PERSPECTIVE
https://doi.org/10.1038/s41562-020-0889-7

nature human behaviour

How behavioural sciences can promote truth, autonomy and democratic discourse online

Philipp Lorenz-Spreen[1]✉, Stephan Lewandowsky[2,3], Cass R. Sunstein[4] and Ralph Hertwig[1]

Public opinion is shaped in significant part by online content, spread via social media and curated algorithmically. The current online ecosystem has been designed predominantly to capture user attention rather than to promote deliberate cognition and autonomous choice; information overload, finely tuned personalization and distorted social cues, in turn, pave the way for manipulation and the spread of false information. How can transparency and autonomy be promoted instead, thus fostering the positive potential of the web? Effective web governance informed by behavioural research is critically needed to empower individuals online. We identify technologically available yet largely untapped cues that can be harnessed to indicate the epistemic quality of online content, the factors underlying algorithmic decisions and the degree of consensus in online debates. We then map out two classes of behavioural interventions—nudging and boosting— that enlist these cues to redesign online environments for informed and autonomous choice.

To the extent that a "wealth of information creates a poverty of attention" (p. 41)[1], people have never been as cognitively impoverished as they are today. Major web platforms such as Google and Facebook serve as hubs, distributors and curators[2]; their algorithms are indispensable for navigating the vast digital landscape and for enabling bottom-up participation in the production and distribution of information. Technology companies exploit this all-important role in pursuit of the most precious resource in the online marketplace: human attention. Employing algorithms that learn people's behavioural patterns[3–6], such companies target their users with advertisements and design users' information and choice environments[7]. The relationship between platforms and people is profoundly asymmetric: platforms have deep knowledge of users' behaviour, whereas users know little about how their data is collected, how it is exploited for commercial or political purposes, or how it and the data of others are used to shape their online experience.

These asymmetries in Big Tech's business model have created an opaque information ecology that undermines not only user autonomy but also the transparent exchange on which democratic societies are built[8,9]. Several problematic social phenomena pervade the internet, such as the spread of false information[10–14]—which includes disinformation (intentionally fabricated falsehoods) and misinformation (falsehoods created without intent, for example, poorly researched content or biased reporting)—or attitudinal and emotional polarization[15,16] (for example, polarization of elites[17], partisan sorting[18] and polarization with respect to controversial topics[19,20]). Some disinformation and misinformation involve public health and safety; some of it undermines processes of self-governance.

We argue that the behavioural sciences should play a key role in informing and designing systematic responses to such threats. The role of behavioural science is not only to advance active scientific debates on the causes and reach of false information[21–25] or on whether mass polarization is increasing[26–28]; it is also to find new ways to promote the Internet's potential to bolster rather than undermine democratic societies[29]. Solutions to many global problems—from climate change to the coronavirus pandemic—require coordinated collective solutions, making a democratically interconnected world crucial[30].

Why behavioural sciences are crucial for shaping the online ecosystem

More than any traditional media, online media permit and encourage active behaviours[31] such as information search, interaction and choice. These behaviours are highly contingent on environmental and social structures and cues[32]. Even seemingly minor aspects of the design of digital environments can shape individual actions and scale up to notable changes in collective behaviours. For instance, curtailing the number of times a message can be forwarded on WhatsApp (thereby slowing large cascades of messages) may have been a successful response to the spread of misinformation in Brazil and India[33].

To a substantial degree, social media and search engines have taken on a role as intermediary gatekeepers between readers and publishers. Today, more than half (55%) of global internet users turn to either social media or search engines to access news articles[2]. One implication of this seismic shift is that a small number of global corporations and Silicon Valley CEOs have significant responsibility for curating the general population's information[34] and, by implication, for interpreting discussions of major policy questions and protecting civic freedoms. Facebook's recent decision to declare politicians' ads off-limits to their third-party fact checkers illustrates how corporate decisions can affect citizens' information ecology and the interpretation of fundamental rights, such as freedom of speech. The current situation, in which political content and news diets are curated by opaque and largely unaccountable third parties, is considered unacceptable by a majority of the public[35,36], who continue to be concerned about their ability to discern online what is true and what is false[2] and who rate accuracy as a very important attribute for social media sharing[37].

