Citizens Versus the Internet: Confronting Digital Challenges With Cognitive Tools

Anastasia Kozyreva\textsuperscript{1*}, Stephan Lewandowsky\textsuperscript{2,3}, & Ralph Hertwig\textsuperscript{1}

\textsuperscript{1}Center for Adaptive Rationality, Max Planck Institute for Human Development
\textsuperscript{2}School of Psychological Science, University of Bristol
\textsuperscript{3}School of Psychological Science, University of Western Australia

*Corresponding author

Contact: Anastasia Kozyreva, Max Planck Institute for Human Development, Lentzeallee 94, 14195, Berlin, Germany. E-mail: kozyreva@mpib-berlin.mpg.de

Acknowledgements: The authors are grateful to Gerd Gigerenzer, Stefan Herzog, Philipp Lorenz-Spreen, Sam Wineburg, and Mark Leiser for their comments and suggestions. We also thank Deb Ain for editing the manuscript.
**Abstract:** The Internet has evolved into a ubiquitous digital environment in which people communicate, seek information, and make decisions. Online environments are replete with smart, highly adaptive choice architectures designed primarily to maximize commercial interests, capture and sustain users’ attention, monetize user data, and predict and influence future behavior. This online landscape holds multiple negative consequences for society, such as a decline in human autonomy, rising incivility in online conversation, the facilitation of political extremism, and the spread of disinformation. Benevolent choice architects working with regulators may curb the worst excesses of manipulative choice architectures, yet the strategic advantages, resources, and data remain with commercial players. One way to address this imbalance is with interventions that empower Internet users to gain some control over their digital environments, in part by boosting their information literacy and their cognitive resistance to manipulation. Our goal is to present a conceptual map of interventions that are based on insights from psychological science. We begin by systematically outlining how online and offline environments differ despite being increasingly inextricable. We then identify four major types of challenges that users encounter in online environments: persuasive and manipulative choice architectures, AI-assisted information architectures, distractive environments, and false and misleading information. Next, we turn to how psychological science can inform interventions to counteract these challenges of the digital world. After distinguishing between three types of behavioral and cognitive interventions—nudges, technocognition, and boosts—we focus in on boosts, of which we identify two main groups: (1) those aimed at enhancing people’s agency in their digital environments (e.g., self-nudging, deliberate ignorance) and (2) those aimed at boosting competences of reasoning and resilience to manipulation (e.g., simple decision aids, inoculation). These cognitive tools are designed to foster the civility of online discourse and protect reason and human autonomy.
against manipulative choice architectures, attention-grabbing techniques, and the spread of false information.

**Key words:** algorithms, artificial intelligence, attention economy, decision autonomy, behavioral policy, boosting, choice architecture, cognitive tools, decision aids, digitalization, disinformation, false news, Internet, nudging, online environments, online manipulation, online reasoning, self-nudging, technocognition.

Word count (excluding figures): 26,900
COGNITIVE TOOLS FOR THE DIGITAL WORLD

Contents

Contents ......................................................................................................................................................... 4
1. The Role and Responsibility of Psychological Science in the Digital Age................................. 6
2. Systematic Differences Between Online and Offline Environments ........................................ 11
   2.1. Differences in Structure and Functionality ..................................................................................... 14
   2.2. Differences in Perception and Behavior ......................................................................................... 18
3. Challenges in Online Environments .................................................................................................... 24
   3.1. Persuasive and Manipulative Choice Architectures ........................................................................ 24
   3.2. AI-Assisted Information Architectures ................................................................................................. 29
   3.3. Distractive Environments ................................................................................................................. 34
   3.4. False and Misleading Information ..................................................................................................... 39
4. Behavioral Interventions Online: Nudging, Technocognition, and Boosting ............................ 48
5. Boosting Cognitive Competences in Online Environments ............................................................ 55
   5.1. Self-Nudging: Boosting Control Over One’s Digital Environment .................................................. 56
   5.2. Deliberate Ignorance as Information Management Device ............................................................ 62
   5.3. Simple Decision Aids: Boosting Digital Information Literacy ......................................................... 66
   5.4. Inoculation: Boosting Cognitive Resilience to Misinformation and Manipulation ....................... 71
6. Conclusion: From Psychological Science to the Internet for Citizens ............................................. 74
References ..................................................................................................................................................... 80

Tables and figures

Table 1 ......................................................................................................................................................... 11
Table 2 ......................................................................................................................................................... 14
Table 3 ......................................................................................................................................................... 24
Table 4 ......................................................................................................................................................... 55
Table 5 ......................................................................................................................................................... 72

Figure 1. Entry points for policy interventions in the digital world. ......................................................... 7
Figure 2. Categories and types of dark patterns. ....................................................................................... 27
Figure 3. Examples of AI-assisted information architectures online. ..................................................... 30
Figure 4. Four classes of schedules of reinforcement. ............................................................................. 37
Figure 5. Types of false and misleading information in the digital world. ............................................. 41
Figure 6. Sources and strategies of false and misleading information in the digital world. ...42
Figure 7. Types of behavioral and cognitive interventions for the digital world. ...............48
Figure 8. Self-nudging interventions in online environments. ........................................60
Figure 9. Simple rules for online reasoning. ..................................................................68
Figure 10. “Can you trust this information?” .................................................................70
Figure 11. Bad News game .............................................................................................73
Figure 12. Map of challenges and boosts in the digital world........................................77
1. The Role and Responsibility of Psychological Science in the Digital Age

In 1969, the year Neil Armstrong became the first person to walk on the moon, the Internet—then known as ARPANET—was brought online. The first host-to-host message was sent from a computer at UCLA to a computer at Stanford University and it read “lo.” The network crashed before the full message, “login,” could be transmitted. Fast forward half a century from this first step into cyberspace, and the Internet has evolved into a ubiquitous global digital environment, populated by more than 4 billion people\(^1\) and entrenched in nearly all aspects of their professional, public, and private lives.

The evolution of digital technologies has given rise to possibilities that were largely inconceivable in 1969, such as instant worldwide communication, a mostly unfettered and constant access to information, democratized production and dissemination of information and digital content, and the ability to coordinate global political movements. But as the popular adage goes, there is no such thing as a free lunch. Digital technology has also introduced challenges that imperil the well-being of individuals and the functioning of democratic societies, such as the rapid spread of false information and online manipulation of public opinion (e.g., Bradshaw & Howard, 2019; Kelly, Truong, Shahbaz, Earp, & White, 2017), as well as new forms of social malpractice such as cyberbullying (Kowalski, Giumetti, Schroeder, & Lattanner, 2014) and online incivility (Anderson, Brossard, Scheufele, Xenos, & Ladwig, 2014). Moreover, the Internet is no longer an unconstrained and independent cyberspace but, notwithstanding appearances, a highly controlled environment. Online, whether people are accessing information through search engines or social media, their access is regulated by algorithms and design choices made by corporations in pursuit of profits and with little transparency or public oversight. Government control over the Internet

---

\(^1\) See International Telecommunication Union (2018) and We Are Social, DataReportal, & Hootsuite (2019).
is largely limited to authoritarian regimes (e.g., China, Russia); in democratic countries, technology companies have accumulated unprecedented resources, market advantages, and control over people’s data and access to information (Zuboff, 2019).

This hidden commercial regulation has been brought into sharp focus by several scandals implicating the social media giant Facebook in unethical dealings with people’s data (Cadwalladr & Graham-Harrison, 2018). Regulators and the general public have awakened to the extent to which digital technologies and tech companies can infringe on people’s privacy and control access to information; these scandals revealed the manipulative power of techniques such as “dark ads” (advertising messages that are visible only to those who are targeted by them) and microtargeting (customizing advertisements to particular individuals), which influence people’s decision making and voting behavior by exploiting their psychological vulnerabilities and personal identities (e.g., Matz, Kosinski, Nave, & Stillwell, 2017). There is no panacea for solving these problems. Instead, there are multiple entry points for addressing the existing and emerging challenges (Figure 1; see also Lazer et al., 2018). We argue that psychology must play a key role in this process.

**Figure 1.** Entry points for policy interventions in the digital world.
The first entry point for interventions comes from the realm of law and ethics; this includes legislative regulations and ethical guidelines (e.g., Ethics guidelines for trustworthy AI by the High Level Expert Group on Artificial Intelligence, 2019). Regulatory interventions can, for instance, introduce transparent rules for data protection (e.g., the European Union’s General Data Protection Regulation, GDPR) or for political campaigning on social media, and impose significant costs for violating them; it can also implement serious incentives (and disincentives) for tech firms and the media to ensure that shared information is reliable and online conversation is civil. Regulatory initiatives should strive to create a coherent user-protection framework instead of the fragmentary legislative landscape currently in place (e.g., see Jaursch, 2019 for Germany and the EU).² The second entry point for interventions is technological: Structural solutions are introduced into online architectures in order to mitigate adverse social consequences. For example, social media platforms can take technological measures to remove fake and automated accounts, ensure transparency in political advertisement, and detect and limit the spread of fake news using automated or outsourced fact checking (e.g., Harbath & Chakrabarti, 2019; Rosen, Harbath, & Gleicher, 2019). However, such measures are mainly self-regulatory, depend heavily on the company’s good will, and are often only introduced following considerable public, political, and regulatory pressure.

The third entry point for interventions is educational. These interventions are directed at the public as recipients and producers of information—for example, school curricula for digital information literacy that teaches students how to search, filter, evaluate, and manage

² The present regulatory framework is fragmented and disparate and focuses on the types of actors online instead of providing a more coherent form of protection that covers the entirety of the online experience (Leiser, 2019). Moreover, the EU and the US are likely to pursue different regulatory approaches to specific problems since the First Amendment of the U.S. Constitution will take precedent over many other goals (e.g., privacy).
data, information, and digital content (e.g., Breakstone, McGrew, Smith, Ortega, & Wineburg, 2018; McGrew, Smith, Breakstone, Ortega, & Wineburg, 2019). Finally, the fourth entry point for interventions comes from psychological science and encompasses behavioral and cognitive interventions: Here, nonregulatory, nonmonetary policy measures are implemented to empower people and steer their decision making towards greater individual and public good. In online behavioral and cognitive policy making, there are three notable approaches to designing interventions. The first is nudging, which aims to guide people’s behavior through the design of choice architectures (Thaler & Sunstein, 2008). The second is boosting, which seeks to improve people’s cognitive and motivational competences (Hertwig & Grüne-Yanoff, 2017). The third is technocognition, which aims to design technological solutions resting on and informed by psychological principles identified in the study of human cognition (Lewandowsky, Ecker, & Cook, 2017).

The four entry points for interventions—coming from law, technology, education, and psychology—are interrelated and can inform each other. For example, regulations on the ethical design of digital technologies should inform technological, educational, and behavioral interventions. At the same time, behavioral and cognitive insights from psychological science can be useful for designing both educational and technological tools. In this article we are concerned specifically with behavioral and cognitive interventions that can be successfully applied to counter the challenges people encounter in digital environments. Indeed, our main aim is to present a conceptual map of a single type of cognitive intervention in the digital world: boosts. We focus on boosts for several reasons. First, although the call to increase people’s ability to deal with the challenges of online environments is growing louder (e.g., Independent High Level Group of Fake News and Online Disinformation, 2018; Lazer et al., 2018), there has been no systematic account of interventions based on insights from psychological science that could form the foundation of future efforts. Second, the Internet is
a barely constrained playground for commercial policy makers and choice architects acting in accordance with financial interests; in terms of power and resources, benevolent public choice architects are at a significant disadvantage. It is therefore crucial to ensure that psychological and behavioral sciences are employed not to manipulate users for financial gain, but instead to empower the public to detect and resist manipulation. Finally, boosts are the least paternalistic measures in the toolbox of public policy makers and potentially the most resilient in the face of rapid technological change as they aim to foster lasting and generalizable competences in users.

We begin by comparing online environments with offline environments in order to fill any gaps in the current understanding of new digital environments (Section 2). In Section 3, we consider the challenges people encounter in the digital world and show how they impact users’ cognitive and motivational abilities. We distinguish four types of challenges: persuasive and manipulative choice architectures (Section 3.1), AI-assisted information architectures (Section 3.2), distractive environments (Section 3.3), and false and misleading information (Section 3.4). We then turn to the question of how to counteract these challenges. After a brief review of the types of behavioral and cognitive interventions that can be applied to the digital world (nudges, technocognition, and boosts; Section 4) we focus on boosts (Section 5). Here we identify four types of boosts: self-nudging, which aims at enhancing people’s agency in their digital environments (Section 5.1); deliberate ignorance, which can be used as a tool for information management (Section 5.2); simple decision aids, which can help people accurately assess content they encounter online (Section 5.3); and inoculation, which is a preemptive intervention that aims to boost people’s resilience to online misinformation and manipulation (Section 5.4). These tools are designed to foster the civility of online discourse and protect reason and human autonomy against manipulative choice architectures, attention-grabbing techniques, and the spread of false information.
Table 1

**Glossary of Technical Terms**

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithms</td>
<td>In the context of digital environments, computer programs that order, classify, generalize, predict, and filter information online. Algorithms can be rule-based (i.e., instructions are programmed by humans) or self-learning programs (“machine learning”).</td>
</tr>
<tr>
<td>Artificial intelligence (AI)</td>
<td>In the context of digital environments, mainly refers to self-learning computer programs (“machine learning”) that analyze people’s personal data and digital footprints in order to customize their online experience. Also includes fields such as robotics, knowledge representation and reasoning, planning, and computer vision.</td>
</tr>
<tr>
<td>AI-assisted information architectures</td>
<td>AI-powered algorithmic tools that filter and mediate information online (e.g., targeted advertising, personalized recommender systems, algorithmic filtering in search engines, personalized curation of news feeds on social media; see Figure 3 for an overview).</td>
</tr>
<tr>
<td>Recommender (also: Recommendation) systems</td>
<td>Information-filtering and associations-finding algorithms that suggest products based on users’ past activities and preferences as well as the activities and preferences of other users with similar tastes.</td>
</tr>
<tr>
<td>Personalized or targeted advertising</td>
<td>A type of online advertising that shows ads to people based on their online activity as well as both stated and inferred characteristics (e.g., gender, age, interests, political views, personality traits).</td>
</tr>
<tr>
<td>Data privacy</td>
<td>Online, a set of rules for how Internet companies collect, share, and use personal information. An important aspect of data privacy concerns the choice of users to reveal or protect their personal information.</td>
</tr>
<tr>
<td>Dark patterns</td>
<td>Designs of user interfaces employed to steer people’s choices towards unintended decisions in the service of commercial interests.</td>
</tr>
<tr>
<td>Choice architecture</td>
<td>Design of the external environments within which people make decisions (e.g., location of subway exits, presentation of foods in a cafeteria, display of search results for local restaurants on Google Maps, privacy settings on Facebook). Strategically organizing the external context in which people make decisions is one way to affect their choices.</td>
</tr>
<tr>
<td>Internet</td>
<td>A global system of interconnected computer networks that includes several applications—e.g., the web, e-mail, messenger systems, and mobile applications—for communication between devices and for access to the information contained within these networks.</td>
</tr>
<tr>
<td>Web (a.k.a. World Wide Web or www)</td>
<td>A standardized system for accessing and navigating information on the Internet; it requires web browsers for access.</td>
</tr>
<tr>
<td>Social web</td>
<td>Web-based social networks and other collaborative platforms (e.g., forums). Advanced social web: Global social media websites (e.g., Facebook, Twitter).</td>
</tr>
</tbody>
</table>

### 2. Systematic Differences Between Online and Offline Environments

The Internet and the devices people use to access it represent not just new technological achievements, but also entirely new artificial environments. Much like people’s physical surroundings, these are environments in which people spend time, communicate with each
other, search for information, and make decisions. Yet the digital world is a recent phenomenon: The Internet is 50, the web is 30, and the advanced social web is merely 15 years old (see Table 1 for definitions). New adjustments and features are added to these environments on a continuous basis, making it nearly impossible for most users, let alone regulators, to keep abreast of the inner workings of their digital surroundings.

