-markets-act-details

Vizoso, A., Vaz-Álvarez, M., & López-García, X. (2021). Fighting deepfakes: Media and internet giants' converging and diverging strategies against hi-tech misinformation. *Media and Communication, 9*(1), 291–300. https://doi.org/10.17645/mac.v9i1.3494

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication, 37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, August 19–23, London. ACM, New York, Article 4. https://doi.org/10.1145/3219819.3219903.

Wardle, C., & Derakshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe report, DGI(2017)09. Retrieved 13 Apr 2022, from https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

Winston, B., & Winston, M. (2021). *The roots of fake news: Objecting to objective journalism*. Routledge.

Wittenberg, C., & Berinsky, A. J. (2020). Misinformation and its correction. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 163–198). Cambridge University Press.

YouTube. (2021). *Misinformation policies*. Retrieved 13 Apr 2022, from https://support.google.com/youtube/answer/10834785#zippy=%2Cmanipulated-content

Zello Staff. (2021, January 13). *Zello takes action against militias*. Zello. Retrieved April 13, 2022, from https://blog.zello.com/zello-takes-action-against-militias

CHAPTER 9

Strengthening the Civic Body as the Bandwidth for Optimised Emotion Expands

We now understand the nature of the complex, interconnected environment where communication and technologies operate to spread false information, impacting individuals and society. We diagnose that it is incubated especially by the *economics of emotion* (namely, the optimisation of datafied emotional content for financial gain) and the *politics of emotion* (namely, the optimisation of datafied emotional content for political gain). To reach this understanding, we integrated and shaped a wealth of literature from numerous disciplines on the deployment of false information, emotion, profiling and targeting. We illustrated this with case examples from across the world while reflecting on arising social and democratic harms to the *civic body* and multi-stakeholder solutions. Throughout, we have focused on global digital platforms, especially social media platforms, as these are the dominant purveyors of emotional AI globally today. Yet, far greater datafication of emotion is presaged worldwide through a plethora of more emergent emotional AI technologies. In this final chapter we draw out more substantive answers to strengthen the *civic body* as the bandwidth for the datafication, and optimisation, of emotion expands.

First, we tease out core shifts discernable from a backward glance. This allows us to identify that while false information, emotion, profiling and targeting are hardly new phenomena in citizen-political communications, the scale of contemporary profiling is unprecedented. As such, a prime site of concern is the automated industrial psycho-physiological profiling of

V. Bakir, A. McStay, *Optimising Emotions, Incubating Falsehoods*,
https://doi.org/10.1007/978-3-031-13551-4_9

the *civic body* to understand affect and infer emotion for the purposes of changing behaviour. Exploring this, we look to near-horizon futures. This is an important angle given the rapid onset, scale and nature of contemporary false information online; the rising tide of deployment of emotional analytics across all life contexts; and what we see as the greater role that biometrics will play in everyday life. Peeking over the horizon line allows us to distil our core protective principle of *protecting mental integrity*. This is necessary to strengthen the *civic body* to withstand false information in a future where optimised emotion has become commonplace.

Looking Backwards: Core Shifts

Reflecting on the rise of false, emotive information online, our chapters on false information (Chap. 4) and affect, emotion and mood (Chap. 5) highlight that such phenomena are enduring features of citizen-political communications, spanning thousands of years and attuned to shifts in media environments. Yet, if contemporary false information online simply makes use of classical propaganda techniques, why the current furore? The most obvious changes to the wider environment have been wrought by introduction of new forms of media, profiling techniques, systems that judge humans and their behaviour, and the search to monetise these phenomena. Indeed, the *scale* of contemporary profiling is unprecedented, notwithstanding the fact that profiling itself has a long history.

As shown in our discussion of adtech and corporate profiling in political communication (especially Chap. 6), the private sector led improvements in classification and quantification of populations using a panoply of approaches to identify audiences and record feedback. Originating in the USA over a century ago, and subsequently adopted by media owners internationally (McStay, 2011), this close monitoring of behaviour, consumption and geo-demography brought order to understanding of preferences, attitudes, civic feeling and disposition. Indeed, the fundamental principles of societal management and control through data had been essentially completed by the late 1930s through ad-testing, retail patterns, surveys and media engagement trends, among other data sources (Beniger, 1986).

Similarly, pre-empting automated real-time A/B testing used in commercial and political digital campaigning, key figures in the history and practice of advertising, such as Daniel Starch (1914) and Claude Hopkins (1998 [1923]), were insistent that advertising should be treated as a

science, using feedback to understand and identify those techniques that worked. The championing of datafied campaigning and voter profiling increasingly evident worldwide has discursive roots 100 years old. Indeed, while the feedback logics of contemporary false information online might be said to have neo-behaviourist characteristics, the earliest large advertising agencies (such as the J. Walter Thompson agency) were hiring behaviourist scientists to study their advertising, audiences and their states of mind, emotion and reactivity, in a systematic, data-first manner (McStay, 2011).

While we are wary of technological determinism, the use of technology does alter things, as witnessed through numerous seismic changes to mediated life. The printed press, radio, television, Internet, mobile telephony and their modalities of audience profiling are of profound importance. As we look towards the horizon line for how the media ecology might evolve, we regard as a prime site of concern the rise of emotional AI (McStay, 2018) and its psycho-physiological profiling of the *civic body* to understand affect and infer emotion.

Looking Forward: Near-Horizon Futures

Social media platforms developed and honed the practice of profiling and targeting individual desires and vulnerabilities, but they are now being joined by more emergent forms of emotional AI that are being trialled by governments worldwide as well as by globally dominant digital platforms themselves. When assisted by technologies that can turn human-state signals into fungible electronic data, identify patterns in small and large datasets, and apply and test rules from one situation to other situations and when this can be done increasingly cheaply, this provides for hitherto unseen scale. This portends nothing less than the *automated industrial psychology of emotional life*, one already attuned for changing behaviour. 'Emotional AI' claims to read and react to emotions through text, voice, computer vision and biometric sensing. This simulates understanding of human emotions via sensing words and images (such as sentiment analysis) and via sensing various bodily behaviours including facial expressions, gaze direction, gestures, voice, heart rate, body temperature, respiration and dermal electrical properties (McStay, 2018). Applicable machine learning and AI techniques deliver outputs that are named emotional states. These are then used for given purposes, such as predicting behaviour.

This is not at all far-fetched, and we do not seek to be alarmist or dystopian as a means of attracting attention. Instead, be this the wearable on our wrist, the cameras and microphones in our mobile phones, home digital assistants, in-car cameras and telematics, and more, their affect- and emotion-aware systems can provide not just novel means of engagement but profile us too (as introduced in Chap. 1). In addition to the human-technology touchpoints, we should also consider the commercial and political motives to better understand feeling and emotion (or at least being able to claim to do so), as elucidated in Chap. 2. Indeed, the body is already playing a role in political profiling, including testing with emotional AI and wider technologies, where ad-testers and political communications specialists use facial coding, electroencephalography (EEG) and other intimate means of analysis to assess bodies and brains for reactions to political messages and advertising. This entails reactions to propositions, types of attention, the role of contrasts and reactions to colour, music and narrative within a given ad (McStay, 2018). In this vein, it should not be missed that microtargeting in politics stems from technological 'innovation' in advertising, so it is reasonable to assume that political communicators will continue to utilise techniques from the commercial advertising sector.