How can citizens and democratic governments be empowered[38] to create an ecosystem that "values and promotes truth"

[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany. [2]School of Psychological Science and Cabot Institute, University of Bristol, Bristol, UK. [3]School of Psychological Science, University of Western Australia, Perth, Australia. [4]Harvard Law School, Cambridge, MA, USA. ✉e-mail: lorenz-spreen@mpib-berlin.mpg.de

Table 1 | Overview of challenges, cues and potential targets of nudging and boosting interventions in three online contexts

Context	Challenges	Cues	Nudging	Boosting
Online articles	Information overload and fragmentation of sources	Cues to epistemic quality, like cited references	…to pay attention to epistemic cues and external evidence.	…procedures to systematically check epistemic cues.
Algorithmic curation	Asymmetry of knowledge and opaque manipulation	Transparent recommendation and sorting criteria	…awareness of factors that shape recommendations and the news feed.	…self-nudging towards quality information.
Social media	Lack of global network information and false consensus effects	Global social cues that include base rates and passive behaviour	…to consider global social cues and accuracy before sharing.	…to infer credibility from social context and history of content.

(p. 1096)[14]? The answers must be informed by independent behavioural research, which can then form the basis both for improved self-regulation by the relevant companies and for government regulation[39,40]. Regulators in particular face three serious problems in the online domain that underscore the importance of enlisting the behavioural sciences. The first problem is that online platforms can leverage their proprietary knowledge of user behaviour to defang regulations. An example comes from most of the current consent forms under the European Union (EU) General Data Protection Regulation: instead of obtaining genuinely informed consent, the dialogue boxes influence people's decision-making through self-serving forms of choice architecture (for example, consent is assumed from pre-ticked boxes or inactivity)[41,42]. This example highlights the need for industry-independent behavioural research to ensure transparency for the user and to avoid opportunistic responses by those who are regulated. The second problem is that the speed and adaptability of technology and its users exceed that of regulation directly targeting online content. If uninformed by behavioural science, any regulation that focuses only on the symptoms and not on the actual human–platform interaction could be quickly circumvented. The third problem is the risk of censorship inherent in regulations that target content; behavioural sciences can reduce that risk as well. Rather than deleting or flagging posts based on judgements about their content, we focus here on how to redesign digital environments so as to provide a better sense of context and to encourage and empower people to make critical decisions for themselves[43–45].

Our aim is to enlist two streams of research that illustrate the promise of behavioural sciences. The first examines the informational cues that are available online[31] and asks which can help users gauge the epistemic quality of content or the trustworthiness of the social context from which it originated. The second stream concerns the use of meaningful and predictive cues in behavioural interventions. Interventions can take the form of nudging[46], which alters the environment or choice architecture so as to draw users' attention to these cues, or boosting[47], which teaches users to search for them on their own, thereby helping them become more resistant to false information and manipulation, especially but not only in the long run.

Digital cues and behavioural interventions for human-centred online environments

The online world has the potential to provide digital cues that can help people assess the epistemic quality of content[48–50]—the potential of self-contained units of information (here we focus on online articles and social media posts) to contribute to true beliefs, knowledge and understanding—and the public's attitudes to societal issues[51,52]. We classify those cues as endogenous or exogenous[53].

Endogenous cues refer to the content itself, like the plot or the actors and their relations. Modern search engines use natural language-processing tools that analyse content[54]. Such tools have considerable virtues and promise, but current results rarely afford nuanced interpretations[55]. For example, these methods cannot reliably distinguish between facts and opinions, nor can they detect irony, humour or sarcasm[56]. They also have difficulty differentiating between extremist content and counter-extremist messages[57], because both types of messages tend to be tagged with similar keywords. A more general shortcoming of current endogenous cues of epistemic quality is that their evaluation requires background knowledge of the issue in question, which often makes them non-transparent and potentially prone to abuse for censorship purposes.

By contrast, exogenous cues are easier to harness as indicators of epistemic quality. They refer to the context of information rather than the content, are relatively easy to quantify and can be interpreted intuitively. A famous example of the use of exogenous cues is Google's PageRank algorithm, which takes centrality as a key indicator of quality. Well-connected websites appear higher up in search results, irrespective of their content. Exogenous cues can indicate how well a piece of information is embedded in existing knowledge or the public discourse.