Online reality tends to be seen as different from the physical world, and computer-mediated social activities are often described as inferior substitutes for real-life or face-to-face interactions (for an overview see Green & Clark, 2015). However, this presumed dualism between online and offline worlds is becoming more problematic—and possibly obsolete—as the line separating the two environments continue to blur. The ubiquitous nature of computing and the integration of digital devices and services into material objects (e.g., cars) and actions in the physical world (e.g., navigation) makes it difficult to delineate when one is truly online or offline—a phenomenon that Floridi (2014) called the “onlife experience” (p. 43). This effect is highly visible in computerized work environments, where more and more of people’s working time is spent online. According to a report by the European Commission (2017), the use of digital technologies has increased significantly in the last 5 years in more than 90% of workplaces in the EU and most jobs now require at least basic computer skills.

That said, the digital world differs from its offline counterpart in ways that have important consequences for people’s online experiences and behavior. We will proceed by outlining several ways in which online ecologies do not resemble offline environments. A systematic understanding is required not only to fill the gaps in knowledge of the psychologically relevant aspects of the digital world, but also to ensure that psychological

---

3 Ubiquitous computing describes technology that, by virtue of its pervasiveness in everyday life, has become invisible to people (Weiser, 1991).
Interventions take into account the specifics of these new environments and the particular challenges that people are likely to face there. First steps have already been made. Marsh and Rajaram (2019) identified 10 properties of the Internet—including accessibility, unlimited scope, rapidly changing content, and inaccurate information—which they organized into three categories: (1) content (what information is available), (2) Internet usage (how information is accessed), and (3) the people and communities that create and spread the content (who drives information). They argued that these properties can affect cognitive functions such as short-term and long-term memory, reading, and social influence. Other relevant classifications summarizing the differences between online and offline environments in the context of social media include those provided by McFarland and Ployhart (2015) and Meschi, Tamir, and Heekeren (2015).

We expand on these classifications by focusing on two broad types of differences between online and offline ecologies: differences in structure and functionality and differences in perception and behavior (i.e., how people perceive the online and offline worlds and how their behavior might differ accordingly). A list of characteristics of online environments can be found in Table 2, which is followed by a detailed discussion of each characteristic.

---

4 McFarland and Ployhart (2015) outline eight discrete ambient stimuli that distinguish social media contexts from nondigital contexts: physicality, accessibility, latency, interdependence, synchronicity, permanence, verifiability, and anonymity.

5 Meschi et al. (2015) propose six features that impact the ways users interact: user identity, information format, text length, network connections, spatial distance, and temporal scale (p. 7).
Table 2

*Characteristics of Online Environments*

<table>
<thead>
<tr>
<th>STRUCTURE AND FUNCTIONALITY</th>
<th>PERCEPTION AND BEHAVIOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group sizes</td>
<td>Social cues and communication</td>
</tr>
<tr>
<td>Amount of information, limitless space and storage</td>
<td>Reliability of information and cues for epistemic quality</td>
</tr>
<tr>
<td>Rapid change and adaptivity</td>
<td>Social calibration</td>
</tr>
<tr>
<td>Intelligence, personalization, and datafication</td>
<td>Self-disclosure and privacy behavior</td>
</tr>
<tr>
<td>Choice architectures and the power of design</td>
<td>Norms of civility</td>
</tr>
<tr>
<td></td>
<td>Perception of reality</td>
</tr>
</tbody>
</table>

2.1. Differences in Structure and Functionality

*Group sizes:* There are currently more than 4.3 billion people and around 26 billion devices connected to the Internet.⁶ Digital technologies have changed the public sphere, connecting people separated in both time and space and creating the “digital public” (Bunz, 2014). Indeed, one of the predominant uses of the Internet is for communication. The social web boasts impressive numbers of users: Facebook alone has 2.4 billion active monthly users (Facebook, 2019), and the Chinese WeChat more than 1 billion (Tencent, 2019). According to Our World in Data, “social media platforms are used by one-in-three people in the world, and more than two-thirds of all internet users” (Ortiz-Ospina, 2019). Online, one can

⁶ See We Are Social, DataReportal, and Hootsuite (2019); IHS (2016). Although digital technologies are increasingly accessible, almost half of the world’s population, mainly in low-income regions, does not yet have access to the Internet. This skewedness also extends to content production, which is mostly generated in developed countries across Europe and North America: “the global North is characterized by the greatest levels of participation and is creating the bulk of digital content, while the global South contributes very little. Africa, in particular, is almost entirely omitted from these processes of digital generativity” (Graham, De Sabbata, & Zook, 2015, p. 97).
broadcast a message to a nearly unlimited audience, whereas in face-to-face communication there are physical limits to how many people can join a conversation (Barasch & Berger, 2014). Yet even though social media enables people to establish larger social networks and profit from greater global connectivity, the structures of communities and the number of close friends people have online do not significantly differ from their offline counterparts (Dunbar, Arnaboldi, Conti, & Passarella, 2015). In online social networks, the average numbers of friends (between 100 and 200) as well as number of friends who are considered to belong to the two closest circles (typically around five and 15, respectively) do not differ from the values of offline inner circles (Dunbar et al., 2015; Dunbar, 2016). This suggests that the cognitive and temporal constraints that “limit face-to-face networks are not fully circumvented by online environments” (Dunbar, 2016, p. 7).

Amount of information, limitless space and storage: Digital environments are not subject to the same constraints on information proliferation and storage found in physical surroundings. Online space is virtually limitless, contains several layers (e.g., surface web and dark web), and can grow at a high pace. Consider that when Sergey Brin and Larry Page launched Google in 1998, they archived 25 million pages. In 2013 that number had grown to 30 trillion and by 2016 to more than 130 trillion individual pages (Schwartz, 2016). At the time of writing in 2019, there were 1.7 billion websites on the Internet and approximately 4.8

7 There is a cognitive limit on the size of natural face-to-face social networks, which is thought to be determined by a combination of constraints on available social time and cognitive bounds. According to this approach (the social brain hypothesis; Dunbar, 1998), cognitive constraints are related to the size of the neocortex and associated information-processing capacity which, in primates, correlates closely with the typical size of social groups. In humans, this suggests a social circle of about 150 people, with hierarchical levels reflecting both emotional closeness and interaction frequency. These layers have values that approximate five (closest friends, a.k.a. “support clique”), 15 (“sympathy group”), 50, and 150, and extend beyond this in at least two further layers to 500 and 1,500 (Dunbar, 2016).
billion Google searches a day. Moreover, the potential for speed and scope of information propagation is much higher online, where the same message can be effortlessly and immediately copied to reach vast audiences. For example, the most shared tweet to date reached 4.5 million retweets, most of which happened in the 24 hours after the initial posting. New technologies have made processing and storing information superior to any previously available storage system (Clowes, 2013). This feature of digital technology also implies that information does not have an expiration date and can be stored more or less indefinitely—a situation that prompted the European Union to establish what is commonly referred to as the “right to be forgotten,” which provides European citizens with a legal mechanism for ordering the removal of their personal data from online databases (General Data Protection Regulation, 2018, Article 17).

**Rapid change and adaptivity:** Digital environments develop at a high rate, especially compared to most offline environments. The document-based Web 1.0 was replaced by the more interactive Web 2.0 in the beginning of the 2000s, and increasingly more sophisticated and AI-powered web of data is being introduced (Aghaei, Nematbakhsh, & Farsani, 2012; Fuchs et al., 2010). Online content can be added, removed, or changed in seconds and digital architectures can rapidly adapt to new demands and challenges. Even small changes in structures of online architectures can have major societal consequences: For example, introducing some friction into the process of sharing information (i.e., increasing the investment in time, effort, or money required to access or spread information) can significantly decrease the likelihood of citizens engaging with the affected sources, as the Chinese government’s attempt to manage and censor information shows (see Roberts, 2018).

---

8 Based on estimates from https://www.internetlivestats.com.

9 It is a tweet by Japanese billionaire Yusaku Maezawa, who promised on January 5, 2019 to give away 100 million yen ($924,000) to be shared among 100 random people (Twitter & Wikipedia, 2019).
Clicks and likes—as insignificant as they may seem individually—can collectively amount to sizable changes (e.g., for election results).

*Intelligence, personalization, and datafication:* The latest developments in the evolution of the Internet increasingly depend on datafication (the transformation of many aspects of the world and people’s lives into data\(^\text{10}\)) and mediation of content by algorithms and other intelligent technologies. Increasing datafication leads to increasing surveillance and control over people’s information diets (Zuboff, 2019), while rapidly developing machine intelligence technology spurs a gradual relinquishing of public control as well as unclarity surrounding the technology itself. For example, search engines and recommender systems (e.g., video suggestions on YouTube) routinely rely on machine-learning systems that outperform humans in many respects (e.g., RankBrain in Google). Such algorithms are both complex and nontransparent—sometimes for designers and users alike (Burrell, 2016). The opacity of machine-learning algorithms stems from their autonomous and self-learning character: They are given input and produce output, but the exact processes that generate these outputs are hard to interpret. This has led some to describe these algorithms as “black boxes” (Rahwan et al., 2019; Voosen, 2017). Modern-day online environments, unlike their offline counterparts, possess autonomous intelligence—be it pure domain-specific machine intelligence, crowdsourced human intelligence, or a powerful combination of both.

*Choice architectures and the power of design:* Another feature that distinguishes online environments from physical surroundings is the ubiquity and the power of the design that mediates people’s online experience. The design of an interface where people encounter the complexity of interconnected information online—the “human interface” (Berners-Lee,

\(^\text{10}\)“This means not just demographic or profiling data, but also behavioural metadata, such as those automatically derived from smartphones, like time stamps and GPS-inferred locations” (Kennedy, Poell, & van Dijck, 2015, p. 1).
Cailliau, Groff, & Pollermann, 1992)—presupposes that it has a decisive role in how people perceive the information presented. In other words, there is no Internet without ubiquitous choice architectures that constrain, enable, and steer user behavior (see Table 1). This very nature of online platforms affords quick design of choice architecture: It might take several years to make a city bike-friendly (e.g., by building new bike lanes), but adjusting default settings on online pages or introducing friction into the process of information sharing can take less than a day. However, the same flexibility and adaptability of online choice architectures that benevolent choice architects can use to promote positive behavior can also be manipulated by commercial and ill-meaning actors.

2.2. Differences in Perception and Behavior

Social cues and communication: Online communication differs from face-to-face communication in several ways, including the potential for anonymity and asynchronicity, the ability to broadcast to multiple audiences, and the availability of audience feedback (Misoch, 2015). Another characteristic of online communication that was emphasized in early research into Internet communication concerns the lack of nonverbal or physical cues—such as body language or vocal expressivity—that are important for conveying and understanding emotion in face-to-face communication. This raised concerns that increased use of computer-mediated communication would lead to impoverished social interaction (the reduced social cues model; e.g., Kiesler, Siegel, & McGuire, 1984). However, it has now been recognized that users adapt to the medium and substitute the lack of nonverbal cues in digital communication with other verbal cues, thereby achieving equal levels of affective content (Walther, Loh, & Granka, 2005; Walther, Van Der Heide, Ramirez, Burgoon, & Peña, 2015). Online environments also contribute to the development of social cues, offering additional nonverbal cues such as emoticons, “likes,” and shares to enrich online communication. However, social cues can mean different things to users and platforms: To a
user, a “like” button signifies appreciation or attention; to a tech firm it is a useful data point. In addition, digital social cues can leak more, and more sensitive, information than people intend to share (e.g., sexual orientation, personality traits, political views), including information that can be exploited to psychologically target and manipulate users (Kosinski, Stillwell, & Graepel, 2013; Matz et al., 2017).

Reliability of information and cues for epistemic quality: Information available online often lacks not only the typical social cues found in face-to-face interaction, but also the cues to its epistemic quality that are generally available offline, such as an indication of sources or authorship. One reason for this is that the Internet—“an environment of information abundance”—is no longer subject to traditional filtering through professional gatekeepers (Metzger & Flanigan, 2015, p. 447). Modern-day digital media replaces expert gatekeeping with either crowdsourced gatekeeping (e.g., Wikipedia) or automated gatekeeping (e.g., algorithms on social media; Tufekci, 2015). Although some online platforms deliberately construct information ecosystems that favor indicators of quality (e.g., references to sources, fact-checking) and have rules for content creation (e.g., Wikipedia, 2019), much of the content shared on social networks and online blogs does not give users sufficient cues to judge its reliability. For example, the length of messages on Twitter “encourages short declarative statements absent of supporting arguments [so that] users do not become suspicious of unreferenced assertions” (NATO StratCom, 2017, p. 16). Moreover, manipulative use of certain cues can lead to dubious or outright false claims and ideas being disseminated—for instance, by creating fake news websites, impersonating well-known sources and social media accounts, inflating emotional content (Crockett, 2017), or creating an illusion of consensus (Yousif, Aboody, & Keil, 2019).

Social calibration: The Internet can also affect social calibration—that is, perceptions about the prevalence of opinions in the general population. Offline, one gathers information
about how others think based on the limited number of people one interacts with, most of whom live nearby. In the online world, physical boundaries cease to matter; one can connect with people around the world. One consequence of this global connectivity is that small minorities of people can form a seemingly large, if dispersed, community online. This in turn can create the illusion that even extreme opinions are widespread—thereby contributing to “majority illusion” (Lerman, Yan, & Wu, 2016) and “false consensus” effects (the perception of one’s views as relatively common and of opposite views as uncommon; Leviston, Walker, & Morwinski, 2013; Ross, Greene, & House, 1977). It is difficult to meet people in real life who believe the Earth is flat, whereas online among Facebook’s billions of users there are some who do share this belief—or other equally exotic ones—and they can now find and connect with each other.