Extending longstanding practices of sentiment analysis and classification of online emotion-type and disposition, we point to increasing inclusion of data about bodies. For example, Spotify (the world's largest music streaming service provider, with over 381 million monthly active users in 2021) has long profiled emotions and moods and has arrangements with advertising conglomerates (McStay, 2018). Signalling intention, Spotify's patent logged in 2021 to register taste attributes from audio signals is important, given the ubiquity of Spotify's service. Their goal is to improve speech-enabled recommender services by potentially simultaneously processing data from voice tone, speech content, background noise, user history, explicit indications of taste, profiles of Spotify user friends and environmental data from background noise. With this example alone, one easily sees how biometrics (through voice and speech) can begin to inform targeting processes for coming iterations of political advertising. Similar can be said for in-world profiling in Meta's foray into the metaverse (discussed further below). This will be dependent on physical profiling, especially of the face (via cameras or worn lenses with sensors around the mask), thereby rendering emotion expressions for in-world interactions (McStay, 2022).

Whether emotional AI technologies can deliver on their promises to accurately gauge human emotion has attracted much scholastic and industrial attention. Methodological flaws of determining emotions from biometrics (especially from facial coding) have particularly suffused this critique (McStay, 2019). For instance, Barrett et al. (2019) demonstrate in an authoritative meta-analysis that the 'basic emotions' approach that sees emotions as universal and informs much of the emotional AI industry fails to capture how people convey, or interpret, emotion on faces. Illustrating both accuracy and systemic racist bias, Rhue (2018), for example, compares emotional analysis components of Chinese face recognition company Megvii's Face++ software to Microsoft's Face Application Programming Interface when applied to a database of headshots of White and Black male professional basketball players in the USA. It finds that facial recognition software interprets emotions differently based on race, with Black players interpreted as angrier than White players by Megvii's Face++, and Microsoft interpreting Black players as more contemptuous when their facial expressions are ambiguous, compared to White players.

Yet, Barrett et al.'s (2019) damning and authoritative methodological critique of the use of facial coding to determine emotions also suggests solutions that engage more with context. Such context could be a 'cultural context, a specific situation, a person's learning history or momentary physiological state, or even the temporal context of what just took place a moment ago' (Barrett et al., 2019, p. 47). Indeed, industry leaders, such as Microsoft, are now advocating a turn to social context to more accurately gauge users' emotions. As signalled in the Spotify example above, McStay and Urquhart (2019) predict that this will inevitably involve a turn to *more* data so that the profiling analyst can know contextually more about a person and the scenario. In countries where profiling to infer sensitive attributes such as sexual orientation or political opinions is not well regulated, or where being of the 'wrong' sexuality or political tribe can be dangerous to life chances, or even to life itself, this increased optimisation of emotional life is alarming. Furthermore, we observe (and expand later in this chapter) that this sort of contextual data is precisely what globally dominant social media and technology platforms are very good at supplying through their profiling technologies. The suggestion, then, is not that biometric emotional AI will be foolproof (it will not be). Yet, in-house testing through biometric reactions, and potentially multimodal collection of biometric data about reactivity to stimuli, will make a significant difference to how *civic bodies* are understood, profiled, represented and targeted.

Despite methodological concerns, emotional AI is being used worldwide in a wide variety of governance contexts that impact the *civic body*. Its deployment for the purposes of governance varies according to different countries' societal goals, social organisation, and regulatory and cultural norms of privacy and agency. For instance, since 2016, in authoritarian United Arab Emirates, the smart city initiative of Smart Dubai uses sensors and analytics that feed a centralised monitoring and management layer to tell city analysts how residents, visitors, commuters and tourists feel about municipal matters, from transport to shopping and health. Presaged on opening personal data silos to the state, the Smart Dubai programme presents this as 'a globally unique, science-based approach to measuring and impacting people's happiness, fuelling the city's transformation' (McStay, 2018, p. 156). Notably, although Dubai's citizens have privacy rights, they constitute a small proportion of the overall population: residents and tourists have no such rights. Also noteworthy is that Dubai is well positioned to export its smart city model and emotional capture technologies globally (McStay, 2018). Not content with emotionally profiling populations, emotional AI is deployed worldwide to tell if we are lying. Fifty countries, including over 65 American law enforcement agencies and nearly 100 worldwide, already deploy US firm Converus' EyeDetect that uses software to track involuntary eye movements to detect lies (Lisbona, 2022, January 31). Universities are meanwhile developing lie detectors that rely on speech (content and tone of voice), body language and other physiological measures such as changes in facial muscle movements (Shuster et al., 2021). Although facial emotion expressions are far from universal, emotional AI technology companies have already sold facial recognition cameras across the world to surveil and police schools and cities (Article 19, 2021; McStay, 2018).

What then might the near future hold, and what does it portend for the spread of false information online? We consider three near-horizon futures as the bandwidth for profiled, datafied emotions expands.

Scenario 1: The Ministry of Optimised Moods

As a vehicle to consider connections between emotional AI technologies and the *civic body*, we could ask, 'What would political strategists such as Dominic Cummings make of them?' Cummings was a data-focused campaign strategist for Vote Leave, a campaigning organisation that, against all expectations, won the 2016 referendum campaign for Britain to leave

the European Union on what was regarded as a disinformation-heavy campaign. After Boris Johnson was elected UK Prime Minister in July 2019, Cummings was appointed to the new role of Chief Adviser to the Prime Minister. As COVID-19 ravaged the UK across 2020, Cummings was on hand to advise on adaptive strategies that (the government emphasised) followed the data and science (see Chap. 5).

Cummings embraces the role of data, engineering and management. In his blog, he proudly claims that Vote Leave innovated, 'the first web-based canvassing software that actually works properly in the UK and its integration into world-leading data science modelling to target digital advertising and ground campaigning' (Cummings, 2017, January 30). From the heart of the government, rather than relying on stories and authority, Cummings championed data-informed politics and novel modes of visualising complex information (across time, as well as contemporary complexity) to enhance decision-making. This includes a high-level interest in data and computer science, systems theory, psychology of persuasion, game theory, AI and machine learning, and the intersection of technology and storytelling. Cummings also champions sciences of prediction that are *dynamic* in nature (such as from weather forecasting and epidemiology), new technologies and interface design, difficult-to-control modern communications and cybernetic government (error-correction paths and prediction). His championing of interface design heavily draws on (and supports) Bret Victor, whose company, Dynamicland, builds computers and interfaces that people can *handle*. Cummings laments the UK government's Cabinet Room where important decisions are made without data-informed insight or dynamic representation of ongoing events and longitudinal trends. This contrasts with his enthusiasm for Dynamicland where computing (not just data representations) is embedded in surfaces of walls and objects. These new 'cognitive technologies' provide 'a new way of seeing and thinking' (Cummings, 2019a, 26 June). Cummings (2019a, 26 June) posits: 'Imagine discussing … possible post-Brexit trading arrangements with the models running like this for decision-makers to interact with'.