From here on we focus on exogenous cues and how they can be enlisted by nudging[46] and boosting[47]. Let us emphasize that a single measure will not reach everyone in a heterogeneous population with diverse motives and behaviours. We therefore propose a range of measures that differ in their scope and in the level of user engagement required. Nudging interventions shape behaviour primarily through the design of choice architectures and typically require little active user engagement. Boosting interventions, in contrast, focus on creating and promoting cognitive and motivational competences, either by directly targeting competences as external tools or indirectly by enlisting the choice environment. They require some level of user engagement and motivation. Both nudging and boosting have been shown to be effective in various domains, including health[58,59] and finances[60]. Recent empirical results from research on people's ability to detect false news indicate that informational literacy can also be boosted[61]. Initial results on the effectiveness of simple nudging interventions that remind people to think about accuracy before sharing content[37] also suggest that such interventions can be effective in the online domain[62]. While empirical tests and evidence are urgently needed, the first step is to outline the conceptual space of possible interventions and make specific proposals.

Table 1 examines three online contexts: articles from newspapers or blogs, algorithmic curation systems that automatically suggest products or information (for example, search engines or algorithmic curation of news feeds), and social media that display information about the behaviour of others (for example, shared posts or social reactions such as comments or 'likes'). Each is associated with a unique set of challenges, cues and potential interventions. Next, we review the challenges and cues in Table 1 and detail some interventions in the subsequent sections.

Online articles: information overload and epistemic cues

The capacity to transfer information online continues to increase exponentially (average annual growth rate: 28%)[63]. Content can

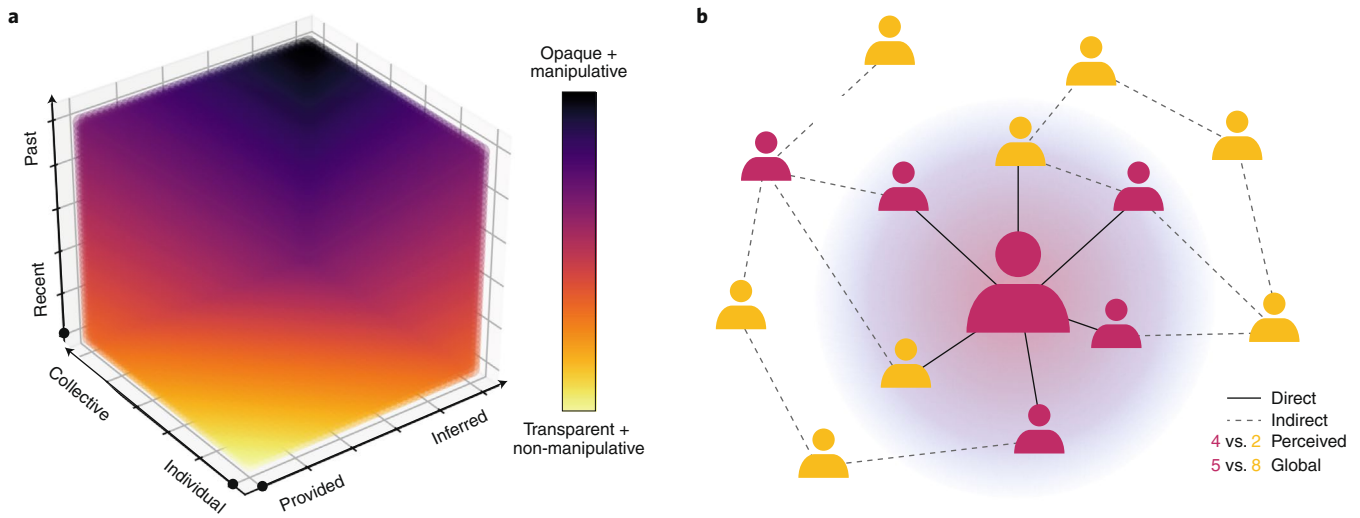

Fig. 1 | Challenges in automatically curated environments and on social media platforms. a, Dimensions of knowledge that platforms can acquire with information technology, which make their recommendations continuously opaque and manipulative. **b**, Perceived group sizes versus the actual global sizes, from the viewpoint of one user (head icon in the centre) in a homophilic social network.