*Self-disclosure and privacy behavior:* The emergence and development of new online environments has consequences not only for how people communicate with others or how they evaluate information but also for the way they disclose information about themselves. Early studies on self-disclosure (revealing personal information to others) reported higher levels of sharing in visually anonymous computer-mediated communication than in face-to-face communication (Joinson, 2001; Tidwell &Walther, 2002). People also tend to be more willing to disclose sensitive information in online surveys with reduced social presence of the surveyor (Joinson, 2007). A systematic literature review by Nguyen, Bin, and Campbell (2012) reported mixed evidence: While most experimental studies (four of six) that measured self-disclosure showed more disclosure in online than in face-to-face interactions, in survey studies participants reported more disclosure and willingness to share information with their offline friends (six of nine surveys). One may speculate that while it is the level of closeness, trust, and depth of interactions that prompts people to disclose personal information in offline relationships, the anonymity afforded by online communication is what can enhance people’s
willingness to share. The benefits of online anonymity include the elimination of hierarchical markers (e.g., gender and ethnicity) that may trigger hostility (Young, 2002) and a sense of control people have over the information they share that stems from a belief that it will not be linked to their real personas. However, this sense of control can backfire. For example, one study showed that increasing individuals’ perceived control over the release and access of private information can increase their willingness to disclose sensitive information (“the control paradox”; Brandimarte, Acquisti, & Loewenstein, 2013). Another paradox in people’s privacy behavior online is the “privacy paradox”: On one hand, people claim to care a great deal about their online privacy, but on the other, they show little concern for it in their actual behavior (Acquisti, Brandimarte, & Loewenstein, 2015; see Kokolakis, 2017 for a review). One potential reason for this discrepancy between what people say about online privacy and what they actually do is the lack of transparency and understanding surrounding how online platforms collect and use people’s data and what can be inferred from that data (Kosinski et al., 2013).

*Norms of civility:* The “online disinhibition effect” describes “a lowering of behavioral inhibitions in the online environment” (Lapidot-Lefler & Barak, 2012, p. 434) that is not seen offline. Online disinhibition can be both benign and toxic (Suler, 2004): It can inspire acts of generosity and help shy people socialize, but it can also lead to increased incivility in online conversations—a behavior “that can range from aggressive commenting in threads, incensed discussion and rude critiques, to outrageous claims, hate speech, and more severe forms of harassment such as purposeful embarrassment and physical threats” (Antoci, Bonelli, Paglieri, Reggiani, & Sabatini, 2018, p. 3). One of the most common examples of incivility is trolling, a type of online harassment that involves “posting inflammatory malicious messages in online comment sections to deliberately provoke, disrupt, and upset others” (Craker & March, 2016, p. 79). Trolling can be used strategically to disrupt the
possibility of constructive conversation (see also Section 3.4). Incivility is pervasive online: A survey by the Pew Research Center revealed that 44% of Americans have personally experienced online harassment, while 66% have witnessed it being directed at others (Duggan, 2017). Although incivility in online comments can polarize how people perceive issues in the media (Anderson et al., 2014) and can disproportionately affect female politicians and public figures (Rheault, Rayment, & Musulan, 2019), it seems to be perceived as the norm, rather than the exception, for online interaction (Antoci, Bonelli, Paglieri, Reggiani, & Sabatini, 2019). Relatively little is known about the causes of these phenomena. One may speculate that actions in the online sphere might be perceived as less impactful: For instance, insulting and even threatening anonymous users in online forums may be perceived as less harmful and consequential for both the victim and the perpetrator than threatening someone to their face.

Perception of reality: In contrast to the offline world, the Internet and social media are immaterial, virtual environments that do not exist outside of the human-created technology that supports them (McFarland & Ployhart, 2015). One consequence of this separation is the potential to construct multiple realities for different audiences and media online (Waltzman, 2017), so that any reference to the objective truth and shared reality is replaced by alternative narratives (e.g., “systemic lies” created to promote a hidden agenda; McCright & Dunlap, 2017). The impact of the Internet on the media landscape—along with several other factors, such as rising economic inequality and growing polarization—is likely to have contributed to the emergence of the “post-truth” environment, an alternative epistemic space “that has abandoned conventional criteria of evidence, internal consistency, and fact-seeking” (Lewandowsky, Ecker & Cook, 2017, p. 360). In this alternative post-truth reality, deliberate falsehoods can be described as “alternative facts” and politicians and media figures (on both sides of the Atlantic) can claim that “objectivity is a myth which is proposed and imposed on
us” (Yaffa, 2014), that “there is no such thing as fact anymore” (Holmes, 2016) or that “truth isn’t truth” (Pilkington, 2018, see also Lewandowsky & Lynam, 2018; Lewandowsky, in press-b). These environments are conducive to the dissemination of false news and rumors, which in turn undermines public trust in any information and erodes the basis of shared reality (Watts & Rothschild, 2017), thereby creating an atmosphere of doubt that serves as a fertile ground for conspiracy theories (more on this in Section 3.4).

To summarize, online and offline worlds differ in psychologically and functionally relevant ways. The online world appears to trigger perceptions that can render it different from the offline world. When people and online architectures are brought into contact (without much public oversight and democratic governance), pressure points will emerge. We next review four such challenges: persuasive and manipulative choice architectures, AI-assisted information architectures, distractive environments, and the proliferation of false and misleading information.
3. Challenges in Online Environments

Table 3

Map of Challenges

<table>
<thead>
<tr>
<th>Persuasive and manipulative choice architectures</th>
<th>AI-assisted information architectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Persuasion versus manipulation</td>
<td>• Personalization and customization</td>
</tr>
<tr>
<td>• Dark patterns</td>
<td>• Algorithmic gatekeeping</td>
</tr>
<tr>
<td>• Privacy-intruding defaults</td>
<td>• Algorithmic biases</td>
</tr>
<tr>
<td></td>
<td>• User profiling and targeted advertising</td>
</tr>
<tr>
<td>Section 3.1; Figure 2</td>
<td>Section 3.2; Figure 4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Distractive environments</th>
<th>False and misleading information</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Information overload and attention economy</td>
<td>• Types and sources of false and misleading information</td>
</tr>
<tr>
<td>• Virtual Skinner boxes and schedules of reinforcement</td>
<td>• Strategies for content creation and dissemination</td>
</tr>
<tr>
<td>• Multitasking and distracted minds</td>
<td>• Specifics of false information dissemination in online networks</td>
</tr>
<tr>
<td>Section 3.3; Figure 4</td>
<td>Section 3.4; Figures 5,6</td>
</tr>
</tbody>
</table>

3.1. Persuasive and Manipulative Choice Architectures

Modern online environments are replete with smart, persuasive choice architectures that are designed primarily to maximize financial return for the platforms, capture and sustain users’ attention, monetize user data, and predict and influence future behavior (Zuboff, 2019). For example, Facebook’s business model relies on exploiting user data to the benefit of advertisers; the goal is to maximize the likelihood that an ad captures its target’s attention. In order to stretch the time people spend on the platform (thus producing behavioral data and watching ads), Facebook employs a variety of design techniques that aim to change users’ attitudes and behavior by means of persuasive choice and information architectures (e.g., Eyal, 2014; Fogg, 2003). It is no coincidence that notifications are red; the color incites a sense of urgency. The “like” button triggers a quick sense of social affirmation. The
bottomless news feed, with no structural stop to scrolling (i.e., infinite scroll), prompts people to consume more without noticing. These examples illustrate that persuasive choice architectures rely on an understanding of human psychology and extensive use of commercial nudging. Benefiting from an abundance of data on human behavior, these architectures are continuously being adapted to offer ever more appealing user interfaces in order to compete for human attention (e.g., Harris, 2016).

The main ethical ambiguity of persuasive choice architectures and commercial nudging resides in their close ties to other types of influence, such as coercion and, in particular, manipulation. Coercion is a type of influence that does not convince its targets, but rather compels them by eliminating all options except for one (e.g., take-it-or-leave-it choices). Manipulation is a hidden influence that attempts to interfere with people’s decision-making processes in order to steer them toward the manipulator’s ends. It neither persuades people nor deprives them of their options; instead, it exploits their vulnerabilities and cognitive shortcomings (Susser, Roessler, & Nissenbaum, 2018). Manipulation thus undermines both people’s control and their autonomy over their decisions—that is, their sense of authorship and their ability to identify with the motives of their choices (e.g., Dworkin, 1988). It also prevents people from choosing their own goals and pursuing their own interests. Not all persuasive choice architectures are manipulative—only those that exploit people’s vulnerabilities in a nontransparent, covert manner. Below we consider two cases where persuasive design in online environments borders on manipulation: dark patterns and hidden privacy defaults.

“Dark patterns”—a term coined by designer and user-experience researcher Harry Brignull (see Brignull, 2019; Gray, Kou, Battles, Hoggatt, & Toombs, 2018; Mathur et al., 2019) are a manipulative and ethically questionable use of persuasive online architectures. “Dark patterns are user interface design choices that benefit an online service by coercing,
steering, or deceiving users into making unintended and potentially harmful decisions” (Mathur et al., 2019, p. 1). One notorious example of dark patterns is the “roach motel,” unglamorously named after devices used to trap cockroaches. The roach motel makes it easy for users to get into a certain situation, but difficult to get out (in Figure 2 it falls under the type “hard to cancel”). Many online subscription services function that way. For instance, creating an Amazon account requires just a few clicks, but deleting it is difficult and time-consuming. The user must first hunt for the hidden option of deleting an account, then request this procedure by writing to customer service. This asymmetry in the ease of getting in and out is persuasive and retains customers. Another example is “forced continuity”: subscriptions that, after an initial free trial period, continue on a paid basis without notifying users in advance and without giving them an easy way to cancel the service.¹¹

Dark patterns are anything but rare. In a recent large-scale study, Mathur et al. (2019) tested automated techniques that identified dark patterns on a sizeable set of websites. They discovered 1,818 instances of dark patterns from 1,254 websites in the data set of 11,000 shopping websites. Mathur et al.’s findings revealed 15 types of dark patterns belonging to seven broader categories (see Figure 2), such as misdirection, applying social pressure, sneaking items into the user’s shopping basket, and inciting a sense of urgency or scarcity (a strategy often used by hotel booking sites or airline companies).

¹¹ Further examples can be found at https://www.darkpatterns.org/types-of-dark-pattern.
<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sneaking</td>
<td>Sneak into basket, adding additional products to users’ shopping carts without their consent.</td>
</tr>
<tr>
<td></td>
<td>Hidden costs, revealing previously undisclosed charges to users right before they make a purchase.</td>
</tr>
<tr>
<td></td>
<td>Hidden subscription, charging users a recurring fee under the pretense of a one-time fee or a free trial.</td>
</tr>
<tr>
<td>Urgency</td>
<td>Countdown timer, indicating to users that a deal or discount will expire using a countdown timer.</td>
</tr>
<tr>
<td></td>
<td>Limited-time message, indicating to users that a deal or sale will expire soon without specifying a deadline, thus creating uncertainty.</td>
</tr>
<tr>
<td>Misdirection</td>
<td>Confuishment, using language and emotion (shame) to steer users away from making a certain choice.</td>
</tr>
<tr>
<td></td>
<td>Visual interference, using style and visual presentation to steer users towards or away from certain choices.</td>
</tr>
<tr>
<td></td>
<td>Trick questions, using confusing language to steer users into making certain choices.</td>
</tr>
<tr>
<td></td>
<td>Pressured selling, preselecting more expensive variations of a product, or pressuring the user to accept the more expensive variations of a product and related products.</td>
</tr>
<tr>
<td>Social proof</td>
<td>Activity messages, informing the user about the activity on the website (e.g., purchases, views, visits).</td>
</tr>
<tr>
<td></td>
<td>Testimonials of uncertain origin, testimonials on a product page whose origin is unclear.</td>
</tr>
<tr>
<td>Scarcity</td>
<td>Low-stock message, indicating to users that limited quantities of a product are available, increasing its desirability.</td>
</tr>
<tr>
<td></td>
<td>High-demand message, indicating to users that a product is in high demand and likely to sell out soon, thereby increasing its desirability.</td>
</tr>
<tr>
<td>Obstruction</td>
<td>Hard to cancel, making it easy for the user to sign up for a recurring subscription but cancelation requires e-mailing or calling customer care.</td>
</tr>
<tr>
<td>Forced action</td>
<td>Forced enrollment, coercing users to create accounts or share their information to complete their tasks.</td>
</tr>
</tbody>
</table>

*Figure 2. Categories and types of dark patterns.*

Another case of persuasive design that borders on manipulation is hidden default settings. Hidden defaults present a particularly strong challenge because they trick people into accepting settings without being fully (if at all) aware of the consequences. For example, online platforms are often designed to make it difficult to discontinue personalized advertising or choose privacy-friendly settings. Default data-privacy settings do not even have to follow dark patterns strategies: Most users, lacking the time or motivation to go several clicks deep into the settings labyrinth, will not change their defaults unless they have a specific reason to do so. Hidden defaults raise clear ethical concerns, but these practices continue despite the introduction of the GDPR in Europe in 2018, which stresses the importance of privacy-respecting defaults and insists on a high level of data protection that does not require users to actively opt out of the collection and processing of their personal data (General Data Protection Regulation, 2018, Article 25).

Not everyone complies. According to a report by the Norwegian Consumer Council (2018), tech companies such as Google, Facebook, and, to a lesser extent, Microsoft, use design choices in “arguably an unethical attempt to push consumers toward choices that benefit the service provider” (p. 4). On the topic of privacy, the report’s key findings include the use of privacy-intrusive default settings (e.g., Google requires that the user actively go to the privacy dashboard in order to disable personalized advertising), framing and wording that nudges users towards a choice by presenting the alternative as ethically questionable or highly risky (e.g., on Facebook: “If you keep face recognition turned off, we won’t be able to use this technology if a stranger uses your photo to impersonate you”), giving users the illusion of control (e.g., Facebook allows users to control whether Facebook uses data from partners to show them ads, but not whether the data are collected and shared in the first place), take-it-or-leave-it choices (e.g., a choice between accepting the privacy terms or
deleting an account), and design of choice architectures where choosing the privacy-friendly option requires more effort from the users (Norwegian Consumer Council, 2018).

In sum, persuasive designs and commercial nudges can go far beyond transparent persuasion and enter the territory of hidden manipulation when they rely on dark patterns (Mathur et al., 2019), default settings that intrude on user privacy (Norwegian Consumer Council, 2018), and the exploitation of people’s biases and vulnerabilities (Susser et al., 2018). These practices impact not only how users access information but also what information they agree to share. Moreover, online manipulation undermines people’s control and autonomy over their decisions by nudging them toward behaviors that benefit commercial actors, or by hiding relevant information (e.g., settings for discontinuing personalized advertisement).

3.2. AI-Assisted Information Architectures

Another challenge of online information and choice architectures comes with the use of machine learning and smart algorithms. We use the term *AI-assisted information architectures* to describe a variety of AI-powered algorithmic tools that filter and mediate information online. These tools include personalized targeted advertising, personalized recommender systems, algorithmic filtering in search engines, and customized news feeds on social media (see Figure 3 for an overview). Algorithmic filtering and personalization are not inherently malicious technologies—on the contrary, they are helpful tools that allow people to navigate the overwhelming amount of information on the Internet. Instead of showing countless random results for search queries, search engines aim to offer the most relevant results. Googling “Newcastle” in Sydney, Australia, should prioritize information about the city that is 200 km to the north, not its distant British namesake. In a similar vein, news feeds on social media strive to show news that is interesting to users. Recommender systems offer content suggestions based on users’ past preferences and the preferences of users with similar
tastes (e.g., video suggestions on Netflix and YouTube). Besides selecting information based on its personalized relevance, algorithms can also filter out information that is considered to be harmful or unwanted, for instance by automatically filtering spam or flagging hate speech and disturbing videos. There are countless examples of why filtering information on the Internet is indispensable and helpful and why automation makes this daunting process more efficient. Automated algorithmic systems act as buffers between the abundance of information and the scarcity of human attention. However, they are not without some notable problems.