Beyond such 'Seeing Rooms' (Cummings, 2019a, June 26), given the rising tide of interest across society in emotional AI, it is not a stretch to see how citizen feeling might be modelled with multiple predictive scenarios of novel variables to consider outcomes and policies. The UK's Office for National Statistics (2021) already tracks national well-being data, but consider this dynamically visualised at granular levels in real time

in the Cabinet Room, using multiple sensors across cities, transport, workplaces, wearables, mainstream media and social media. One has to be careful not to overreach, but there is a clear appetite in being able to gauge the *civic body*, predict it (and its parts), know what the public will accept (such as restrictions on specific freedoms for the civic good) and use these insights to model public infrastructure initiatives. Arguably, before COVID-19, such datafication of the emotionalised *civic body* might have seemed unthinkable in liberal democracies, but COVID-19 has shown there to be keen appetite to know the public mood for governance purposes. As surveillance systems become even more normalised to protect the public and to police desired behaviour changes during pandemics, governments have a vested interest in understanding how the nation or specific groups are feeling, in order to hone targeted messages and other behavioural interventions and to cultivate a desired emotional state among the population (see Chap. 5). Add this to Cummings' interest in more intuitive forms of computing that facilitate new ways of doing politics, such as through 'neural interfaces' (Cummings, 2019a, June 26).

One might write off Cummings as an eccentric, someone who does not actually understand technology, as someone who misunderstands social complexity and the irreducibility of qualitative life to quantitative form. Dismissal ignores that these beliefs themselves matter given Cummings' prominent positions within UK politics and government at momentous times (the architect of the official Vote Leave referendum campaign and governing the UK during the first year of COVID-19, until leaving office in November 2020). Previous UK government advisors such as Alastair Campbell (Downing Street Press Secretary (1997–2000) and Downing Street Director of Communications and Strategy (2000–2003) for Prime Minister Tony Blair) perhaps belong to the age of news and rhetoric. By contrast, Cummings exists in a discourse of neuroscience, biohacks, datafication and predictive analytics. The test of whether this is a serious proposition is based in value: if there is deemed to be commercial or political value in optimising the mood of the *civic body* for the purposes of governing, it is a proposition that engaged citizenry should take seriously, however outlandish it may seem for liberal democracies. As a minimum, the convergence of emotion, commercial biometrics and politics is something that should be recognised and guarded against. Again, if the connection between emotional AI and political discourse seems too tenuous, we might remember that a central architect of Brexit, and advisor to the

British government, saw emotion and data as key to his successes, albeit in this case using online behavioural technologies built for advertising.

Turning from liberal democracies to the one-party state of China, of note is its 14th Five-Year Plan for National Informatisation. Aiming to promote innovation and application at scale of AI, it plans to 'launch cutting-edge intersectional research on artificial intelligence and basic disciplines such as neuroscience, cognitive science, psychology, social science' (Central Commission for Cybersecurity and Informatisation, 2021, December 28, p. 48). When married with its planned projects to experiment with AI for social governance purposes covering areas like public health, urban management, education and building 'social governance big data and virtual inference scientific research platforms' (p. 34), it is likely that emotional AI will play an increasing role in governance. Indeed, through experimenting with facial recognition technologies in schools and for policing, China has already started down this route, as observed by international human rights organisation, Article 19 (2021).

Scenario 1, then, is where the *civic body* is empathically optimised so that governments may better manage populations. It offers potential to be in touch with the disposition and emotional state of the *civic body* of one's country (or even that of another country). This scenario may appeal to those desiring more compliant populations (for instance, to instil prosocial public health behaviour during pandemics). However, those who prioritise individual agency above being dictated to by a wider, or leading, group are unlikely to view this scenario positively. The potential for honing disinformation by bad actors and for information warfare is also profound: it would super-charge the ability of an adversarial state or bad actor to achieve its goals by better understanding how to manipulate the emotions of targeted individuals or groups in other countries.

Scenario 2: Campaigns That Optimise Embodied Emotions

How would political or advocacy groups seeking to win elections or referenda, or promote their cause, behave in this brave new world of automated industrial psycho-physiological profiling of the *civic body*? Recent history shows sometimes psychopathic political levels of desire to win, willingness to break rules and to use all available data and new technologies to exploit psycho-emotionally sensitive points of the *civic body*. We posit that many campaigners would embrace this profiling to empathically optimise their messages to resonate with target audiences, regardless of

what social, cultural and technological norms are broken. Indeed, Chap. 4 already documents advocacy groups worldwide making powerful demands by putting words we want to hear into political leaders' mouths (such as apologising for failing to avert climate change) and resurrecting the dead (such as bringing back a murdered journalist to demand that state-backed violence against the press ends). Chapters 3, 5 and 6 document optimised emotive political campaigning and information warfare, where emotive, deceptive, microtargeted political campaigns have been offered, attempted or delivered, taking advantage of the affordances of social media and mobile apps. Chapter 6 highlights linguistically optimised deepfakes with politicians seeking to generate closer emotional connections with targeted voters by artificially, through AI, speaking their dialects.

Such profiling and targeting opportunities and claims continue to develop. Of note is recent research by Kosinski, given prior interest by (now defunct) political consultancy Cambridge Analytica in his work on Facebook 'Likes' to predict psychological characteristics and political inferences (see Chap. 6). Arguing that he is exposing societal threats rather than building new tools for harm, Kosinski (2021) claims that an open-source facial recognition algorithm can expose individuals' political orientation from a single naturalistic facial image taken from US Facebook profiles or from a popular dating website in the USA, UK and Canada. According to Kosinski, facial expression, self-presentation and facial morphology contain potential cues. For instance, in the US Facebook sample, Kosinski reports that liberals tend to face the camera more directly, are more likely to express surprise and are less likely to express disgust. Political orientation was correctly classified in 72% of liberal-conservative face pairs. Kosinski (2021) posits that even higher accuracy would likely arise from using higher resolution and multiple images per person; training custom neural networks aimed at political orientation; or including non-facial cues such as hairstyle. He also notes that even modestly accurate predictions can be impactful when applied to large populations in high-stakes contexts, such as elections. Unsurprisingly, given its biological deterministic bent, similar research by Kosinski (for instance, that AI can distinguish gay from straight people in photos (Wang & Kosinski, 2018)) has attracted stinging critiques, rightly invoking the racist and junk science of physiognomy, especially Kosinski's connecting of personality with facial morphology. From the point of view of physiognomy and the political *civic body*, warning from history could not be any louder, given the keen interest of Nazism in morphology, anthropometrics and physiognomy (Gray, 2004;

McStay, 2022). Regardless of whether Kosinski's research on AI's ability to expose political or sexual orientation from a facial image is realistic, that the question is being asked means political strategists and advocacy groups will be interested. This portends a direction of travel towards biometric profiling of the political *civic body*.