be distributed more rapidly and reaches an audience faster[64]. This increasing pace has consequences. In 2013, a hashtag on Twitter remained in the top 50 most popular hashtags worldwide for an average of 17.5 h; by 2016, a hashtag's time in the limelight had dropped to 11.9 h. The same declining half-life has been observed for Google queries and movie ticket sales[65]. This acceleration, arguably driven by the finite limits of attention available for the ever-increasing quantity of topics and content[66] alongside an apparent thirst for novelty, has significant but underappreciated psychological consequences. Information overload makes it harder for people to make good decisions about what to look at, spend time on, believe and share[67,68]. For instance, longer-term offline decisions such as choosing a newspaper subscription (which then constrains one's information diet) have evolved into a multitude of online microdecisions about which individual articles to read from a scattered array of numerous sources. The more sources crowd the market, the less attention can be allocated to each piece of content and the more difficult it becomes to assess the trustworthiness of each—even more so given the demise and erosion of classic indicators of quality[69] (for example, name recognition, reputation, print quality, price). For this reason, new cues for epistemic quality that are readily accessible even under information overload are necessary. Exogenous cues can highlight the epistemic quality of individual articles, in particular by showing how an article is embedded in the existing corpus of knowledge and public discourse. These cues include, for instance, a newspaper article's sources and citation network (i.e., sources that cite the article or are cited by it), references to established concepts and topical empirical evidence, and even the objectivity of the language.

Algorithmic curation: asymmetry of knowledge and transparency

To help users navigate the overabundance of information, search engines automatically order results[70,71], and recommender systems[72] guide users to content they are likely to prefer[73]. But this convenience exacts a price. Because user satisfaction is not necessarily in line with the goals of algorithms—to maximize user engagement and screen time[74]—algorithmic curation often deprives users of autonomy. For instance, feedback loops are created that can artificially reinforce preferences[75–78], and recommender systems can eliminate context in order to avoid overburdening users. To stay up-to-date and engaging, algorithms can trade recency for importance[79] and, by optimizing on click rates, trade 'clickbait' for quality.

Similarly, aggregated previous user selections make targeted commercial nudging—and even manipulation—possible[80,81]. For example, given just 300 Facebook likes from one person, a regression model can better predict that person's personality traits than friends and family[82]. There are at least three dimensions of knowledge where platforms can far exceed individual human capabilities (Fig. 1a): data that reaches further back in time (for example, years of location history on Google Maps), information about behaviour on a collective rather than an individual level (for example, millions of Amazon customers with similar interests can be used to recommend further products to an individual) and knowledge that is inferred from existing data using machine-learning methods (for example, food preferences inferred from movement patterns between restaurants).

Moving further along these dimensions, it becomes more difficult for a user to comprehend the wealth and predictive potential of this knowledge. Automatic customization of online environments that is based on this knowledge can therefore be opaque and manipulative (Fig. 1a). Recent surveys in the USA and Germany found that a majority of respondents consider such data-driven personalization of political content (61%), social media feeds (57%) and news diets (51%) unacceptable, whereas they are much more accepting of it when it pertains to commercial content[35,36]. To rebalance the relationship between algorithmic and human decision making and to allow for heterogeneous preferences across different domains, a two-step process is required. First, steps should be taken toward the design and implementation of more transparent algorithms. They should offer cues that clearly represent the data types and the weighting that led to a system's suggestions as well as offer information about the target audience. Second, users should be able to adapt these factors to their personal preferences in order to regain autonomy.

Social media: network effects and social cues

More than two thirds of all internet users (around 3 billion people) actively use social media[83]. These platforms offer information about the behaviour of others (for example, likes and emoticons)[84] and new opportunities for interaction (for example, follower relationships and comment sections). However, these signals and interactions