One general problem is that decision making is being delegated to a variety of algorithmic tools without clear oversight, regulation, or an understanding of the mechanisms underlying the resulting decisions. For example, ranking algorithms and recommender

Figure 3. Examples of AI-assisted information architectures online.
systems are considered proprietary information and therefore neither individual users nor society in general has a clear understanding of why information in search engines or social media feeds is ordered in a particular way (Pasquale, 2015). Other factors contribute further to the lack of transparency, such as the inherent opacity of machine-learning algorithms (the black box problem) and the complexity of algorithmic decision-making processes (de Laat, 2018; Diakopoulos, 2015; Turilli & Floridi, 2009). Delegating decision making this way not only results in impenetrable algorithmic decision-making processes, it also precipitates people’s gradual loss of control over their personal information and a related decline in human agency and autonomy (Anderson & Rainie, 2018; Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016; Zarsky, 2016).

Consistent delegation of choice and shifting autonomy from users to algorithms leaves open the question of responsibility and accountability. Since artificial agents are capable of making their own decisions and since no one has decisive control over their actions, it is difficult to assign responsibility for the outcomes (e.g., the responsibility gap; see Matthias, 2004). Consider the decisions of a recommender system employed on YouTube (boasting about 1.6 billion users, it is the second most visited website in the US and worldwide12). The recommender algorithm—based on deep neural network architecture—offers video recommendations to YouTube users with the predominant purpose of increasing watching time (Covington, Adams, & Sargin, 2016). However, one unintended consequence happened to be that the system promoted videos that tended to radicalize their viewers with every step. For example, Tufekci (2018) reported how after showing videos of Donald Trump during the 2016 presidential campaign, YouTube started to recommend and autoplay videos featuring white supremacists and Holocaust denialists. After playing videos of Bernie Sanders,

---

12 See eMarketer & ExchangeWire, 2018; Alexa, 2019.
YouTube suggested videos on left-wing conspiracies (e.g., that the U.S. government was behind the September 11 attacks). There is now evidence suggesting that these algorithms may have actively contributed to the rise and unification of right-wing extremists in the US (Kaiser & Rauchfleich, 2018), Germany (Rauchfleisch & Kaiser, 2017), and Brazil (Fisher & Taub, 2019); it is unlikely that these are the only affected countries. Who, then, should be held accountable for decisions made by autonomous recommender systems that suggest ever more radical content on YouTube: the developers of the algorithms, the owners of the platforms, or the content creators? YouTube recently vowed to limit recommending conspiracy theories on its platform (Wong & Levin, 2019), in a move that highlights the tech industry's unilateral power to shape their users' information diets.

Another closely related concern is the impact of AI-driven algorithms on choice architectures—for instance, when algorithms function as gatekeepers, deciding what information should be presented and in what order (Tufekci, 2015). Be it personalized advertising or filtering information to present the most relevant items, the results directly impact people’s choices by both narrowing their options (Newell & Marabelli, 2015) and steering their decisions in a particular direction. The consequences loom large for societies as a whole as well as for individuals: Epstein and Robertson (2015) argued that search engine rankings that favor a particular political candidate can shift voting preferences of undecided voters by 20% or more.

Microtargeted advertisement on social media, especially in the context of political campaigning, is another case in point. This method relies on automated targeting of messages based on people’s personal characteristics (as extracted from their digital footprints) and a use of private information that stretches the notion of informed consent (e.g., psychographic profiling; see Matz et al., 2017). The resulting microtargeted political messages, which are seen only by the targeted user, can exploit people’s psychological vulnerabilities while
evading public oversight. The impact of this manipulation on the outcomes of the Brexit vote and the 2016 U.S. election is a major cause for concern and an argument for stricter regulation of online platforms (e.g., Digital, Culture, Media and Sport Committee, 2019; Jamieson, 2018; Persily, 2017). A majority of social-media users in the US (62%, n = 4,594) agree that it is not acceptable for social media platforms to use their data to deliver customized messages from political campaigns (Smith, 2018b). The impact of microtargeting is often exacerbated by the lack of transparency in political campaigning on social media: It is nearly impossible to trace how much has been spent on microtargeting and what content has been shown (e.g., Dommett & Power, 2019).

Another challenging consequence of algorithmic filtering is algorithmic bias (e.g., Bozdag, 2013; Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017; Fry, 2018). Here ethical concerns touch upon both the generation of biases in data processing and the societal consequences of implementing biased algorithmic decisions, such as discrimination (Mittelstadt et al., 2016; Rahwan et al., 2019). One particularly disturbing set of examples concerns deeply rooted gender or racial biases that can be picked up by data-processing algorithms. One study of personalized Google advertisements demonstrated that setting the gender to female in simulated user accounts resulted in fewer ads related to high-paying jobs than did setting it to male (Datta, Tschantz, & Datta, 2015). Another study found that online searches for “black-identifying” names were more likely to be associated with advertisements suggestive of arrest records (e.g., “Looking for Latanya Sweeney? Check Latanya Sweeney’s arrests”; Sweeney, 2013). Striking examples of racial biases in algorithmic decision making are not limited to online environments; they also have consequential effects offline, for instance in policing and health (e.g., Obermeyer, Powers, Vogeli, & Mullainathan, 2019).

Algorithms are designed by human beings, and they rely on existing data generated by human beings. They are therefore likely to not only generate biases due to technical
limitations but also reinforce existing biases and beliefs (Bozdag, 2013), which in turn can deepen ideological divides and exacerbate political polarization. Relatedly, it has been argued that personalized filtering on social media platforms may be instrumental in creating “filter bubbles” (Pariser, 2011) or “echo chambers” (Sunstein, 2017). Echo chambers are information environments “in which individuals are exposed only to information from like-minded individuals” (Bakshy, Messing, & Adamic, 2015, p. 1130) while filter bubbles refer to content selection “by algorithms according to a viewer’s previous behaviors” (p. 1130). Both echo chambers and filter bubbles tend to amplify the confirmation bias—a way to search for and interpret information that reinforces preexisting beliefs and increases political polarization (e.g., Bail et al., 2018) and radicalization.13

3.3. Distractive Environments

We now turn to still another challenging aspect of online environments: the way they shape not only information search and decision making but also people’s ability to concentrate and allocate their attention efficiently. As early as 1971, Herbert Simon understood that in an information-rich world, the abundance of information goes hand in hand with a scarcity of attention on the part of the individuals and organizations that consume information: “A wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it” (Simon, 1971, pp. 40–41). Information overload and scarcity of attention became even more salient with the rapid evolution and proliferation of the Internet and media technologies. The original goals behind the web were first, to create a user interface that would facilitate access to

---

13 Not everyone agrees about the existence of filter bubbles (Lewandowsky, Ecker, & Cook, 2017), arguing, for instance, that news audience fragmentation is less prevalent than it is often assumed (Flaxman, Goel, & Rao, 2016) or that face-to-face interaction is currently even more segregated (Gentzkow & Shapiro, 2011).
information and second, to simplify the process of information accumulation in the interconnected online space (Berners-Lee, et al., 1992). Organizing information and making it accessible is also Google’s official mission statement (Google, 2019).

However, as new informational environments evolved and business models of Internet companies were refined, the goals and incentives of Internet design shifted as well. Human collective attention became a profitable market resource for which different actors compete. Fierce competition for human attention has led to the growing fragmentation of collective attention, with ever greater proliferation of novelty-driven content and shorter attention intervals allocated to particular topics (Lorenz-Spren, Mønsted, Hövel, & Lehmann, 2019). By analyzing the dynamics of collective attention that is spent on cultural items like Twitter hashtags, Google queries, or Reddit comments, Lorenz-Spren, Mønsted, et al. (2019) showed that across the last decades, the rate at which the popularity of items decreased or increased has grown. For example, in 2013, a hashtag on Twitter was popular on average for 17.5 hours; in 2016 it lasted only 11.9 hours. The authors’ explanation is that when the excess of information meets limited attentional capacities, the thirst for novelty leads to accelerated ups and downs for each item and a higher frequency of alternating items. In other words, the amount of collective attention allocated to each single topic is decreasing and more topics are attended to in the same amount of time.

Changing economic incentives in the zero-sum race for finite human attention also affects the way modern Internet technologies are designed to be appealing, addictive, and distractive (see also Harris, 2016). Take, for instance, Facebook, which provides users with many types of rewards, including positive feedback in the form of “likes” and shares, social reinforcements in messages and comments, and friend requests. As Meshi et al. (2015) noted, “even minimalistic cues of social success such as these may activate our brain’s reward system, and keep us coming back to Facebook for more” (p. 774)—not unlike in Skinner’s
operant-conditioning experiments with rats and pigeons (“virtual Skinner boxes”; Davidow, 2013), but this time with humans as the subjects. There is ongoing speculation that Internet companies are using behaviorist research on operant conditioning and schedules of reinforcement (e.g., Ferster & Skinner, 1957) to reward and maintain desired online behavior (e.g., playing video games or checking updates on social media). According to the operant-conditioning approach, the strength of behavior depends not only on the reinforcement, but also on the intervals or schedules upon which rewards are delivered. While fixed schedules depend on rewards being delivered at predictable time intervals (fixed-interval schedules) or after a certain number of attempts (fixed-ratio schedules), in variable-interval schedules reinforcements are delivered at time intervals that are unpredictable from a subjective perspective (e.g., checking text messages that arrive at unpredictable times). Variable-ratio schedules involve reinforcement after an average (but not fixed) number of responses (e.g., winning a prize after a variable number of attempts; see Figure 4). Both variable schedules are known to create a steady rate of responding, with variable-ratio schedules producing the highest rates of responding and variable-interval schedules producing moderate response rates (Domjan, 2018, p. 119). It seems that if rewards are difficult to predict, people tend to increase the rate of a particular behavior, perhaps hoping to eventually attain the desired reward.
**Figure 4.** Four classes of schedules of reinforcement.

The operant conditioning chamber (also known as the Skinner box) was used to study animal behavior by teaching an animal (e.g., a rat) to perform certain actions (e.g., pressing a lever) in response to a controlling stimulus (e.g., a light signal) reinforced by a reward (e.g., food). Different schedules of reinforcement were studied to see which would create steady and high rates of response behavior. By analogy, “virtual Skinner boxes” such as social media or online gaming offer their users rewards (e.g., likes or reaching another level in a game) at varying intervals to reinforce and maintain the desired behavior.

Gambling, slot machines, and lottery games are typically evoked as examples of variable-ratio schedules that maintain behavior efficiently (e.g., Thorens, Wullschleger, Khan, Achab, & Zullino, 2012). Online gaming is another example of the principles of behavioral conditioning being used to maximize user commitment (Ducheneaut, Yee, Nickell, & Moore, 2006). Wu (2016), Harris (2016) and several others have argued that the
same intermittent reinforcement mechanisms are also used by online applications such as Facebook or Twitter. Jonathan Badeen, cofounder of online dating app Tinder, recently acknowledged that its algorithms were inspired by this behaviorist approach (Reynolds, 2019). Reinforcements in those cases are messages, likes, matches, comments, or any desirable content that is delivered at irregular intervals and that prompts users to constantly refresh their feeds and check their inboxes.

Furthermore, while the amount of information generated on the Internet has increased exponentially, people’s ability to process this information has not (see also Bozdag, 2013). Research on limited attention shows that people’s information processing is inevitably limited and selective, so that allocating attention to one task tends to decrease attention to another (e.g., Dukas, 2004; Kahneman, 1973). In the online world, the ability to concentrate becomes even more compromised when one’s surroundings are full of distractive stimuli that, by buzzing, ringing, or flashing, constantly call for attention. Moreover, digital environments are no longer constrained to desktop screens but are becoming increasingly integrated in people’s daily routines through a variety of smart devices. Unsurprisingly, these environments, which breed distraction and interruption, lead to “distracted minds” (Gazzaley & Rosen, 2016). Even the mere presence of a smartphone can occupy attentional resources and reduce cognitive ability (Ward, Duke, Gneezy, & Bos, 2017), and smartphone notifications disrupt performance on attention-demanding tasks even when people are not actively attending to their phone, arguably due to mind-wandering (Stothart, Mitchum, & Yehnert, 2015).

Media multitasking—simultaneously attending to several media sources, such as TV, text messages, and websites—is becoming more and more common among not only younger but also older generations (Rosen, 2008). Studies of high school and university students showed that the typical student could not stay focused on a task for more than 3 to 5 minutes
without checking their messages or browsing the web (Rosen, Carrier, & Cheever, 2013); in addition, multitasking is particularly pronounced when people read digitally rather than in print (Carrier, Rosen, Cheever, & Lim, 2015). Studies on multitasking show that switching attention between tasks instead of concentrating on one specific task not only increases the time spent on a task but also negatively affects performance (Uncapher & Wagner, 2018).

3.4. False and Misleading Information

A final challenge presented by online environments and social networks is the increasing speed and scope of false information proliferation and its resulting threat to the rationality and civility of public discourse and, ultimately, to the very functioning of democratic societies. A recent report by Bradshaw and Howard (2019) reveals the extent of the problem:

In the last two years alone, the number of countries with disinformation campaigns has more than doubled, and Facebook remains the main platform for those campaigns. There is much concern that the spread of false news and rumors on Facebook and Twitter influenced the outcomes of the U.S. presidential election and the Brexit referendum in 2016 (see House of Commons, 2019; Persily, 2017). For example, the infamous Pizzagate conspiracy alleging that Hillary Clinton and her top aides were running a child-trafficking ring out of a Washington pizzeria was floated during the 2016 presidential campaign on Reddit, Twitter, and fake news websites. It led to repeated harassment of the restaurant’s employees and

---

14 The evidence, however, is mixed: The link between multitasking and cognitive control (distractibility) turns out to be weaker than previous studies have suggested (see Wiradhany & Nieuwenstein, 2017 for a replication study and a meta-analysis). The general discrepancy in findings is also highlighted in the literature review on minds and brains of media multitaskers by Uncapher & Wagner (2018). At the same time, the authors of the review concluded that in light of the current evidence (both convergent and divergent), heavier media multitaskers exhibit poorer performance in a number of cognitive domains (e.g., working memory; Uncapher & Wagner, 2018).
eventually prompted an armed 28-year-old man to open fire inside the pizzeria (Aisch, Huang, & Kang, 2016). Even more disturbingly, the Myanmar military orchestrated a propaganda campaign on Facebook that targeted the country’s Muslim Rohingya minority group, inciting violence that forced 700,000 people to flee (Mozur, 2018). Encrypted messenger networks such as WhatsApp are also vulnerable to manipulation: False rumors about child kidnappers shared in Indian WhatsApp groups in 2018 incited at least 16 mob lynchings, leading to the deaths of 29 innocent people (Dixit & Mac, 2018).