We also note that, beyond multiple emotional AI start-ups, several globally dominant companies already offer emotion recognition services based on analysis of facial expressions, including Microsoft, Amazon (Rekognition), Facebook, Apple and Google (Cloud Vision API) (McStay, 2018; Wright, 2021). Social media platforms already offer granular profiling and microtargeting tools to influence unsuspecting users, as Chap. 6 demonstrated. Their deployment of biometric emotion recognition services can only add further layers of granularity, and presumably, accuracy, to their suite of services for influence.

Manipulation of embodied emotions by political and advocacy groups is of particular concern where such groups engage in deceptive practices. For instance, deepfake synthetic media can elicit more emotional responses, as well as collapsing language barriers and reaching the illiterate (as deepfakes can deliver messages in any language or dialect that the deepfaker desires). While providing short-term wins for the campaigning group that has persuaded people by establishing greater personal connection, it is likely to further damage belief in the indexicality of the audiovisual image. Already, public figures are denying the authenticity of past incriminating video clips, allowing them to avoid accountability (see Chap. 4). If deepfakes, or the very idea of them, become more commonplace, then people will likely demand further proof of veracity, as seeing will no longer be believing. Given the biometric turn, this may involve biometric indicators to (a) prove that the campaigner is who they say they are and (b) that they mean what they are saying. This would represent a societal shift for would-be persuaders to 'prove' their authenticity (of self or message) by strapping themselves up to biometric lie detectors or other indicators of affect and emotion. An arms race, not just to increase citizens' digital literacy to spot false information but also to identify authenticity of emotions, and from that to infer the persuader's intent, may be on the near horizon too, despite concerns about the accuracy of such technology.

Scenario 2, then, is one where biometrics as a proxy for the *civic body*'s emotions are gauged so that campaigning groups can better connect with target audiences to influence votes, donations or behaviour. With the rise of machine learning on bodies and disposition, and as industry leaders

advocate a turn to ingesting and understanding social context (namely, wider forms of data) so that profiling analysts can know more about a person and the scenario, optimisation endeavours are likely to increase to be both more effective and affective. This lays the ground for undue influence and manipulation at important moments in the life of the *civic body*. This is of particular concern where campaigning groups engage in deceptive practices to achieve their aims.

Scenario 3: Profiting from Optimising Fellow-Feeling

As this book has demonstrated, emotional profiling is already deployed to manipulate us for profit by '*feeling-into*' online conversations and creating content and headlines on social media to resonate with, or trigger, specific groups within the *civic body* (see Chaps. 2 and 3). Furthermore, automated journalism can already automatically (with little human intervention beyond the initial programming phase) dig into reams of data to find patterns, such as using algorithms to sift through the leaked Panama Papers (Schapals & Porlezza, 2020); and it can offer insights to journalists on what the most important story element is (Cools et al., 2021). On top of this, the ability to automatically enable tone-optimised and geo-tailored news stories is already at hand for newsrooms willing to experiment. Using automated insights, algorithms can determine the emotional tone of a story and can tailor news stories for local audiences, for instance, on local sports results or local election outcomes, enabling highly personalised news feeds (Bakir & McStay, 2018; Graefe, 2016). Indeed, the phenomenon of *empathically optimised automated news* (of fake and real events alike) is on the near horizon, given the current state of automated journalism, sentiment analysis and language modelling.

To create empathically optimised automated fake news, the process would be to understand key trigger words and images among target groups; create fake news (itself normally comprising shorter and less informative content oriented towards disgust and anger [as discussed in Chap. 4]) and measure its engagement; and then have machines learn in an evolutionary capacity from this experience to create stories with more potency to increase engagement and thereafter advertising revenue (Bakir & McStay, 2018).

Should this appear unrealistic, consider the practices of Open AI, an American company whose mission is to ensure that artificial general intelligence (namely, highly autonomous systems that outperform humans at

most economically valuable work) benefits all of humanity (Open AI, 2022). In 2020, Open AI launched GPT-3 that uses deep learning to produce humanlike text. Within a year, over 300 applications were delivering GPT-3-powered search, conversation, text completion and other advanced AI features through their Application Programming Interface, involving tens of thousands of developers worldwide (Open AI, 2021, March 25). Such capacity has been noticed by political strategists. Dominic Cummings regularly wore an Open AI tee shirt and cites Open AI on his blog: for instance, how output from its large-scale unsupervised language model 'feels close to human quality' (Cummings, 2019b, March 1).

While one might counter that people would not be fooled by AI-generated text, this cannot be assumed. By way of illustration, Google engineer, Blake Lemoine, published transcripts in June 2022 that seemed to indicate that the AI chatbot generator system he was working on (Google's LaMDA (Language Model for Dialogue Applications)) had become sentient, with Lemoine claiming that it has the perception of, and ability to express thoughts and feelings equivalent to, a seven- or eight-year old human (Tiku, 2022, June 11). Google disagrees with Lemoine's assessment: LaMDA's abilities are based on pattern recognition rather than understanding meaning; and those familiar with chatbots can easily detect LaMDA's chatbot qualities, such as speaking in general ways that lack specificity, depth or originality (Ray, 2022, June 18). Reading the LaMDA transcripts (see Lemoine, 2022, June 11), if the reader has no awareness that the AI is using machine learning (transformer-based neural language models) to put the right words in the right order based on vast amounts of training data (trillions of words from the Internet) and the help of human crowd workers conscripted to engage in thousands of chats with the programme, the conversation looks convincingly humanlike.

Despite Open AI's and Google's stated commitments to Responsible AI, dangers to the *civic body* are in plain sight if it becomes impossible to distinguish human-generated text from AI-generated text. Google's research paper on LaMDA acknowledges that 'adversaries could potentially attempt to tarnish another person's reputation, leverage their status, or sow misinformation by using this technology to impersonate specific individuals' conversational style' (Thoppilan et al., 2022, p. 18). The architects of disinformation would surely add this tool to their arsenal if there is monetary or other gain to be made in doing so.

We have already seen how profiting from optimising fellow-feeling manifests throughout the contemporary disinformation supply chain. Money is made by digital influence mercenaries and trolls supplying false content (financed by propagandists or their clients); by creators of fake news websites (from associated online advertising on their sites); by clickbait-oriented news organisations (who earn money from more click-throughs of misleading headlines); and by the dominant digital platforms themselves (who sell profiles of engaged audiences to advertisers). Unfortunately, whistleblowing accounts detailed in Chap. 2 show that by designing algorithms that gave outsize weight to emotional Reactions and engaging posts, communities sharing false, extremist information were generated and consolidated on Facebook. Chapter 2 also observes that other social media platforms are similarly *emotional by design*, and Chap. 5 documents studies of the virality of emotional content on multiple social media platforms.