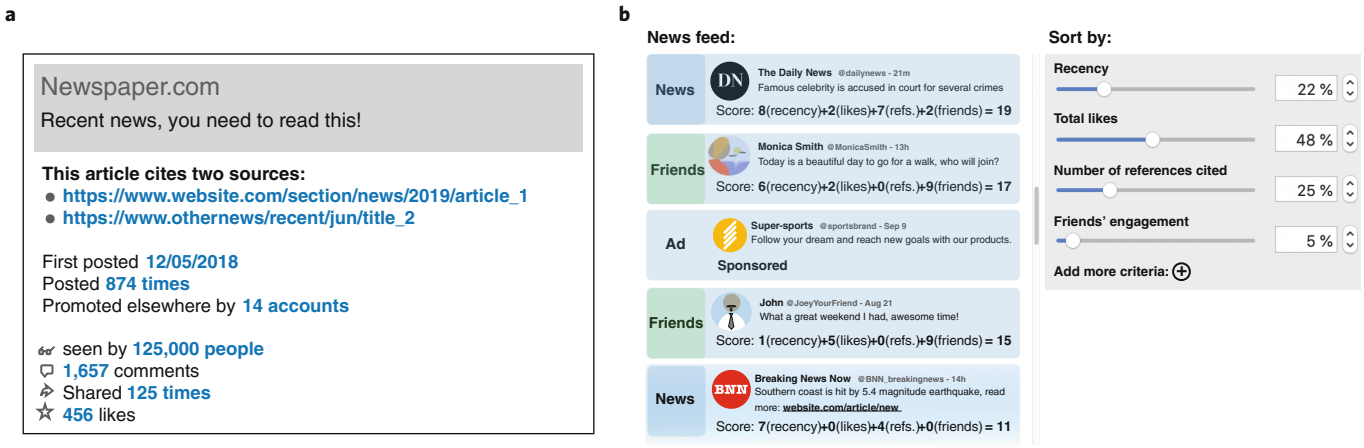

Fig. 2 | Nudging interventions that modify online environments. a, Examples of exogenous cues and how they could appear alongside a social media post. **b**, Example of a transparently organized news feed on social media. Types of content are clearly distinguished, sorting criteria and their values are shown with every post, and users can adjust weightings.