As these examples illustrate, dangerously misleading online content might proliferate due to deliberate attempts to manipulate public opinion or emerge as an unintended consequence of sharing unverified rumors and false news. Wardle and Derakhshan (2017) used information falseness and intent to mislead to distinguish between three types of “information disorders”\(^\text{15}\): misinformation (false or misleading content shared without malicious intent), disinformation (false, fabricated, or manipulated content shared with intent to mislead or cause harm), and mal-information (genuine information shared with intent to cause harm, e.g., hate speech and leaks of private information). While this classification establishes some useful general distinctions, the landscape of online falsehoods and propaganda is much more complicated. For example, the difference in intent between misinformation and disinformation is often hard to establish and the real consequences of both can be equally harmful. Hence, they are usually both considered to be false information—or, if presented as news, referred to by the now-common moniker “fake news.” Moreover, there are additional categories of misleading content, such as online political propaganda and “systemic lies” (McCright & Dunlap, 2017); the latter are created and curated by organized groups with vested interests (e.g., fossil fuel companies denying climate science). Creating

\(^{15}\) Reference to “disorder” in this context does not mean that the blame is shifted to the user but rather denotes that information disorders are not there by design; they emerge as malfunctions in a healthy flow of information.
and disseminating false information relies on several common practices, which can be catalogued and used to develop tools to counteract misinformation (e.g., inoculation; see Roozenbeek & van der Linden, 2019 and Section 5.4). Here we list the main categories (Figure 5) of false and misleading information in the digital sphere, as well as sources and strategies used for its creation and dissemination (Figure 6), which we have compiled based on a wide range of sources (indicated in the figures).

**Figure 5.** Types of false and misleading information in the digital world.
Propaganda, rumors, conspiracy theories, and other kinds of misleading information are not new phenomena, nor are they exclusive to online environments (see Uberti, 2016). In 1275, England’s First Statute of Westminster outlawed spreading false news, stating that “none shall report slanderous news, whereby discord may arise” (c. 34). Numerous fake news stories were published in newspapers in the 19th century, including the Great Moon hoax published in New York tabloid The Sun in 1835. However, what distinguishes online propaganda and misinformation is the new medium itself. Besides having the capacity to spread misinformation further and faster, online environments offer new tools for computational propaganda that rely on the combination of algorithms and automation (e.g., bots) with human curation to flood social media networks with misleading and polarizing content (Bradshaw & Howard, 2019; Woolley & Howard, 2017). Incentives for creating this
content are often financial: Recent findings by the Global Disinformation Index (2019) showed that online ad spending on disinformation domains amounted to $235 million a year.

The scope and speed with which false information proliferates online is deeply connected to the nature of online networks (Bounegru, Gray, Venturini, & Mauri, 2018). Recent research has found that false rumors on Twitter spread faster, deeper, and broader than does truth, arguably because their novel, highly emotional nature appeals to people (Vosoughi, Roy, & Aral, 2018). This is in line with findings showing that people are more likely to share messages featuring moral-emotional language (Brady, Wills, Jost, Tucker, & Van Bavel, 2017). As Crockett (2017) argued, in the context of competition for human attention, digital media may promote the expression of negative emotions such as moral outrage “by inflating its triggering stimuli, reducing some of its costs and amplifying many of its personal benefits” (p. 769). In addition, the digital context exacerbates the problem due to the absence of reliable cues to the epistemic quality of information and shifts in the meaning of habitual social cues (see Section 2.2).

Unfortunately, people are not particularly skilled at evaluating the trustworthiness of information they encounter online. Research in education suggests that young people struggle with many aspects of finding reliable information online—from selecting search results to judging whether a site is trustworthy. Students tend to ignore the source of information, have difficulties distinguishing between traditional news and sponsored content, and evaluate sites based on superficial features such as graphic design and authoritative logos (Barzilai & Zohar, 2012; Wineburg & McGrew, 2019; McGrew, Breakstone, Ortega, Smith, & Wineburg, 2018; Wiley et al., 2009; Wineburg, McGrew, Breakstone & Ortega, 2016).

---

16 One may also speculate that false news is so attractive because it exploits the negativity bias—that is, the tendency for humans to react more strongly to negative than to positive information (Soroka, Fournier & Nir, 2019).
Young people—as “digital natives”—may nonetheless have an advantage over their parents when it comes to dealing with rapidly evolving technological landscapes and learning new digital skills. A recent study of fake news on Facebook found that Americans over age 60 were much more likely to visit fake news sites compared to younger people (Guess, Nagler, & Tucker, 2019). The same study found that 60% of visits to fake news websites were made by just 10% of Americans. A small fraction of users consuming and disseminating false information online is also found on Twitter: Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer (2019) showed that in the 2016 U.S. election campaign, fake news on Twitter “accounted for nearly 6% of all news consumption, but it was heavily concentrated—one% of users were exposed to 80% of fake news, and 0.1% of users were responsible for sharing 80% of fake news” (p. 374). The vast majority of both shares of and exposures to fake news were attributable to relatively small fractions of the population (Grinberg et al., 2019). It should be noted that in this study, the threshold for a news item to be considered fake was high; the studies therefore left out many manipulative sources, half-truths, and other misleading information techniques.

It is important to distinguish the susceptibility or even commitment to sharing false news and conspiracy theories found in a small part of the population from ignorance or mere lack of skills. “Ignorance rarely leads to strong support for a cause, in contrast to false beliefs based on misinformation, which are often held strongly and with (perhaps infectious) conviction. For example, those who most vigorously reject the scientific evidence for climate change are also those who believe they are best informed about the subject” (Lewandowsky, Eckert, Seifert, Schwarz, & Cook, 2012, p. 108). Crucially, not all people who lack knowledge about an issue are overconfident, but the opposite is usually the case: for example, people who tend to be overconfident about their knowledge about autism and vaccines are
usually the ones who are the least informed about these topics and doubt expert knowledge the most (Motta, Callaghan, & Sylvester, 2018).

Despite the somewhat uplifting finding that outright false content still represents only a small fraction of online information, even a minority opinion can set agendas in “its ability to ‘push’ or ‘drive’ the popularity of issues in the broader online media ecosystem” (Vargo, Guo, & Amazeen, 2018, p. 2043). Vargo et al.’s study showed that although fake news media did not appear to control the whole media landscape in 2014–2016, American partisan media (e.g., Fox News) and fake news media were intertwined and influenced each other’s agendas across a wide range of topics, including the economy, education, the environment, international relations, religion, taxes, and unemployment. During the 2016 election partisan media were especially susceptible to fake news agendas. Lewandowsky, Jetter, and Ecker (2019) showed that the agenda-setting power of misleading news also extends to the American president’s tweets. In response to media coverage he considers threatening (e.g., the Mueller investigation), President Trump strategically deploys distraction (e.g., tweeting about China, jobs, or immigration) to set the agenda for the mainstream media (e.g., the New York Times and ABC News), thus prompting less coverage of the issues he wishes to remove from public discourse.

Finally, people’s perceived exposure to misinformation online is high: In the European Union, “in every country, at least half of respondents [in the sample of 26,576] say they come across fake news at least once a week” (Directorate-General for Communication, 2018). In the United States, “about nine-in-ten U.S adults (89%, n = 6,127) say they often or sometimes come across made-up news intended to mislead the public, including 38% who do so often” (Mitchell, Gottfried, Stocking, Walker, & Fedeli, 2019). This fosters dissatisfaction
with democracy and erodes trust in democratic institutions. Fifty percent of Americans (n = 6,127) view made-up news as a major issue in their country and 68% say that fake news impacts confidence in government institutions (Mitchell et al., 2019). In Europe, 85% of respondents (n = 26,576) think that fake news is a problem in their country and 83% say it is a problem for democracy in general (Directorate-General for Communication, 2018).

In this section we have distinguished four groups of challenges to human cognition and motivation in online environments. While our focus here has been on urgent challenges to people’s agency, self-control, and autonomy of choice, as well as the civility and rationality of public discourse and ultimately the functioning of democratic societies, there are many other challenges raised by online environments and digital technology that deserve psychologists’ attention, such as the nature of the association between social media use and individual well-being. The four issues we reviewed are:

- Human-made, ubiquitous, persuasive and manipulative designs, which rely on dark patterns and hidden defaults, challenge the human capacity to exercise autonomous and informed choice. These practices affect not only how people access information but also—as is the case with privacy-intruding defaults—what information they agree to share.

- AI-assisted information architectures that filter information on the Internet and shape personalized information environments reduce agency and autonomy, amplify biases, and introduce obscurity into the automated decision-making processes.

---

17 “Across 27 countries polled, a median of 51% are dissatisfied with how democracy is working in their country; just 45% are satisfied” (Wike, Silver, & Castillo, 2019).
• The distractive nature of digital media triggers people’s emotional responses and consumes their limited attention, thereby hindering people’s capacity for concentration and self-control.

• False and misleading information disseminated through social networks and digital media challenge people’s ability to distinguish truth from falsehood, to detect when information has been shared with malicious intent, and to participate in civil conversation.
4. **Behavioral Interventions Online: Nudging, Technocognition, and Boosting**

Although challenges loom large, they are not insurmountable. Insights and evidence from psychological science permit to point the way to potential remedies. In this section, we summarize three types of behavioral and cognitive interventions inspired and informed by evidence from psychology that can be applied to the digital world: nudges, technocognition, and boosts (Figure 7).

**Figure 7.** Types of behavioral and cognitive interventions for the digital world.
**Nudging**

Nudging is a popular approach to behavioral policy that harnesses the power of choice environments and the knowledge of human psychology to design choice architectures in ways that steer people’s decisions toward a greater individual or public good (Thaler & Sunstein, 2008). Nudging is based on the insight that it is possible to change people’s behavior without changing their minds. Nudging does not block, fence off, or significantly burden choices (as laws can do); rather, it proposes interventions that are easy, reversible, and cheap to implement. It thus represents a form of soft paternalism, also called libertarian paternalism.

The target of these interventions is choice architectures (see Table 1 for definition). Nudging can be achieved by varying the order in which options are presented, thus changing their physical and cognitive accessibility. For example, rearranging food options in a cafeteria so that healthier foods are more accessible is meant to increase healthy food consumption (for a systematic review see Broers, De Breucker, Van den Broucke, & Luminet, 2017; Bucher et al., 2016). The preselected default option is another paradigm of nudging and one of the choice architect’s most widely employed tools. Due to the mechanisms of endorsement (defaults are seen as signaling what the choice architect wants the decision maker to do) or endowment (defaults are perceived to reflect the status quo), people are likely to accept a preselected option (Jachimowicz, Duncan, Weber, & Johnson, 2019). Benevolent choice architects can harness this tendency for causes serving the public good, such as increasing organ donation rates (Johnson & Goldstein, 2003; but see Arshad, Anderson, & Sharif, 2019), or the good of the individual, such as saving more money for retirement through automatic enrollment (Thaler & Benartzi, 2004).

The same architectural principles, however, can also be used to build online and offline choice architectures that benefit service providers rather than consumers. Commercial nudging can drive people to inadvertently subscribe to undesirable content or consent to
privacy settings that are inconsistent with their stated best interests (Section 3.1; see also Thaler, 2018 on “sludge”). The success and ethical permissibility of nudging thus largely depend on the goals of the choice architects (commercial or public good) and their alignment with the goals and values of individuals. Difficulties arise not only in determining people’s best interests or true preferences but also in maintaining a balance between what is best for different actors (individual decision makers, commercial bodies, political institutions) and society at large. Another limitation of nudging is the assumed and actual role of human autonomy. Nudges do not eliminate available options and are easily reversible (Thaler & Sunstein, 2008, p. 236). Yet they substitute autonomous choice with preselected “rational” decisions in order to overcome people’s cognitive biases and inadequate decision-making competences. As Rebonato (2012) argued, even though nominal autonomy might be preserved, the effective autonomy (as demonstrated by choices made in the presence of defaults) is reduced (but see also Sunstein, 2015).

A category of nudging that is explicitly respectful of human autonomy is known as educative nudges (Sunstein, 2015). As the name indicates, these interventions involve some form of education, for instance in the form of additional information such as the nutritional quality of foods or the risks of smoking (Sunstein, 2016a; 2016b). In contrast to noneducative nudges, these interventions are transparent to people, engage their deliberate faculties and preserve autonomy of choice—which may be why people prefer educative nudges. According to a nationally representative survey in the United States, a majority of people (between 55% and 74% across four topics, n = 430) consistently preferred educative versions of nudges when no information about their comparative effectiveness to noneducative nudges was presented (Sunstein, 2016a).
Technocognition

Technocognition is an approach proposed by Lewandowsky, Ecker, and Cook (2017) that offers a “cognitively-inspired design of information architectures” (Lewandowsky, Cook, & Ecker, 2017, p. 419). It suggests that a combination of insights from cognitive science and appropriate interventions in digital architectures can help in designing technological safeguards against the spread of false information or targeted manipulation. In digital environments, the power of choice and information architectures on users’ behavior is even more significant than in the offline world. No online choice is ever made without predesigned context. Technocognition considers this design context through the lens of cognitive science. Cognitively inspired technological interventions can, for instance, introduce friction into the process of commenting on or sharing of information. As a response to the problem of toxic commenting, the Norwegian broadcaster NRK launched an experiment: Before readers could post a comment on an article, they had to pass a brief comprehension quiz on what they’d read (Lichterman, 2017; Lewandowsky, in press-a). The friction created by increasing the entry cost for participating in online discussions is meant to foster deliberate thinking. Crucially, no one is censored in the process; once a person passes the quiz they are free to comment as usual. Yet this measure, unlike a nudge, does fence off certain behaviors unless the quiz is answered correctly. “Trolls,” for example, are unlikely to expend the effort required to pass the quiz.

A simpler version of friction can be used to prevent uncontrolled sharing cascades of false and misleading information. Instagram introduced an AI-powered feature in June 2019 that delays posts containing offensive comments by notifying users that their comment may be considered offensive and allowing them to cancel the post (Instagram Info Center, 2019). Messaging app Telegram recently introduced a “slow mode” that enables group administrators to impose a wait period before users respond (Telegram Blog, 2019). And
WhatsApp Messenger’s reportedly successful response to mob lynchings in India (see Section 3.4) was to limit the number of times a message can be shared to five chats—a feature that now applies to all users (WhatsApp Blog, 2018). The underlying cognitive insights in these cases are twofold: First, limiting the number of chats to which a message can be forwarded and removing the share button from media posts introduced a delay, or cooling-off period. Cooling-off periods are known to affect people’s willingness to engage in an activity (e.g., see Luca, Malhotra, & Poliquin, 2017 for the effect of cooling-off periods on gun violence in the United States). Second, identifying a forwarded message as such provided a cue to users that the message originated not from a (potentially trusted) contact but from elsewhere. These interventions in the information architecture of social media, while small and easy to implement technologically, can have significant effects given the scale of these platforms—a promising point for designing appropriate technocognitive solutions in digital environments.

Let us highlight, however, that similar techniques can also be used to restrict freedom of choice and communication on the Internet, as can be seen in the case of authoritarian regimes that use friction to limit citizens’ access to information (Roberts, 2018). It is therefore important to ensure that technocognitive interventions are designed with people’s best interests in mind and with public oversight.

**Boosting**

Boosting is another class of cognitive interventions from the psychological science. It responds to the challenge of rapidly changing digital environments by aiming to foster lasting and generalizable competences in users (see also Hertwig, 2017; Hertwig & Grüne-Yanoff, 2017; Hertwig & Ryall, 2019). Boosts target individual cognitive and motivational competences rather than immediate behavior (the target of nudges) and aim to empower people to make better decisions for themselves in accordance with their own goals and
preferences. Boosting interventions can be directed at domain-specific (e.g., understanding health information) and domain-general competences (e.g., statistical literacy). They can target human cognition (e.g., decision strategies), the environment (e.g., information representation), or both (Hertwig & Grüne-Yanoff, 2017, p. 977). Moreover, in contrast to nudges, boosts specifically aim not only to preserve but also to foster and extend human agency and autonomy. Boosts are also necessarily transparent because they require an individual’s active cooperation.