As an empirically grounded book, we have focused primarily on globally dominant digital platforms (especially social media); how their exploitation of datafied emotions maximises user engagement that can be monetised; and how this drives viral, false information. Looking to the future, however, the world's globally dominant social media platform, Facebook (rebranded as Meta in late 2021), is also turning to wider bandwidths of data collection, including biometrics. In late 2021 Mark Zuckerberg outlined plans for Meta as a metaverse company, a realisation of cyberspace where people move between virtual reality, augmented reality and familiar web-based platforms. Although the so-called metaverse is subject to much scepticism by well-placed commentariat, this would see the capacity for emotional profiling and targeting already afforded by social media platforms to connect with that afforded by biometrics. Keeping in mind that alongside Facebook, Instagram and WhatsApp, Meta also own Oculus (that produces virtual reality devices) and that they have long been researching in-world detection of emotion in virtual reality, one begins to discern Meta's direction of travel. As early as 2014, Zuckerberg regarded virtual reality as the next globally significant platform, capable of sharing precious, personal experiences (Levy, 2020, p. 328). Seven years later, Facebook Reality Labs Research predicted that virtual reality and augmented reality will 'become as universal and essential as smartphones and personal computers are today' and that they will involve 'optics and displays, computer vision, audio, graphics, brain-computer interface, haptic interaction, full body tracking, perception science, and true telepresence' (Tech@FACEBOOK, 2021, March 18).

This portends a profoundly granular control system built on an expanded bandwidth of data collection. As a minimum, in-world profiling will include data about facial expressions and reactivity stimuli and others (whether generated from desktop cameras or worn sensors, such as around a virtual reality head unit mask tracking muscle movement). Neural input technology is steadily moving towards everyday experience, such as Facebook's wristband that uses haptics to measure hand and finger gesture (Tech@FACEBOOK, 2021, March 18). As such, there is clear scope for ocular- and affect-based interactions to create and track engagement with virtual objects. Meta, of course, is not the only company seeking to realise long-promised visions of the neuro-enhanced 'human-machine', but unlike start-up companies such as Elon Musk's Neuralink that is developing brain chips, Meta has global scale. The significance of augmentation is the scope to sense and measure, or *feel-into*, electrical impulses (such as through electromyography) in the body to gauge human intention.

Scenario 3, then, is one where individuals and companies profit by *feeling-into* the *civic body* and creating content to resonate with specific groups to increase engagement and thereafter advertising revenue. This has already proven lucrative to the architects of disinformation across *emotional by design* social media platforms. The nature of future instantiations of profiting by *feeling-into* the *civic body* is not at all clear given hype and the diverse technologies and practices in play, but we foresee a near-horizon future where citizens' online and offline behaviour is registered by much more granular means, representing a biometric future for communication with and through the *civic body*. That we may be turned into perpetually targeted data pools to be exploited and managed by architects of disinformation and influence is not a scenario that accords with one of human dignity and flourishing.

Protecting Citizens in the Coming Era of Optimised Emotions

That citizens could be more intensely emotionally profiled and targeted for manipulation by individuals, pressure groups, companies, political parties, governments and nation-states has raised concerns at the highest of levels. Published attention sharpened in 2021, for example, with the United Nations Committee on the Rights of the Child publishing 'General Comment 25' that addresses children's rights in the digital age. This contains multiple mentions of emotion analytics (see §42, 62, 68), finding

them to interfere with children's right to privacy, freedom of thought and belief. It also flags the importance 'that automated systems or information filtering systems are not used to affect or influence children's behaviour or emotions or to limit their opportunities or development' (United Nations Convention on the Rights of the Child, 2021, March 2, §62). Moreover, also in 2021, the United Nations Human Rights Council formally adopted the Resolution titled 'Right to privacy in the digital age' where §3 notes need for safeguards for emotion recognition (United Nations General Assembly, 2021). The Council of Europe (2021) likewise called for strict limitations and bans regarding emotion profiling in areas of education and the workplace. Also in 2021, the European Data Protection Board and the European Data Protection Supervisor issued a joint statement declaring use of AI to infer emotions of a natural person as highly undesirable and that it should be prohibited, except for specified cases, such as some health purposes (European Data Protection Board, 2021). Related, 2021 also saw the release of a draft of the proposed European Union AI Act, a risk-based piece of legislation that classifies emotion recognition as both risky and high risk, depending on the use case (European Commission, 2021, April 21).

Indeed, beyond interest in emotion recognition systems, the proposed European Union AI Act is unequivocal about the need to protect against the capacity of AI (especially that using biometric data) for *undue influence* and *manipulation*. To create an ecosystem of trust around AI, its proposed AI regulation bans use of AI for manipulative purposes; namely, that 'deploys subliminal techniques … to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm' (European Commission, 2021, April 21, Title II Article 5, p. 43). While it is not yet clear what current applications this might include, it is highly likely to cater for neural and in-world environmental manipulation of the sort that would be facilitated if Meta and Neuralink's developments are realised.

Furthermore, in April 2022, proposed amendments to the draft AI Act included the proposal from the Committee on the Internal Market and Consumer Protection, and the Committee on Civil Liberties, Justice and Home Affairs, that 'high-risk' AI systems should include AI systems used by candidates or parties to influence, count or process votes in local, national or European elections (to address the risks of undue external interference and of disproportionate effects on democratic processes and democracy). Also proposed as 'high risk' are machine-generated complex

text such as news articles, opinion articles, novels, scripts and scientific articles (because of their potential to manipulate, deceive or expose natural persons to built-in biases or inaccuracies) and deepfakes representing existing persons (because of their potential to manipulate the natural persons that are exposed to those deepfakes and harm the persons they are representing or misrepresenting) (European Parliament, 2022, April 20, Amendments 26, 27, 295, 296, 297). Classifying them as 'high risk' would mean that they would need to meet the Act's transparency and conformity requirements before they could be put on the market: these requirements, in turn, are intended to build trust in such AI systems.

Mindful that people have generally low awareness about emotion profiling, since 2015, we at the Emotional AI Lab have carried out studies into the British public's views on established and emergent emotional AI use cases. Our recent survey shows that a majority of British adults dislike use of emotional AI technologies where there is capacity for *undue influence* in situations that they are *powerless to control* and where it affects *important moments in civic life or a person's own life chances.* This demographically representative omnibus online survey (n > 2000 adults conducted in January 2020 by ICM Unlimited) explores levels of concern about five use cases for emotion-sensing technologies in everyday life (see Table 9.1). Of these, it finds that people are most concerned about social media profiling in political campaigns utilised to find out which political ads or messages are most engaging for specific audiences and to personalise and target what political ads we see (66% are 'not OK' with any form of such data collection). A majority (58%) are also concerned about biometrics in the workplace to track employees' emotions. A small majority are concerned about biometrics in schools to track students' facial expressions to work out their emotional states and attention levels in order to tailor teaching. Large minorities are concerned about automated understanding of the emotional and affective behaviour of drivers (45%) and usage in out-of-home advertising to gauge reactivity to ads (45%) (see Table 9.1). We include this survey snapshot because it is notable that people appear to be more concerned about undue political influence and manipulation though social media in politics than biometric profiling. Given sensitivities around the body especially in relation to questions of workplaces, this finding was unexpected.