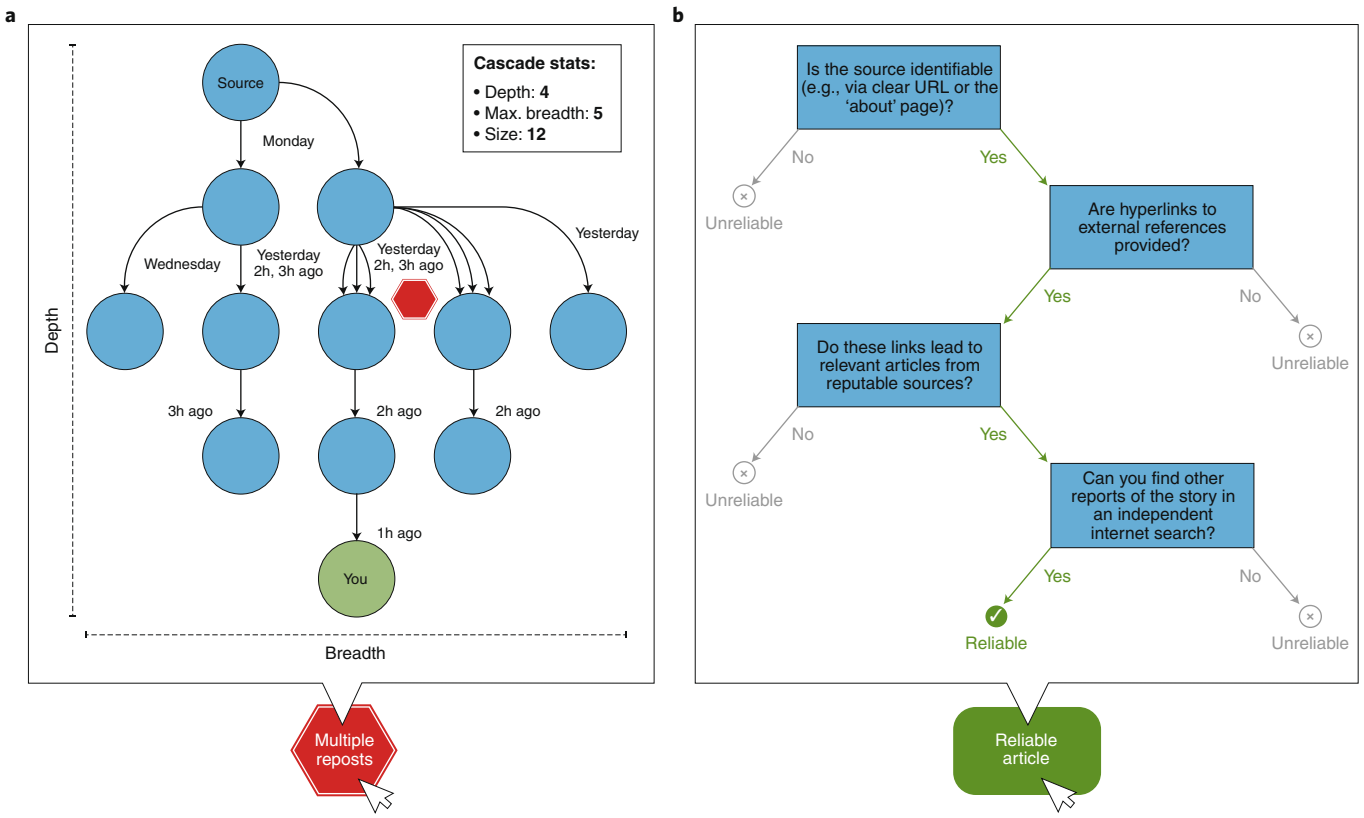

Fig. 3 | Illustrations of boosting interventions as they could appear within an online environment or as external tools. a, Visualization of a sharing cascade. Alongside metrics, like the depth or the breadth of a cascades, a pop-up window on social media can provide a simple visualization of a sharing cascade that shows who (if the profile is public) and when others have shared content before it reached the user. **b**, A fast-and-frugal decision tree as an example of a boosting intervention. A pop-up or an external tool can show a fast-and-frugal decision tree alongside an online article that helps a reader check criteria to evaluate the article's reliability, where the criteria were adapted from professional fact checkers and primarily point to checking external information[90].

are often one-dimensional, represent only a user's immediate online neighbourhood and do not distinguish between different types of connections[85]. These limitations can have drastic effects, such as dramatically changing a user's perception of group sizes[86,87] and giving rise to false-consensus effects (i.e., the majority opinion in an individual's neighbourhood leads people wrongly to believe it reflects the actual majority opinion; Fig. 1b). When people associate with like-minded others from a globally dispersed online community, their self-selected social surroundings (known as a homophilic social network) and the low visibility of the global state of the

network[88,89] can create the illusion of broad support[90] and reinforce opinions or even make them more extreme[91,92]. For instance, even if only a tiny fraction (for example, one in a million) of the more than two billion Facebook users believe that the Earth is flat, they could still form an online community of thousands, thereby creating a shield of like-minded people against corrective efforts[93–96].

Although large social media platforms routinely aggregate information that would foster a realistic assessment of societal attitudes, they currently do not provide a well-calibrated impression of the degree of public consensus[97]. Instead, they show reactions from others as asymmetrically positive—there typically is no 'dislike' button—or biased toward narrow groups or highly active users[98] to maximize user engagement. This need not be the case. The interactive nature of social media could be harnessed to promote diverse democratic dialogue and foster collective intelligence. To achieve this goal, social media needs to offer more meaningful, higher-dimensional cues that carry information about the broader state of the network rather than just the user's direct neighbourhood, which can mitigate biased perceptions caused by the network structure[99]. For instance, social media platforms could provide a transparent crowd-sourced voting system[100] or display informative metrics about the behaviour and reactions of others (for example, including passive behaviour, like the total number of people who scrolled over a post), which might counter false-consensus effects. We note that some platforms have taken steps in the directions we suggest.

Nudging interventions to shape online environments

Nudging interventions can alter choice architectures to promote the epistemic quality of information and its spread. One type of nudge, educative nudging, integrates epistemic cues into the choice environment primarily to inform behaviour (as opposed to actively steering it). For instance, highlighting when content stems from few or anonymous sources (as used by Wikipedia) can remind people to scrutinize content more thoroughly[101,102] and simultaneously create an incentive structure for content producers to meet the required criteria. Such outlets can be made more transparent, for example by disclosing the identity of their confirmed owners. Similarly, pages that are run by state-controlled media might be labelled as such[103]. Going a step further, adding prominent hyperlinks to vetted reference sources for important concepts in a text could encourage a reader to gain context by perusing multiple sources—a strategy used by professional fact checkers[104].