One example of boosting is a risk-literacy boost that can be employed to quickly educate people about relative versus absolute risks in, for instance, the health domain (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007). While benefits of drugs are often expressed in relative terms (e.g., “Drug X reduces the chance of stroke by 48%”), this information is incomplete and does not permit the user to judge the magnitude of the effect—but it suggests that the drug is highly effective. Absolute risk information, in contrast, provides easy-to-understand information about the magnitude of the drug’s benefit, such as: “Drug X reduces the chance of stroke from 28 in 1,000 to 15 in 1,000”. In this framing, the absolute reduction of stroke due to the drug is 13 in 1,000 people, or merely 1.3%. This risk-literacy boost is a simple, memorable rule—“always ask for health statistics to be translated into absolute numbers”—that can help people make more informed decisions about their health.

Boosting cognitive competences via the environment might involve changing the way information is presented to users or providing additional cues to existing information in order to improve the epistemic quality of online content (Lorenz-Spreen, Lewandowsky, Sunstein, & Hertwig, 2019). For example, such informational boosts can draw on research in
algorithmic detection of false rumors (e.g., Vosoughi, Mohsenvand, & Roy, 2017) in order to design visual aids (e.g., information icons on social media posts) that represent hidden cues such as a post’s propagation dynamics—that is, to make cues to the veracity of information both visible and transparent. This type of intervention can improve the credibility of digital information and help people make better decisions (for more examples, see Lorenz-Spreen, Lewandowsky et al., 2019). Importantly, this additional information introduced in the environment is easily accessible but does not restrict users’ choices or activities. People can decide for themselves how much they want to engage with these information labels. In contrast to boosts that aim to foster long-term competences, information labels are short-term interventions that provide quick and context-appropriate information. However, if one encounters them repeatedly, the development of long-term competences could be spurred.

To summarize, one can distinguish between a range of interventions, all informed by psychological science and behavioral sciences, that can be harnessed to respond to the four challenges of online environments outlined earlier. Conceptualizing and studying these interventions is a task of the highest order. As long as regulators fall behind the speed of change in digital environments and are hamstrung by the political power of Big Tech, interventions informed by scientific evidence will be crucial. We next turn to a map of boosting interventions in digital environments.

18 Authors of this study identified three categories of cues predictive of veracity of online information (in this case, information shared on Twitter): people (who spreads the news), linguistic content (what words are used), and propagation dynamics (the shape of an information cascade). The most predictive cue in the study was propagation dynamics—a cue rarely detected by humans.
5. Boosting Cognitive Competences in Online Environments

Table 4

Map of Boosts

<table>
<thead>
<tr>
<th>Boosting agency and autonomy</th>
<th>Self-nudging</th>
<th>Deliberate ignorance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Self-imposed interventions in one’s proximal digital choice architecture.</td>
<td>Conscious choice to not seek information. A strategic way of filtering information in order to reduce information overload while staying informed.</td>
</tr>
<tr>
<td></td>
<td>Section 5.1, Figure 8</td>
<td>Section 5.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Boosting reasoning and resilience to manipulation</th>
<th>Simple decision aids</th>
<th>Inoculation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simple strategies and decision aids to improve people’s ability to efficiently analyze the information they encounter on the web.</td>
<td>A preemptive intervention that boosts people’s cognitive resistance to misinformation and online manipulation.</td>
</tr>
<tr>
<td></td>
<td>Section 5.3, Figures 9, 10</td>
<td>Section 5.4, Table 5, Figure 11</td>
</tr>
</tbody>
</table>

The interventions we review here are designed to satisfy two constraints simultaneously: to remedy specific problems in the digital world while also building on existing or fostering new competences. While some interventions focus on people’s internal cognitive and motivational competences, others target people’s competences to restructure the external choice architecture of the digital environment in a way that enhances their own agency and autonomy. An important point is that different tools are adapted to counter specific challenges. For instance, social media exploit humans’ reward sensitivity to cultivate hard-to-control habits that platforms capitalize on. Knowing this, the best response to manipulative and persuasive choice architecture might be to become one’s own choice architect (self-nudging; Section 5.1) or to restrict engagement with certain information sources (deliberate ignorance; Section 5.2), rather than attempting to exercise a superhuman ability to detect and resist all attempts at influence. By contrast, false information and AI-powered persuasive techniques such as targeted political advertisement can best be met by people exercising
existing competences (e.g., reasoning) or learning new ones (e.g., lateral reading; section 5.3). We thus identify two main groups of cognitive boosting tools: (1) those aimed at enhancing people’s agency and autonomy in their digital environments (e.g., self-nudging and deliberate ignorance) and (2) those aimed at boosting reasoning and resilience to manipulation in order to accurately assess content encountered online (e.g., simple decision aids, inoculation, lateral reading). The effectiveness of some of these boosts has already been demonstrated experimentally; others are supported by evidence collected from neighboring areas of research in behavioral and cognitive sciences (e.g., research on nudging, self-control, and the use of simple heuristics in decision making under uncertainty). These evidence-based and evidence-informed interventions can be presented to users, educators, and policy makers in the form of fact boxes, apps, and policy recommendations.

5.1. Self-Nudging: Boosting Control Over One’s Digital Environment

Design of choice architectures that make online environments open to manipulating user behavior can also be used by people to foster self-control and motivation. Online environments permit—although rarely encourage—a relatively high level of control over one’s choice architecture, such as setting one’s own defaults, adjusting notifications, installing ad blockers, and organizing one’s digital environment in a way that hinders interruptions and undesirable triggers. Users can take control over their digital surroundings and exercise freedom and agency by not being passive vis-à-vis their environment.

Accordingly, successful interventions in persuasive and attention-maximizing environments should aim to enhance people’s autonomy and their ability to control and shape their digital environments in ways that are consistent with their own goals.

One type of behavioral intervention that is particularly well suited for this challenge is self-nudging, which is a cognitive boost that builds competences by actively enlisting the proximate environment (Reijula & Hertwig, 2019). While nudging redesigns choice
architectures to prompt a behavioral change, self-nudging empowers people to act as their own choice architects. For example, one can choose to implement a nudge in one’s own kitchen by moving tempting but undesirable foods to harder-to-reach places. In Duckworth, Milkman, and Laibson’s (2018) classification of self-control strategies, self-nudging falls into the category of self-deployed situational strategies. The approach of self-nudging draws inspiration from three sources. First and foremost, it has roots in nudging and its emphasis on choice architecture but aims to share the psychological knowledge built into nudges with the individual. Self-nudging can therefore benefit from the accumulated evidence on nudges such as defaults (e.g., Jachimowicz et al., 2019) or changes in cognitive and spatial accessibility (Thaler & Sunstein, 2008). Another inspiration for self-nudging comes from research on commitment devices stemming from economic theory (Bryan, Karlan, & Nelson, 2010; Rogers, Milkman, & Volpp, 2014; Schelling, 1978) and used predominantly to solve self-control problems. “Commitment devices attempt to enforce people’s voluntarily imposed restrictions until they have accomplished their goals, or their voluntarily imposed penalties for failing to accomplish their goals” (Rogers et al., 2014, p. 1). In other words, a commitment device is a way to lock oneself into doing something that one may otherwise not be able to follow through with. One example is to define a health goal such as weight loss and to tell as many people as possible about when the goal must be reached and the penalty for not reaching it on time (e.g., donating to a political campaign one deeply dislikes).

Finally, self-nudging also is related to the notion of behavioral stimulus control, employed, for instance, in cognitive behavioral therapy to treat insomnia or substance abuse (e.g., Edinger, Wohlgemuth, Radtke, Marsh, & Quillian, 2001; see also Griffiths, Kuss, & Demetrovics, 2014 for online addiction). Here, strategic changes are introduced in the environment in order to manage one’s exposure to stimuli that exercise control over one’s behavior. For instance, if a person is triggered by hyperpalatable stimuli (e.g., sugary food or
distractive YouTube videos), removing them from the proximate environment or making them less accessible should strengthen the person’s ability to control their urges. The same rationale can also be applied to one’s information diet. In the words of Wendy Wood (2019), the key to self-control in the digital domain is in taking control over the contextual cues that activate people’s use of technology (e.g., smartphones) and adding friction to make undesirable actions (e.g., excessive phone use) more difficult (pp. 234–235). In what follows, we briefly review three types of self-nudges that can be enlisted by people to nudge themselves away from distracting sources or make their desired options more easily available.

*Self-nudging by adapting cognitive accessibility.* The Center for Humane Technology (2019) suggests several steps that people can take to exercise more control over the time they spend on their devices. For example, the variable reinforcement schedule of notifications (see Figure 4) can turn checking one’s phone into a powerful habit. People can control these distracting stimuli by turning off notifications for anything not coming directly from other people (e.g., news apps) or even only allowing notifications from apps used by their most important contacts (e.g., enabling notifications for messenger apps they use with friends and family but disabling e-mail notifications). They could set specific times in which messages can be received, thereby reserving periods of time for concentrated work (see also Newport, 2016). This measure can also help convert variable schedules of receiving messages to fixed-interval schedules (which is known to elicit the lowest rates of responding), thereby potentially reducing messages’ addictive character. Further advice includes deliberately separating applications that, by one’s own standards, improve the quality of time spent online (e.g., educational podcasts) and those that do not. This can be achieved by rearranging one’s smartphone home screen so that only useful apps (e.g., podcasts and meditation apps, as well as tools such as calendars and maps) are displayed on the front page while others (e.g., social
media, games) are tucked away in folders (see Figure 8, “Adaptive cognitive accessibility”). Other self-imposed interventions in one’s digital choice architecture include removing social media apps from one’s mobile devices entirely and accessing them from one’s home computer only or leaving devices out of sight to reduce the cognitive accessibility of the most distracting platforms.

**Self-nudging by adjusting defaults:** Defaults are one of the most widely employed tools in the choice architecture toolbox and they have considerable impact on decisions (Jachimowicz et al.’s. 2019 meta-analysis produced a medium-sized effect of \( d = .68 \)). Defaults are fertile ground for self-nudging. People can take control of their digital default settings, including privacy settings on social media and settings for personalized advertisements (e.g., https://myactivity.google.com). While it might initially require some time, effort, and possibly even guidance to understand how default settings work and their considerable effect on people, self-command over defaults may prove to be a powerful way to return agency and autonomy to users (see Figure 8, “Adjusting defaults”).

**Self-nudging with the help of technology:** One can also make use of external apps (e.g., Digital Wellbeing, Cold Turkey, Freedom, and Boomerang; see Figure 8, “Using technology”) that allow users to control how much time they spend on their phone, to schedule e-mails, or to block all notifications for a period of time in order to maintain focus. Being in control, the self-nudger decides which goals and tools to prioritize and which to move to the background.
**Self-nudging interventions online**

<table>
<thead>
<tr>
<th>ADAPTING COGNITIVE ACCESSIBILITY</th>
<th>USE GRAY SCALE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KEEP YOUR HOME SCREEN FOR TOOLS ONLY</strong></td>
<td><strong>Use gray scale to decrease the attractiveness of sensory stimuli, such as the color of notifications or the appealing color design of most apps.</strong></td>
</tr>
<tr>
<td>- Limit your home screen to just tools—the apps you use for quick in-and-out tasks like maps, camera, calendar, notes.</td>
<td></td>
</tr>
<tr>
<td>- Move the rest of your apps, especially those you wish to deprioritize, off the home screen and into folders.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ADJUSTING DEFAULTS</th>
<th>CONTROL PRIVACY DEFAULTS IN YOUR BROWSER</th>
<th>CONTROL PRIVACY AND PERSONALIZATION SETTINGS IN YOUR ONLINE ACCOUNTS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SET DEFAULT &quot;DO NOT DISTURB&quot; TIMES</strong></td>
<td><strong>Configure your browser (e.g., Firefox) to erase your cache and cookies automatically every time you close it. Enable tracking protection and Do Not Track to block online trackers from collecting your browser data.</strong></td>
<td><strong>Review your privacy and advertisement settings on your online accounts, such as Google, Facebook, Twitter, and Amazon (e.g., <a href="https://myaccount.google.com/activecontrol">https://myaccount.google.com/activecontrol</a>).</strong></td>
</tr>
<tr>
<td>- Adjust the times when your phone will automatically switch to silent mode (e.g., during your sleeping hours). In this mode, sounds will still ring or vibrate but calls and messages will not.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>USING TECHNOLOGY</th>
<th>CONTROL YOUR SCREEN TIME</th>
<th>USE PRIVACY PROTECTING BROWSERS AND SEARCH ENGINES</th>
<th>USE A VPN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CONTROL YOUR SCREEN TIME</strong></td>
<td><strong>Track the time you spend on your device. For Apple, use Screen Time. For Android, consider installing ActionDish, Digital Wellbeing, or Moment.</strong></td>
<td><strong>Use privacy-focused browsers, such as Firefox or Brave.</strong></td>
<td><strong>Use a VPN especially on public Wi-Fi that encrypts your web traffic so no one can see where sites you’re visiting or what you’re doing while you’re there (e.g., NordVPN).</strong></td>
</tr>
<tr>
<td>- Selectively control and block apps on your phone using Freedom or Boomerang. Temporarily block your phone using Cold Turkey.</td>
<td><strong>Use privacy-focused search engines, such as DuckDuckGo or Startpage.</strong></td>
<td><strong>Prevent ads from being loaded to your browser with an ad blocker (e.g., UBlock, AdBlock Plus, AdBlock).</strong></td>
<td></td>
</tr>
</tbody>
</table>

Figure 8. Self-nudging interventions in online environments.

A summary of potential self-nudging interventions to enhance people’s control over their digital environments and their privacy protection online. Based in part on Center for Humane Technology (2019) and Epstein (2017).

All these interventions aim to enhance users’ control over their digital environments and digital lives. Self-nudging is particularly suited to situations when exercising self-control or resisting temptation is difficult, or when a choice environment is toxic (i.e., when choice architects design highly addictive environments with nonbenevolent goals in mind; see Hertwig, 2017). Moreover, self-nudging enhances autonomy, as it aims to put people in charge of determining their own goals and choice environments, thus bypassing the paternalism that can accompany the kind of nudging that capitalizes on people’s deficiencies rather than attempting to educate them. “Self-nudging means that people intentionally nudge themselves in order to self-regulate their behavior and break self-destructive habits […].
When the nudger and the nudged are one and the same person, as in the case of self-nudging, autonomy and agency remain intact” (Hertwig, 2017, p. 155). A sense of agency is crucial for self-nudging, as is an awareness of the existing environmental challenges and a certain level of understanding of psychological mechanisms involved. As a Pew Research Center survey shows, more users who think they have more control over their news feeds than users who think they have no control have attempted to influence the content of their Facebook feeds (Smith, 2018a).\(^{19}\)

Let us also highlight the potential limitations of self-nudging. Although a perceived sense of control is crucial for exercising agency, it should be backed up by appropriate affordances in the environment. As discussed previously, persuasive design can create an illusion of control while still nudging users away from privacy-friendly choices (e.g., Norwegian Consumer Council, 2018; Section 3.1). Moreover, increasing individuals’ perceived control over the release and access of private information can, paradoxically, increase their willingness to disclose sensitive information (“the control paradox”; Brandimarte et al., 2013). For example, users can have a strong feeling of control on Facebook because they can change their default privacy settings and adjust who will see what type of information in their profiles. At the same time, they have very little control over the way in which the information they share will be used by the platform, by third-party applications, or even by their friends. We hold that self-nudging efforts should be

---

\(^{19}\) “Half (50%) of Facebook adopters who think users have a lot of control over their news feeds have themselves attempted to influence the content they see there. But that share falls to 40% among those who think users have only a little control over the content of their feeds, and to 24% among those who think they have no control” (Smith, 2018a).
Cognitive tools for the digital world are complemented by reasonable regulations and online tools\textsuperscript{20} that not only give users more control over their digital environments but also ensure that personal information is protected regardless of users’ actions.