It remains to be seen if and how uses of emotion recognition will scale and whether seemingly low-stake emotional AI interactions with people (such as via outdoor ads and in cars) will increasingly feature without

Table 9.1 UK adult attitudes to different form of emotional profiling

'Not OK' with any form of data collection	*Where?*	*How?*	*By whom?*	*For what purpose?*
66%	Social media	Profiles of social media posts	Political ad companies	To find out which political ads/messages are most engaging for specific audiences and to personalise and target what political ads we see
58%	Workplace	Sentiment of emails and social media. Cameras to record facial expressions, gesture and behaviour. Audio recorders to measure voice. Wearables	Employers	To track employees' emotions
52%	School classroom	Cameras to track students' facial expressions	Ed tech companies or schools	To work out students' emotional states and attention levels and to tailor teaching
45%	Car	Sensors to detect stress, anger or frustration	Car manufacturers	To understand drivers' emotional behaviour, to monitor fatigue and distraction and to personalise driver experience
45%	Outdoor spaces	Cameras in outdoor ads	Advertising agencies	To scan onlookers' facial expressions to work out their emotions towards an outdoor ad, so that the ad changes itself to be more appealing

Source: ICM Unlimited UK-based survey, $n > 2000$ adults, January 2020

significant societal pushback (see McStay & Urquhart, 2022). For now, people (at least in the UK) are clearly not keen on higher stake emotional AI interactions (such as for political influence or that affect the workplace

and schools). One might safely wager that people would not be 'OK' with emotion-based biometric insights from their engagements with devices being used for political purposes, such as data generated by longitudinal profiling of interaction with home voice assistants, facial expression data collected by phones, or in-world tracking of emotion and behaviour. In addition to well-known problems of embedded values and biases in socio-technical systems, and the methodological and conceptual flaws of emotional AI technologies (AI Now Institute, 2018; McStay, 2018; Russell, 1994; Stark & Hutson, 2021), we suggest a need for greater recognition of the potential for biometric profiling to spill into political profiling. This recognition would alert us to the need not just for individual protections but also for those of a collective and civic sort. This would involve being alert to organisational justifications for aggregation of biometric affinity data, where profiling does not occur directly but through people's 'affinity' with a group defined by such data (Wachter, 2020) and their biometrics and reaction types. The consequence of this is that, paradoxically, while the data points collected about a person may be relatively few, when they are assembled alongside indirect inferences and assumed dispositions, profiling and targeting becomes, and may feel, much more personal.

Mindful that proposed transparency obligations, bans on undue influence and specification of what is deemed 'high risk' may be diluted via lobbying before the European Union AI Act is passed, we also note the increasing clamour for human-centric design for emotional AI and empathic technologies by industry critics, the basic tenet of which is to design to benefit humankind rather than to exploit it (Institute of Electrical and Electronics Engineers, 2019; McNamee, 2019; McStay & Pavliscak, 2019). Yet, seen most charitably, as this book has shown, companies cannot always foresee, nor are prepared to adequately remedy, real-world harmful uses of their technologies, especially if such remedies damage their engagement-driven business model (as evident in the case of false information and digital platforms). As such, global technology standards board, the Institute of Electrical and Electronics Engineers (IEEE), comprising multinational volunteers from academia, industry and government, formed the IEEE P7014 Working Group in 2019 to try to standardise the ethical design associated with empathic technologies and its tools, frameworks and processes (Soper et al., 2020). However, while useful as a means of identifying and promoting good behaviour, adherence to standards is voluntary, so lacking force of law. Mindful of ongoing legislative activity and weakness in technological standards-based

initiatives to protect citizens in the coming era of optimised emotions, we try to crystallise the social problem: the *need to protect mental integrity.*

Protecting Mental Integrity

In the 2021 Reith Lectures, AI expert Stuart Russell observes the pressing need to protect our 'mental integrity' (a right in the Charter of Fundamental Rights of the European Union (European Union, 2012, Article 3)) from the profiling and predictive capacities of AI (Russell, 2021). Neuro-ethicist, Andrea Lavazza (2018), defines mental integrity as an 'individual's mastery of his [sic] mental states and his [sic] brain data so that, without his [sic] consent, no one can read, spread, or alter such states and data in order to condition the individual in any way'. While Lavazza is concerned to protect mental integrity from devices capable of directly interfering with it, such as brain implants and neuro-prosthesis, McStay (2022) urges that we should be similarly concerned with plans, models, processes and potentially ubiquitous systems that seek to automate empathy. This includes emotional AI tools that monitor and condition human emotion.

Emotional AI technologies claim to be able to gauge human emotions for the purposes of influencing, predicting and controlling human behaviour. Yet, if these technologies were judged in human terms, they would be considered psychopathic (McStay, 2022). Despite being marketed under the auspices of empathy and sensitivity to emotion, they do not actually understand our emotions: they only process signals (such as biometrics) and predict outputs (named emotional states). The judgements of emotional AI may display deeply cold-hearted behaviour (such as playing on an audience's fears to maximise engagement with specific content). Ultimately, our relationships with them will be inauthentic and fake (such as deepfaking a political actor's dialect to establish closer connections with target electorates).

To be subjected to profiling by emotional AI systems, we argue, is not just psychopathic but also highly invasive. Emotional AI clearly does not 'understand' first-person outlooks, the phenomenology, or lifeworld of the individual. However, that it can discern and predict proxies of mental life to some degree should raise concerns about human privacy and dignity. If such systems become anywhere near as accurate as their developers claim, we would be stripped of our privacy and dignity as our inner life and feelings would be exposed and mined. As Alegre (2021, May, p. 4) puts it,

we must rapidly work out where we draw the line 'between what we choose to reveal about ourselves and what is being unlawfully inferred about the absolutely protected space inside our heads'.

Beyond the individual, what of *collective mental integrity*? *Feeling-into* the collective may well be useful to optimise societal moods and behaviour change in time of national emergency (such as pandemics). More collectivist societies, such as China, may prefer a more permanent arrangement of *feeling-into* their society, in the name of social cohesion, order and harmony. For them, the social good of such emotional optimisation may outweigh the social harms of an overzealous surveillance state, including its chilling effects on freedom of thought, expression and association. However, such emotional optimisation capabilities can be abused by bad actors, not least hostile states conducting information warfare on unsuspecting populations by fomenting division, dissent and ontological insecurity. Furthermore, if freedom of thought is a fundamental human right that underpins all other human rights (as argued in Chap. 7), then this should lead even collectivist societies to step back from endeavours to optimise the datafied emotions of their collective.