Nudges can also communicate additional information about what others are doing, thereby invoking the steering power of descriptive social norms[105]: For instance, contextualizing the number of likes by expressing them against the absolute frequency of total readers (for example, '4,287 of 1.5 million readers liked this article') might counteract false-consensus effects that a number presented without context ('4,287 people liked this article') may otherwise engender. Transparent numerical formats have already been shown to improve statistical literacy in the medical domain[106]. Similarly, displaying the total number of readers and their average reading time in relation to the potential total readership could help users evaluate the content's epistemic quality: if only a tiny portion of the potential readership has actually read an article, whereas the majority spent just a few seconds on it, it might be clickbait. The presentation of many other cues, including ones that reach into the history of a piece of content, could be used to promote epistemic value on social media. Figure 2a shows a nudging intervention that integrates several exogenous cues into a social media news feed.

Similarly, users could be discouraged from sharing low-quality information without resorting to censorship by introducing 'sludge' or 'friction'—for instance, by making the act of sharing slightly more effortful[107]. In this case, sharing low-quality content may require a further mouse click in a pop-up warning message, alongside additional information about which of the above cues are missing or have critical values.

Another type of nudge targets how content is arranged in browsers. The way a social media news feed sorts content is crucial in shaping how much attention is devoted to particular posts. Indeed, news feeds have become one of the most sophisticated algorithmically driven choice architectures of online platforms[7,108]. Transparent sorting algorithms for news feeds (such as the algorithm used by Reddit) that show the factors that determine how posts are sorted can help people understand why they see certain content; at the very least this nudging intervention would make the design of the feed's architecture more transparent. Relatedly, platforms that clearly differentiate between types of content (for example, ads, news, or posts by friends) can make news feeds more transparent and clearer (Fig. 2b).

Boosting interventions to foster user competences

Boosting seeks to empower people in the longer term by helping them build the competences they need to navigate situations autonomously (for a conceptual map of boosting interventions online, see also ref. [109]). These interventions can be integrated directly into the environment itself or be available in an app or browser add-on. Unlike some nudging interventions, boosting interventions will ideally remain effective even when they are no longer present in the environment, because they have become routinized and have instilled a lasting competence in the user.

The competence of acting as one's own choice architect, or self-nudging, can be boosted[110]. For instance, when users can customize how their news feed is designed and sorted (Fig. 2b), they can become their own choice architects and regain some informational autonomy. For instance, users could be enabled or encouraged to design information ecologies for themselves that are tailored toward high epistemic quality, making sources of low epistemic quality less accessible. Such boosting interventions would require changes to the online environment (for example, transparent sorting algorithms or clear layouts; see previous section and Fig. 2b) and the provision of epistemic cues.

Another competence that could be boosted to help users deal more expertly with information they encounter online is the ability to make inferences about the reliability of information based on the social context from which it originates[111]. The structure and details of the entire cascade of individuals who have previously shared an article on social media has been shown to serve as proxies for epistemic quality[112]. More specifically, the sharing cascade contains metrics such as the depth and breadth of dissemination by others, with deep and narrow cascades indicating extreme or niche topics and breadth indicating widely discussed issues[113]. A boosting intervention could provide this information (Fig. 3a) to display the full history of a post, including the original source, the friends and public users who disseminated it, and the timing of the process (showing, for example, if the information is old news that has been repeatedly and artificially amplified). Cascade statistics teaches concepts that may take some practice to read and interpret, and one may need to experience a number of cascades to learn to recognize informative patterns.

Yet another competence required for distinguishing between sources of high and low quality is the ability to read laterally[104]. Lateral reading is a skill developed by professional fact checkers that entails looking for information on sites other than the information source in order to evaluate its credibility (for example, 'who is behind this website?' and 'what is the evidence for its claims?') rather than evaluating a website's credibility by using the information provided there. This competence can be boosted with simple decision aids such as fast-and-frugal decision trees[114,115]. Employed in a wide range of areas (for example, medicine, finance, law, management), fast-and-frugal decision trees can guide the user to scrutinize relevant cues. For example, users can respond to prompts in a pop-up window (for example, 'are references provided?'), with each answer

leading either to an immediate decision (for example, 'unreliable') or to the next cue until a final judgment about content reliability is reached (for example, 'reliable'; Fig. 3b)[116]. Decision trees can also enhance the transparency of third-party decisions. If reliability is judged by third-party fact checkers or via an automated process, users could opt to see the decision tree and follow the path that led to the decision, thereby gaining insight that will be useful in the long-term. Eventually, fast-and-frugal decision trees may help people establish a habit of checking epistemic cues when reading content even in the absence of a pop-up window suggesting they do so[47].