### 5.2. Deliberate Ignorance as Information Management Device

In 1807, Thomas Jefferson condemned the “polluted vehicle” of newspapers, claiming that “the man who never looks into a newspaper is better informed than he who reads them; inasmuch as he who knows nothing is nearer to truth than he whose mind is filled with falsehoods and errors” (Jefferson, 1999, p. 275). The current challenge of information overload and environments designed to compete for human attention by offering rewards and hyperpalatable stimuli brings new significance to this statement. Moreover, in modern societies many schemes for the cultural production of ignorance have evolved, including the organized campaigns that undermine scientific consensus around climate change in order to divert public attention and policy initiatives (Proctor & Schiebinger, 2008). Technological advances and control over the Internet (and thus over immense audiences) have further bolstered massive, orchestrated attempts to produce ignorance and put the very existence of objective truths into question (see Section 2.2). The flooding technique is one method for producing ignorance. The Chinese government is estimated to create and post about 448 million social media comments per year—not to address controversial issues or even argue with critics of the party and the government, but rather to divert attention from real issues (e.g., natural disasters and the government’s lackluster response) towards trivial and

\textsuperscript{20} Rose-Stockwell (2018) argued that social media could provide users with curation tools for their own algorithmic filtering, such as prioritizing posts from family members, diversifying the spectrum of political news shown, ordering news chronologically, or filtering out posts filled with moral outrage. MIT Social Media Lab has developed a social media aggregator with filters that users can control called Gobo:

scandalous stories that are injected online for the sole purpose of distracting the public from objective coverage of government weaknesses (King, Pan, & Roberts, 2017; Roberts, 2018).

Modern-day readers of digital media face a constant trade-off between staying informed about current events and being exposed to an information environment in which numerous players (e.g., companies, advertisers, media, and policy makers) design hyperpalatable mental stimuli to hijack people’s limited attention. Much as obesogenic environments are replete with foods designed to offer maximal sensory pleasure, informationally fattening environments degrade consumers’ control and autonomy over the information they consume (Crawford, 2015). Take, for illustration, the Kardashian sisters, who found fame on a reality television show. The lifestyles they present in the media are carefully curated to appear as desirable as possible (e.g., signaling wealth and beauty) and lure in viewers. The Kardashians are to information what sugary soda is to nutrition: Neither offers much value to the consumer. When low-quality clickbait stories, conspiracy theories, and fake news masquerade as meaningful information, epistemic abstinence becomes more rational than epistemic indulgence. In other words, more information is not always better. In order to manage information overload, one must ignore a large amount of incoming material and separate useful information from noise, false news, or harmful advice. In this context, deliberate ignorance can be used as a tool for information management (Hertwig & Engel, 2016; Hertwig & Engel, in press).

The idea that deliberate ignorance can be an ecologically rational strategy does not agree with classical ideals of epistemic virtue and rationality (see Kozyreva & Hertwig, 2019), which presume that information and knowledge have intrinsic value for decision makers because they allow them to accumulate more evidence (e.g., Carnap, 1947), acquire better understanding, and ultimately make more informed and rational choices (e.g., Blackwell, 1953; Good, 1967). However, real-world decision makers are boundedly rational
agents, constrained not only by their cognitive limitations but also by their choice environments (Simon, 1990). (This does not prevent people from making good decisions under uncertainty and time constraints; see Hertwig, Pleskac, Pachur, & The Center for Adaptive Rationality, 2019.) Moreover, the expansion of new artificial environments packed with information and noise casts additional doubts on how applicable classical rationality is to modern-day humans (see also Floridi, 2015).

People deliberately ignore information for various purposes: for example, to avoid emotional costs (e.g., choosing not to test for a rare genetic disease), to benefit from strategic ignorance (e.g., in negotiations) or to insure impartiality of judgment (e.g., in blind auditions; for these and many more examples see Hertwig & Engel, 2016, in press). We suggest that using deliberate ignorance as an individual information management device can also be extended to the digital world. As Hertwig and Engel (2016) argue: “For humans, who are hardwired to monitor their environment, the ability to allocate one’s limited attentional resources reasonably is therefore becoming increasingly valuable in today’s world. Indeed, the ability to select a few valuable pieces of information and deliberately ignore others may become a core cultural competence to be taught in school like reading and writing” (p. 364).

One information environment where deliberate ignorance can be a helpful and rational tool is online health information. Facing a flood of low-quality sources, people are at risk of becoming victims of bad health advice or even conspiracy theories.\(^\text{21}\) Tempting, highly unrepresentative, and possibly even misleading environments are difficult to navigate. One ecologically rational strategy in such environments is to abstain from seeking out these narratives, to avoid searching for one’s symptoms in search engines, and to ignore health

\(^{21}\) For instance, online narratives from parents reporting adverse effects of vaccines appear to be an important source of doubt about vaccination safety (the “narrative bias”; Betsch, Haase, Renkewitz, & Schmid, 2015; Haase & Betsch, 2012).
advice from influencers, celebrities, or commenters in online forums. One could instead choose to consult only trusted medical professionals and/or one or two reliable websites that provide evidence-based health information. Perhaps unsurprisingly, high-quality health information is relatively rare online, making the ability to intentionally ignore low-value persuasive sources an important skill. For example, Oxman and Paulsen (2019) identified only three (yet not consistently adapted for the public use and access) websites that met their inclusion criteria for evidence-based aggregation of health information: Cochrane Evidence, Informed Health, and PubMed Health. The next step would be to make the information on these websites publicly accessible and easily understandable (e.g., in fact boxes; see McDowell, Rebitchek, Gigerenzer, & Wegwarth, 2016).

Let us emphasize that we are not advocating for the proliferation of ignorance, echo chambers, and a return to the Dark Ages. An informed public remains the cornerstone of democracy and widespread education is one of its highest achievements. Moreover, the accessibility of information offered by the Internet should be regarded as a public good. Our emphasis in discussing deliberate ignorance as a tool for information management is on its strategic use by consumers of information to shield themselves from the excesses, sticky traps, and information disorders of current modern digital environments. This strategy appears to be particularly suitable for online environments and sources that convey low-quality information. To this end, adaptive and sound deliberate ignorance strategies require, somewhat ironically, knowledge—such as an understanding of what constitutes a reliable indicator of quality and trustworthiness. Auspiciously, Pennycook and Rand (2019) demonstrated that laypeople—on average and across the political spectrum—are quite good at distinguishing between lower and higher quality sources and place more trust in media outlets with stronger editorial norms than in hyperpartisan and fake news sources.
In sum, introducing strategic changes in one’s digital choice architectures (e.g., defaults, arrangement of options) can be an efficient way to counteract challenges associated with persuasive and attention-maximizing environments. It can also be used to improve people’s information diets and diminish their exposure to false information and information that provokes negative emotions such as anger or outrage. Another strategy that aims to preserve autonomy in digital information environments that lure users into surrendering their time and attention spans is deliberate ignorance. When used in an informed way, deliberate ignorance can be a powerful information-management tool that protects people from the excesses of an attention-grabbing information landscape but also shields them from encountering false or manipulative news. Next, we turn to other boosting tools that can help counteract misinformation and online manipulation of public opinion.

5.3. Simple Decision Aids: Boosting Digital Information Literacy

One aspect of digital information literacy is the ability to analyze and evaluate the information people encounter online. Simple strategies and decision aids can help people do this. The idea is to foster good habits that are as simple and automatic as washing one’s hands or scanning the crosswalk before making a turn (Caulfield, 2018). One way to design such simple tools makes use of a skillset of professional fact-checkers, who are experts in evaluating the truthfulness of information. In order to develop a set of rules based on this skillset, researchers from the Stanford History Education group asked participants (professional fact-checkers, n = 10; history professors, n = 10; undergraduate students, n = 25) to evaluate the trustworthiness of information online (Wineburg & McGrew, 2019). Wineburg and McGrew (2019) argued that the key to experts’ success in fact-checking is their strategy of lateral reading, a heuristic rule that allows them to “read less and learn more” by looking to verify the claim outside of the original post. Contrary to the professors and students, who focused on the information source itself, fact-checkers (who were the most
successful group of participants across several fact-checking tasks) spent most of their time verifying the source and the evidence behind the claim by checking information about it on the web.\footnote{“The biggest lesson we learned from watching these experts: They evaluated unfamiliar websites by leaving them. For fact checkers, the direct route to credibility was indirect” (Wineburg & McGrew, 2019, p. 45). “When fact checkers encountered an unfamiliar website, they immediately left it and read laterally, opening up new browser tabs along the screen’s horizontal axis in order to see what other sources said about the original site’s author or sponsoring organization. Only after putting their queries to the open web did checkers return to the original site, evaluating it in light of the new information they gleaned. In contrast, students approached the web by reading vertically, dwelling on the site where they first landed and closely examining its features—URL, appearance, content, and “About” page—without investigating who might be behind this content” (Breakstone et al., 2018, p. 220).} In a similar vein, Graves (2017) attested that the key to professional fact-checkers’ analysis lies “in discovering a claim’s origin and reconstructing its spread” (p. 525).

Drawing inspiration from expert strategies, researchers in the Stanford History Education Group then identified simple rules (see Figure 9) geared at boosting the competence of civic online reasoning. This competence is defined as “the ability to effectively search for, evaluate, and verify social and political information online” (McGrew et al., 2019, p. 2) and it includes three subcompetences: evaluation of the source, evaluation of the evidence, and lateral reading. One way of representing these competences is through simple questions such as those featured in Figure 9 (Breakstone et al., 2018, p. 221). McGrew et al. (2019) found that after two 75-minute lessons on evaluating the credibility of online sources (an extended version of the three rules outlined above), students in the treatment condition (n = 29) were more than twice as likely to score higher at posttest of their online reasoning skills than at pretest, while students in the control condition (n = 38) were equally likely to score higher at posttest than at pretest, indicating that the intervention was successful.
SIMPLE RULES FOR ONLINE REASONING

WHO IS BEHIND THIS INFORMATION?

• Investigate authors, inquire into their motives (commercial, ideological, or otherwise), and decide whether they should be trusted.
• Check the name of the source (e.g., official or impersonating a popular news agency?), spelling in the social media name, cues for advertisement and promoted material.
• If the site or the author of the post is unfamiliar, always use search engines and Wikipedia to check its origin.

WHAT IS THE EVIDENCE?

• Consider whether the claim is a matter of fact, a matter of opinion, or both.
• If it is a matter of fact, then consider whether claims are backed by factual evidence, whether the evidence is trustworthy, and whether the evidence directly supports the claims.
• If this issue is a matter of opinion, then consider the arguments, whether they are relevant to the claim and support it, whether the arguments are sound and reasonable, and whether they are manipulative (e.g., try to provoke emotional response or tend to exaggerate).

WHAT DO OTHER SOURCES SAY?

• Conduct independent Internet searches on the veracity of claims (e.g., place the title of the news item or the name of an organization in quotation marks in the search field to avoid irrelevant results).
• Use fact-checking web-sites to verify claims (e.g., Snopes, Politifact).
• Practice click restraint, i.e., spend some time scrolling through search results and click only on those stemming from credible news organizations.

Figure 9. Simple rules for online reasoning.

Based on research by the Stanford History Education Group: Breakstone, McGrew, Smith, Ortega, & Wineburg, 2018; McGrew, Smith, Breakstone, Ortega, & Wineburg, 2019; Wineburg & McGrew, 2019.

Another example of simple decision aids that can be designed to foster better information literacy online are fast-and-frugal decision trees (FFTs; Martignon, Katsikopoulos, & Woike, 2008; Luan, Schooler, & Gigerenzer, 2011). Already in use in a variety of domains, including medicine, finance, law, and management, FFTs provide comprehensive prescriptive guides for real-world decision making (Hafenbrädl, Waeger, Marewski, & Gigerenzer, 2016). They rank decision criteria in the order of importance and offer a potential exit at each point. To make a decision, a person goes through the cues sequentially. For example, a simple decision tree can be used for triage in hospital emergency rooms in order to quickly categorize patients into those who need immediate medical
attention and those whose treatment can be delayed (one such system, called simple triage and rapid treatment, was used in New York City hospitals during the World Trade Center attack in 2001; see Cook, 2001). Cues in this case are framed as questions: Is the patient walking? If yes, delay treatment; if no, proceed to the following cue. Implementing and understanding an FFT is easy; it requires only that one knows the order of the cues and their exit conditions.

There are not many examples of simple decision aids for the online domain. But one short intervention has already been applied to improve people’s ability to use linguistic cues to distinguish between authentic and fictitious online reviews (Banerjee, Chua, & Kim, 2017). Similarly, FFTs could be designed and tested as decision aids to choices such as whether to trust information encountered online. In Figure 10, we offer an example of a potential decision tree based on the rules for fact-checking identified by Breakstone et al. (2018) and Wineburg and McGrew (2019). The FFT advances through the cues sequentially and ends when the answer is “no,” which indicates that the information is not trustworthy and should not be shared. FFTs work best with strong cues or signals, but in some cases a combination of weak signals can be used (with the help of the tallying strategy), such as the top-level domain (e.g., .com or .gov), how the social media name is spelled, the “about” page, and cues for verified accounts or promoted material. However, all these signals must be taken with caution. For example, a fishy top-level domain (e.g., com.co) is a signal that the source may be untrustworthy, but the opposite is not necessarily true (e.g., a .gov domain does not guarantee trustworthiness). Many cues for trustworthiness can be gamed and fake news websites can appear as genuine and well designed as the websites of real news organizations.

---

23 There is an accessible computerized toolbox for creating, visualizing, and evaluating FFTs called FFTrees (Phillips, Neth, Woike, & Gaissmaier, 2017).
That is why strong negative signals such as an unfamiliar website should be taken seriously, and unfamiliar sources should always be verified using the lateral reading strategy.

**Figure 10.** “Can you trust this information?”

This fast-and-frugal decision tree provides users with three crucial steps for evaluating the trustworthiness of information online, based on the research by the Stanford History Education Group (Breakstone et al., 2018; Wineburg & McGrew, 2019).