Whether at the macro-level (such as protecting elections or health drives) or at the micro-level (such as protecting an individual's freedom to privately think and feel whatever they like without interference), the *civic body* across the world is highly exposed to attempts at undue influence. We suggest that the principle of *protecting mental integrity* can be applied by individualistic societies (such as the USA) and collectivist societies (such as China) alike. Whether it is individual or collective mental integrity that is prioritised by governments, we argue that both are necessary to protect the *civic body*.

The Last Word

In dissecting how emotions are optimised to fuel contemporary false information online, we have reached an understanding of the twin incubators of the *politics of emotion* and the *economics of emotion*; the harms to the *civic body* that have ensued; and the many solutions proposed by diverse stakeholders. Yet, society has yet to tackle the false information media ecology head on as the underpinning business model driving it on social media remains intact. We suggested in Chap. 8 that all other solutions are merely tinkering at the edges.

As emotional AI expands from being the purview mainly of globally dominant social media platforms to a wide range of biometrically oriented forms, we see far greater potential for manipulation and exploitation of the *civic body*. If we are still not prepared to combat global disinformation and misinformation, we are far from ready for the coming era of emotional AI. This chapter outlined three near-horizon futures emanating from the coming *automated industrial psycho-physiological profiling* of the *civic body* to understand affect and infer emotion for the purposes of changing behaviour. None of them are without concern.

Scenario 1, where the *civic body* is empathically optimised so that governments can better manage populations, will concern those who prioritise individual agency above being dictated to by a wider, or leading, group. It will also raise concerns about its enhanced potential for information warfare where an adversarial state or group manipulates the emotions of citizens in its target country.

Scenario 2, where the *civic body* is empathically optimised so that campaigning groups can better connect with their target audiences to influence votes, donations or behaviour, will raise concerns in countries where profiling is poorly regulated and where campaigning groups engage in deceptive practices to achieve their aims. With the rise of machine learning ingesting wider forms of data, the accuracy of such optimisation is likely to increase, and with this, manipulation of the *civic body*.

Scenario 3, where individuals and companies profit by *feeling-into* the online and offline behaviour of the *civic body*, raises the spectre of perpetual surveillance that is perhaps tempting for some (e.g. via the metaverse). However, it will be difficult to resist given the coming ubiquity of smart and augmented environments and the difficulty of fooling context-aware, affect-based recognition tools utilised by complex assemblages of actors that may be monitoring emotions in public spaces. That we may be turned into perpetually targeted data pools does not accord with principles of human dignity and flourishing.

To prevent the perpetuation or intensification of false information as the global *civic body* becomes increasingly awash with datafied, optimised emotion, urgent preventative action is needed. Although emotional AI has raised concerns in the United Nations, and is achieving regulatory attention in the European Union, elsewhere AI and data privacy are far less regulated and deserve immediate attention to protect their citizens and those of other countries (for instance, from information warfare). Failure

to do so will leave the world unprotected from manipulative emotional profiling for commercial and political ends.

The European Union's draft AI regulatory proposals of avoiding undue influence and promoting greater transparency are a good place to start to avoid the harms that may arise where the granularity of online emotional profiling spills offline and becomes the everyday, resigned-to and mundane. However, we propose that this should be underpinned by the principle of *protecting mental integrity*, *both individual and collective*. As demonstrated by the ecology of false information, if the business model pushed by the emotional AI industry is one that exploits our emotions to maximise user engagement, then the battle to ensure that emotional AI is not used for harm will be an uphill one.

For the principle of *protecting mental integrity* to take root across the global *civic body* will require simultaneous effort across stakeholders. This embraces regional prosocial policymakers, ethically minded technologists, innovators, standards bodies and other international policy influencers, through to educators. And as citizens, we should be prepared to learn about the perils, as well as promises, of an emotionally datafied and optimised world. We hope this book helps in this task.

References

AI Now Institute. (2018). *AI now report*. New York University. Retrieved 13 Apr 2022, from https://ainowinstitute.org/AI_Now_2018_Report.pdf

Alegre, S. (2021, May). *Protecting freedom of thought in the digital age*. Policy Brief No. 165. Centre for International Governance Innovation. Retrieved 13 Apr 2022, from https://www.cigionline.org/publications/protecting-freedom-of-thought-in-the-digital-age/

Article 19. (2021). *Emotional entanglement: China's emotion recognition market and its implications for human rights*. Retrieved 13 Apr 2022, from https://www.article19.org/wp-content/uploads/2021/01/ER-Tech-China-Report.pdf

Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism, 6*(2), 154–175. https://doi.org/10.1080/21670811.2017.1345645

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Beniger, J. R. (1986). *The control revolution: Technological and economic origins of the information society*. Harvard University Press.

Central Commission for Cybersecurity and Informatization. (2021, December 28). *14th five-year plan for national informatization*. Retrieved 27 Apr 2022, from https://digichina.stanford.edu/work/translation-14th-five-year-plan-for-national-informatization-dec-2021/

Cools, H., Van Gorp, B., & Opgenhaffen, M. (2021). When algorithms recommend what's new(s): New dynamics of decision-making and autonomy in newsgathering. *Media and Communication, 9*(4), 198–207. https://doi.org/10.17645/mac.v9i4.4173

Council of Europe. (2021). *Consultative committee of the convention for the protection of individuals with regard to automatic processing of personal data*. Retrieved 13 Apr 2022, from https://rm.coe.int/guidelines-on-facial-recognition/1680a134f3

Cummings, D. (2017, January 30). *On the referendum #22: Some basic numbers for the Vote Leave campaign*. Retrieved 13 Apr 2022, from https://dominiccummings.com/2017/01/30/on-the-referendum-22-some-numbers-for-the-vote-leave-campaign/

Cummings, D. (2019a, June 26). On the referendum #33: High performance government, 'cognitive technologies', Michael Nielsen, Bret Victor, & 'Seeing Rooms'. *Dominic Cummings's blog*. Retrieved 13 Apr 2022, from https://dominiccummings.com/2019/06/26/on-the-referendum-33-high-performance-government-cognitive-technologies-michael-nielsen-bret-victor-seeing-rooms/

Cummings, D. (2019b, March 1). On the referendum #31: Project Maven, procurement, lollapalooza results & nuclear/AGI safety. *Dominic Cummings's blog*. Retrieved 13 Apr 2022, from https://dominiccummings.com/tag/openai/

European Commission. (2021, April 21). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*. Brussels, COM(2021) 206 final 2021/0106 (COD). Retrieved 13 Apr 2022, from https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence. Accessed 21 Sept 2021.

European Data Protection Board. (2021). *Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Retrieved 13 Apr 2022, from https://edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf

European Parliament. (2022, April 20). *Draft report 2021/0106(COD) on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending*

certain Union Legislative Acts (COM2021/0206 – C9-0146/2021 – 2021/0106(COD)). Committee on the Internal Market and Consumer Protection Committee on Civil Liberties, Justice and Home Affairs. 2021/0106(COD). Retrieved April 26, 2022, from https://www.europarl.europa.eu/doceo/document/CJ40-PR-731563_EN.pdf

European Union. (2012). Charter of fundamental rights of the European Union. *Official Journal of the European Union*. C 326/391. Retrieved 13 Apr 2022, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT&from=RO

Graefe, A. (2016). *Guide to automated journalism*. Tow Centre for Digital Journalism. Retrieved 13 Apr 2022, from https://www.cjr.org/tow_center_reports/guide_to_automated_journalism.php

Gray, R. T. (2004). *About face: German physiognomic thought from Lavater to Auschwitz*. Wayne State University Press.

Hopkins, C. (1998). *Scientific advertising*. Moore (Original work published 1923).

Institute of Electrical and Electronics Engineers. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*. Retrieved 13 Apr 2022, from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf

Kosinski, M. (2021). Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports, 11*(100). https://doi.org/10.1038/s41598-020-79310-1

Lavazza, A. (2018). Freedom of thought and mental integrity: The moral requirements for any neural prosthesis. *Frontiers in Neuroscience, 12*(82). https://doi.org/10.3389/fnins.2018.00082

Lemoine, Blake. (2022, June 11). Is LaMDA sentient? – an interview. *Medium*. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

Levy, S. (2020). *Facebook: The inside story*. Penguin.

Lisbona, N. (2022, January 31). True story? Lie detection systems go high-tech. *BBC News*. https://www.bbc.co.uk/news/business-60153129

McNamee, R. (2019). *Zucked: Waking up to the Facebook catastrophe*. Harper Collins.

McStay, A. (2011). *The mood of information: A critique of online behavioural advertising*. Continuum.

McStay, A. (2018). *Emotional AI: The rise of empathic media*. Sage.

McStay, A. (2019). Emotional AI and edtech: Serving the public good? *Learning, Media and Technology, 45*(3), 270–283. https://doi.org/10.1080/17439884.2020.1686016

McStay, A. (2022, in press). *Automating empathy: When technologies claim to feel-into everyday life*. Oxford University Press.

McStay, A. & Pavliscak, P. (2019). *Emotional artificial intelligence: Guidelines for ethical use*. Emotional AI Lab and Changesciences. Retrieved 13 Apr 2022,

from https://drive.google.com/file/d/1frAGcvCY_v25V8ylqgPF2brTK9UVj_5Z/view

McStay, A., & Urquhart, L. (2019). 'This time with feeling?' Assessing EU data governance implications of out of home appraisal based emotional AI. *First Monday, 24*(10–7). https://doi.org/10.5210/fm.v24i10.9457

McStay, A., & Urquhart, L. (2022). In cars (are we really safest of all?): Interior sensing and emotional opacity. *International Review of Law, Computers & Technology*. Advance online publication. https://doi.org/10.1080/13600869.2021.2009181

Office for National Statistics. (2021). *Well-being*. Retrieved 13 Apr 2022, from https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing

Open AI. (2021, March 25). *GPT-3 powers the next generation of apps*. Retrieved 13 Apr 2022, from https://openai.com/blog/gpt-3-apps/

Open AI. (2022). *About*. Retrieved 13 Apr 2022, from https://openai.com/about/

Ray, T. (2022, June 18). Sentient? Google LaMDA feels like a typical chatbot. *ZNet*. https://www.zdnet.com/article/match-any-color-you-like-instantly-with-a-mini-color-sensor-for-only-84/

Rhue, L. (2018). Racial influence on automated perceptions of emotions. *SSRN*, November. Retrieved 13 Apr, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281765

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin, 115*(1), 102–141. https://doi.org/10.1037/0033-2909.115.1.102

Russell, S. (2021). The Reith Lectures – Stuart Russell – living with artificial intelligence – AI: A future for humans. *BBC Sounds*. https://www.bbc.co.uk/sounds/play/m0012q21

Schapals, A. K., & Porlezza, C. (2020). Assistance or resistance? Evaluating the intersection of automated journalism and journalistic role conceptions. *Media and Communication, 8*(3), 16–26. https://doi.org/10.17645/mac.v8i3.3054

Shuster, A., Inzelberg, L., Ossmy, O., Izakson, L., Hanein, Y., & Levy, D. J. (2021). Lie to my face: An electromyography approach to the study of deceptive behaviour. *Brain and Behavior, 11*(12), Article e2386. https://onlinelibrary.wiley.com/doi/10.1002/brb3.2386

Soper, R., Bennet, K., Rivas, P., & Mathana (2020). Developing use cases to support an empathic technology ethics standard. *2020 IEEE International Symposium on Technology and Society (ISTAS)*, pp. 25–28, doi: https://doi.org/10.1109/ISTAS50296.2020.9462177.

Starch, D. (1914). Advertising: Its principles, practice, and technique, *Internet Archive*, Retrieved 13 Apr 2022, from http://www.archive.org/download/advertisingitspr00stariala/advertisingitspr00stariala.pdf

Stark, L., & Hutson, J. (2021). *Physiognomic artificial intelligence*, *SSRN*. Retrieved 13 Apr from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3927300

Tech@FACEBOOK. (2021, March 18). *Inside Facebook reality labs: Wrist-based interaction for the next computing platform*. Retrieved 13 Apr 2022, from https://tech.fb.com/inside-facebook-reality-labs-wrist-based-interaction-for-the-next-computing-platform/

Thoppilan, R., De Freitas, D., Hall, J. et al. (2022). *LaMDA: Language Models for Dialog Applications*. Retrieved 13 Sep 2022, from https://doi.org/10.48550/arxiv.2201.08239

Tiku, N. (2022, June 11). The Google engineer who thinks the company's AI has come to life. *The Washington Post*. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

United Nations Convention on the Rights of the Child. (2021, March 2). *General comment No. 25 (2021) on children's rights in relation to the digital environment*. UN Doc CRC/C/GC/25. Retrieved 13 Apr 2022, from https://docstore.ohchr.org/SelfServices/FilesHandler.ashx?enc=6QkG1d%2FPPRiCAqhKb7yhsqIkirKQZLK2M58RF%2F5F0vEG%2BcAAx34gC78FwvnmZXGFUl9nJBDpKR1dfKekJxW2w9nNryRsgArkTJgKelqeZwK9WXzMkZRZd37nLN1bFc2t

United Nations General Assembly. (2021). *Resolution adopted by the Human Rights Council on 7 October 2021*. Retrieved 13 Apr 2022, from https://undocs.org/A/HRC/RES/48/4

Wachter, S. (2020). Affinity profiling and discrimination by association in online behavioural advertising. *Berkeley Technology Law Journal, 35*(2) Available at: papers.ssrn.com/sol3/papers.cfm?abstract_id=3388639

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114*(2), 246–257. https://doi.org/10.1037/pspa0000098

Wright, J. (2021). Suspect AI: Vibraimage, emotion recognition technology and algorithmic opacity. *Science, Technology & Society*, 1–20. Advance online publication. https://doi.org/10.1177/09717218211003411

Index

V. Bakir, A. McStay, *Optimising Emotions, Incubating Falsehoods*,
https://doi.org/10.1007/978-3-031-13551-4