Like any cognitive tool in the toolbox of digital decision makers, simple decision aids must be used under appropriate conditions. For example, lateral reading is an effective tool for verifying the information encountered on a fishy website or social media feed, but it may not be the best strategy for reading trusted material that benefits from concentration and focus on one source. Similarly, decision trees are appropriate tools for dichotomous decisions (e.g., whether to trust or share a news item or not) but they might not be helpful for more complex choices that require more sophisticated deliberation.
5.4. Inoculation: Boosting Cognitive Resilience to Misinformation and Manipulation

Another cognitive intervention against false information and online manipulation is inoculation, also known as “prebunking.” It targets people’s ability to recognize misleading or manipulative strategies before they encounter them face-to-face or online. Metaphorically speaking, if disinformation is a disorder, then inoculation can immunize people against certain strains of false and misleading information. Inoculation is preemptive: It aims to expose people to misleading or manipulative strategies and to neutralize their disruptive potential before people actually encounter them in the world (for more on the inoculation theory see Compton, 2013; McGuire & Papageorgis, 1961). Inoculation differs from debunking strategies, which refute false claims only after they have been seen or heard; it is thus especially valuable, since disinformation is often resistant to debunking after the fact (Cook, Lewandowsky, & Ecker, 2017; Lewandowsky et al., 2012). Furthermore, unlike topic-specific debunking, inoculation aims to instill recipients’ domain-general competence to see through attempts of manipulation (Roozenbeek & van der Linden, 2019), making it a particularly suitable cognitive strategy when fact-checking or evidence-based refutation is costly or unavailable.

According to Cook et al. (2017) there are two components to inoculation: first, an explicit warning about a potential threat of disinformation or manipulation—for example, a warning about attempts to cast doubts on the scientific consensus on climate change that create a chimerical set of “experts” who disagree with the consensus. The second step refutes an anticipated argument, thus exposing the disinformation strategy and rendering its deceptive nature transparent. In our climate change example, this could take the shape of an illustration and an explanation of a particular deceptive technique used to question a scientific consensus or otherwise manipulate the public (Cook et al., 2017, p. 4). In the study by Cook et al. (2017), the inoculation consisted of showing participants the “fake experts” strategy.
used by the tobacco industry in the 1960s (a tobacco ad with the text “20,679 Physicians say ‘Luckies are less irritating’”). The same strategy was used by climate science denialists: The Oregon Petition denied human-caused effects on the Earth’s atmosphere and was signed by 31,000 alleged experts, of whom 99% had no expertise in climate science. By exposing participants to a weakened version\(^{24}\) of disinformation, this intervention provided them with a counterargument. The efficacy of inoculation in preventing acceptance of disinformation has been established in several experiments (Cook et al. 2017; van der Linden, Leiserowitz, Rosenthal, & Maibach, 2017) and inspired the creation of Bad News, an educational game on fake news (Roozenbeek & van der Linden, 2018, 2019).

Table 5

<table>
<thead>
<tr>
<th>INOCULATION Type 1 (passive)</th>
<th>Warning about potential misinformation or manipulation (e.g., about attempts to cast doubts on scientific consensus on climate change)</th>
<th>Refutation of an anticipated argument in a weakened form (e.g., an example and an explanation of the “fake experts” strategy)</th>
<th>Post-intervention exposure (e.g., expose participants to the same strategy used by climate science denialists)</th>
</tr>
</thead>
<tbody>
<tr>
<td>INOCULATION Type 2 (active)</td>
<td>Pre-intervention test (e.g., ratings of fake news credibility)</td>
<td>Active learning (e.g., the Bad News game, which aims to present main disinformation strategies in a weakened, fun way)</td>
<td>Post-intervention test (e.g., credibility ratings of fake news)</td>
</tr>
</tbody>
</table>

The Bad News game study aimed to extend the effects of inoculation beyond a particular topic (such as climate change) and develop a “broad-spectrum vaccine” against disinformation (Roozenbeek & van der Linden, 2019, p. 2). It focused on the tactics commonly used to produce disinformation, rather than on the content of a specific

---

\(^{24}\) Like vaccines, where a weakened version of a virus is administered, a weakened version of disinformation does not entail the same level of risk as encountering such disinformation in an uncontrolled setting might.
disinformation campaign. The study provided an active type of inoculation (see Table 5) by having participants play a game (https://getbadnews.com/) in which they learned six strategies often used to spread disinformation (based on NATO StratCom, 2017): impersonating people or famous sources online, producing provocative emotional content, amplifying group polarization, floating conspiracy theories, discrediting opponents, and trolling (see also Figure 6). The underlying idea of the game is that people train to become expert manipulators by applying different disinformation techniques; in doing so, they develop the competence of detecting manipulation, which will help them realize when manipulative strategies are being applied to them in the future. The game environment represents a weakened form of real-world social media (where people are apt to encounter false information). The inoculation effects of the Bad News game were observed by comparing pre- and post-intervention credibility ratings of various fake news items (n = 14,266; d = .52 average across all items). The effects were most pronounced for individuals who had been more susceptible to fake news headlines in the first place (d = .89).

Figure 11. Bad News game
Inoculation aims to boost cognitive resilience to disinformation and manipulation (van der Linden, Maibach, Cook, Leiserowith, & Lewandowsky, 2017). As is the case with all the interventions we have discussed, it is an efficient strategy when it fits particular challenges in the environment and the cognitive competences involved. Inoculation interventions must be based on an understanding of the manipulative strategies being used online and how they work. Furthermore, people must be willing to be inoculated—that is, to take the time to learn about these techniques. Another limitation of inoculation is that it is ineffective in the face of unexpected or novel deceptive techniques. Thus, as with vaccines in the physical world, it makes sense to be prepared for the most insidious and common methods of online manipulation and to regularly update inoculation techniques. The logic of inoculation could be extended beyond misinformation to other challenges—for instance, helping people detect manipulation through personalized political advertisement that exploits people’s psychological identities and vulnerabilities. This research is currently underway.

6. Conclusion: From Psychological Science to the Internet for Citizens

Technological innovations have frequently been associated with dystopian fears. As far back as 370 BCE, thinkers such as Socrates were deeply concerned about the detrimental consequences of writing:

In fact, it [writing] will introduce forgetfulness into the soul of those who learn it: they will not practice using their memory because they will put their trust in writing, which is external and depends on signs that belong to others, instead of trying to remember from the inside, completely on their own. […] Your invention will enable them to hear many things without being properly taught, and they will imagine that they have come to know much while for the most part they will know nothing (Plato, ca. 370 B.C.E/1997, pp. 551–552).
Today’s concerns about, for instance, the potential effects of Google on memory (e.g., Sparrow, Liu, & Wegner, 2011) and comprehension, or about digital amnesia or digital dementia (e.g., Spitzer, 2012), echo Socrates’s fear of forgetfulness and shallow comprehension. Socrates was not wrong—it might well be the case that the capacity of human memory has fundamentally changed from the time knowledge was transmitted orally. Yet he did not foresee the wide range of benefits—including the invention of the Internet—that were rendered possible by this new form of communication.

Honoring this lesson, we are cognizant of the risk of conjuring up dystopian fears. The current and future benefits of the digital revolution are immense. Yet there is also a growing body of evidence that reveals worrying implications of the digital transformation, with at least four aspects of the transformation causing particular concern. First, unlike previous communication innovations, which permeated societies on time scales of centuries (e.g., writing) or decades (e.g., telephony), today’s digital transformations occur at a breathtaking pace. Apps can appear outdated within a few months, and the lifecycle of information technologies is notoriously short. The comparatively slow pace of academic research, with its cycle of prolonged peer review and revision, cannot fully capture, let alone influence, those transformations. Second, the problem of speed is compounded by the degree of mutation that technology can undergo. Whereas the psychological affordances of writing changed little during the transition from parchment and quill to paper and pencil, new digital technologies can create new psychological affordances in an instant. For example, the seemingly trivial addition of a “retweet” button has made it possible for a small number of people—or indeed, nonhuman “bots”—to trigger global informational cascades (e.g., Bastos & Mercea, 2019). Third, the implications of those mutations cannot be anticipated. WhatsApp did not anticipate that the ease with which material can be shared would contribute to mob killings, and Facebook likely did not anticipate that a platform designed for
staying in touch with friends and family would end up influencing the outcome of elections through dark ads and misinformation (Jamieson, 2018). Finally, and perhaps most troubling, is that this digital transformation is occurring in what is largely a regulatory vacuum. There is nothing to stop platforms from radically altering their interfaces overnight, with unknown consequences for society and democracy—a situation recently brought into focus by Facebook’s decision to allow distribution of false statements in political advertisement under the dubious argument of free speech protection (Facebook Newsroom, 2019).

A recent report by the RAND corporation (Mazarr, Bauer, Casey, Heintz, & Matthews, 2019) condensed those concerns into a number of future scenarios, described under the umbrella term of the “emerging risk of virtual societal warfare.” Mazarr et al. (2019) pointed to associated social trends, such declining faith in institutions that help to sustain generally agreed-upon social truths (e.g., the media), weakened measures of social capital (e.g., social trust and civic engagement), an increase in partisan polarization across many countries, a rise in populist movements and, last but not least, what various scholars (e.g., Specter, 2009, p. 33) have described as a sense of alienation and a loss of agency and ontological security (Giddens, 1991). People’s trust in social institutions, their interpersonal exchanges, the stability and reliability of facts, and even their sense of shared reality are being undermined. One of the future digital scenarios considered by Mazarr et al. is entitled “The Death of Reality.” Envisaged for 2023, it is the point at which the “ability to manufacture seemingly tangible reality from scratch has … become commonplace” (p. 99). Present-day antecedents for this scenario can be found in the radical constructivist ontology of truth employed by practitioners of “post-truth” discourse (Lewandowsky, in press-b). Arguably, this scenario can only materialize within a digital information architecture that permits people to personalize all of reality along with their preferences for deodorants.
The focus of this article has been on challenges that threaten people’s agency, their choice autonomy, and the epistemic quality of their information environment. Many other challenges exist and new ones are quickly emerging, such as the massive amounts of highly plausible but fabricated video and audio material known as “deepfakes” that are further deflating confidence in a shared reality. In an increasingly “onlife” world (Floridi, 2014), psychological science faces important tasks. One is to measure and understand the psychological effects of these revolutionary transformations. Another is to develop and design policy interventions that help people cope with the consequences of those transformations. Focusing on the four challenges of online environments, we outlined various classes of interventions that are informed by the behavioral sciences, then focused on interventions aimed at empowering people; returning a sense of agency to people (e.g., the citizen as a choice architect); and fostering autonomy, self-control, and resistance to being manipulated in the digital world—in other words, interventions meant to cultivate a sense of self-efficacy and ontological security (see Figure 12).

**ENVIRONMENT–COGNITION–TOOLS**

**CHALLENGES**
- Manipulative choice architectures
- AI-assisted information architectures
- Distractive environments
- False and misleading information

**COMPETENCES**
- Autonomous agency
- Attention and self-control
- Evaluation of sources and evidence
- Recognition of manipulation

**BOOSTING TOOLS**
- Self-nudging
- Deliberate ignorance
- Simple decision aids
- Inoculation

*Figure 12. Map of challenges and boosts in the digital world.*
These four types of tools can also be summarized as four simple rules for mindful Internet behavior that could become as routine as washing one’s hands or checking for cars before crossing the street:

1. Control and organize your digital environment. Adapt it to your goals.
2. Learn to ignore and filter out nonessential information.
3. Make a habit of using simple rules for data privacy and information literacy.
4. Immunize yourself against the most common and dangerous types of manipulation and disinformation.

For policy makers these rules can mean:

1. Ensure that users have adequate control over their digital environments and personal data. Make it easy for them to take the reins.
2. Make it easy for people to separate useful information from noise and disinformation—for instance, by mandating clear, intuitive indicators of epistemic quality.
3. With the help of researchers, design simple rules for data privacy and information literacy and provide them to users.
4. Monitor common types of online disinformation and manipulation and provide appropriate and timely inoculations.

We have no illusions. There is no single solution for these and many other challenges. It is very likely that these interventions will be shown to have some benefits, but only for some users. Nevertheless, it is important to start, and soon: Several surveys show that people are concerned about data privacy, the spread of false information, political manipulation, and online harassment (e.g., Directorate-General for Communication, 2018; Mitchell et al., 2019). Any solution will require the orchestrated efforts of regulators, policy makers,
educators, and users—for instance, boosting people’s ability to control the default parameters of their choice architecture should be accompanied by a regulatory framework that takes the heterogeneity of users into account. Specifically, the law could be used preventatively to stop companies from taking advantage of the fact that some citizens fail to take control of their default privacy settings (e.g., the European Union’s GDPR, embracing insights from the behavioral sciences, mandates that data controllers can no longer use opt-out as a default for obtaining consent to data processing).

The rules and design of Internet landscapes are predominately dictated by major corporations and signal a lack of a coherent regulatory framework for transparent and robust user protection. Contrary to promises of the early digital era (e.g., access to information for all, empowered minorities, unsuppressed democratic deliberation), citizens find themselves in a state of constant information overload, surveillance, manipulation, and digital divide. We believe that psychological science must contribute to the long-term goal of designing and fostering the “Internet for citizens,” an online world respectful of fundamental human rights and values that will require users to learn new competences and make active decisions. One may think this is an unrealistic ideal. Yet one need look no further than the digital world itself for evidence of the spectacular human ability to learn: Fifty years ago, Neil Armstrong became the first person to step onto the moon. And yet he would have been unable to navigate the digital world and its technologies with the ease of a nine-year-old today.
References


https://doi.org/10.1007/s13347-017-0293-z

https://doi.org/10.1080/21670811.2014.976411

https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/179102.htm

https://data.europa.eu/euodp/data/dataset/S2183_464_ENG

https://www.buzzfeednews.com/article/pranavdixit/whatsapp-destroyed-village-lynchings-rainpada-india


https://doi.org/10.1111/1467-923X.12687

https://doi.org/10.1177/1555412006292613
https://doi.org/10.1177/1529100618821893


Dunbar, R. I. M. (2016). Do online social media cut through the constraints that limit the size of offline social networks? *Royal Society Open Science, 3*(1), 150292.
https://doi.org/10.1098/rsos.150292

https://doi.org/10.1016/j.socnet.2015.04.005

https://doi.org/10.1016/j.socnet.2015.04.005


Epstein, R. (2017). *Seven simple steps toward online privacy.* Medium.
https://medium.com/@re_53711/seven-simple-steps-toward-online-privacy-20dcb9fa82


https://doi.org/10.1177/2053951715621569


https://doi.org/10.1016/j.cose.2015.07.002


Lichterman, J. (2017, March 1). This site is “taking the edge off rant mode” by making readers pass a quiz before commenting. *NiemanLab*, https://www.niemanlab.org/2017/03/this-site-is-taking-the-edge-off-rant-mode-by-making-readers-pass-a-quiz-before-commenting


Motta, M., Callaghan, T., & Sylvester, S. (2018). Knowing less but presuming more: Dunning-Kruger effects and the endorsement of anti-vaccine policy attitudes. Social Science & Medicine, 211, 274-281. https://doi.org/10.1016/j.socscimed.2018.06.032


https://doi.org/10.1177/1461444814535724

https://doi.org/10.1080/13669877.2018.1443491


https://doi.org/10.1016/j.chb.2012.12.001


Waltzman, R. (2017, April 27). *The weaponization of information: The need for cognitive security.* Testimony presented before the Senate Armed Services Committee, Subcommittee on


**Graphic material:** Images used in figures are licensed at Adobe Stock, except for a “nudging” icon in Figure 7, which is a courtesy of Luis Prado at thenounproject.com (CC license); icons in Figure 2, which are used with permission of the Dark Patterns Project at Princeton University, and images in Figure 11, which are screenshots of the Bad News Game at getbadnews.com.