Finally, the competence of understanding what makes intentionally false information so alluring (for example, novelty and the element of surprise) can be boosted by mental inoculation techniques. Being informed about manipulative methods before encountering them online enables an individual to detect parasitic imitations of trustworthy sources and other sinister tactics[117,118]. Making people aware of such strategies or of their own personal vulnerabilities leaves them better able to identify and resist manipulation. For instance, having people take on the role of a malicious influencer in a computer game has been demonstrated to improve their ability to spot and resist misinformation[61,119]. This inoculation technique can be used in a range of contexts online; for example, learning about the target group of an advertisement can increase people's ability to detect advertising strategies.

Conclusion

Any attempt to regulate or manage the digital world must begin with the understanding that online communication is already regulated, to some extent by public policy and laws but primarily by search engines and recommender systems whose goals and parameters may not be publicly known, let alone subject to public scrutiny. The current online environment has given rise to opaque and asymmetric relationships between users and platforms, and it is reasonable to question whether the industry will take sufficient action on its own to foster an ecosystem that values and promotes truth. The interventions we propose are aimed primarily at empowering individuals to make informed and autonomous decisions in the online ecosystem and, through their own behaviour, to foster and reinforce truth. The interventions are partly conceptualized on the basis of existing empirical findings. However, not all interventions have been tested in the specific context in which they may be deployed. It follows that some of the interventions that we have recommended, and others designed to promote the same goals, should be subject to further empirical testing. Current results identify some interventions as effective[37,119] while also indicating that others are less promising[120]. Both set of results will inform the design of more effective interventions.

In our view, the future task for scientists is to design interventions that meet at least three selection criteria. They must be transparent and trustworthy to the public; standardisable within certain categories of content; and, importantly, hard to game by bad-faith actors or those with vested interests contrary to those of users or society as a whole. We also emphasize the importance of examining a wide spectrum of interventions, from nudges to boosts, to reach different types of people, who have heterogeneous preferences, motivations and online behaviours. These interventions will not completely prevent manipulation or active dissemination of false information, but they will help users recognise when malicious tactics are at work. They will also permit producers of quality information to differentiate themselves from less trustworthy sources. Behavioural interventions in the online ecology can not only inform government regulations, but also signal a platform's commitment to truth, epistemic quality and trustworthiness. Platforms can indicate their commitment to these values by providing their users with exogenous cues and boosting and nudging interventions, and users can choose to avoid platforms that do not offer them these features.

For this dynamic to gain momentum, it is not necessary that all or even the majority of users engage with nudging or boosting interventions. As the first Wikipedia contributors have proven, a critical mass may suffice to allow positive effects to scale up to major improvements. Such a dynamic may counteract a possible drawback of the proposed interventions; namely, widening information gaps between users if only empowered consumers are able to recognise quality information. If a critical mass is created, nudging and boosting interventions might well help to mitigate gaps currently arising from disparities in education or in the ability to pay for quality content. In light of the high stakes—for health, safety and self-governance itself—we err on the side of adopting interventions that empower as many people as possible.

Received: 10 July 2019; Accepted: 23 April 2020;
Published online: 15 June 2020

References

1. Simon, H.A. Designing organizations for an information-rich world. *Computers, Communications and the Public Interest* (ed. Greenberger, M.) 37–72 (1971).
2. Newman, N., Fletcher, R., Kalogeropoulos, A. & Nielsen, R. *Reuters Institute Digital News Report 2019* https://ora.ox.ac.uk/objects/uuid: 18c8f2eb-f616-481a-9dff-2a479b2801d0 (Reuters Institute for the Study of Journalism, 2019).
3. Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl Acad. Sci. USA* **110**, 5802–5805 (2013).
4. Boerman, S. C., Kruikemeier, S. & Zuiderveen Borgesius, F. J. Online behavioral advertising: a literature review and research agenda. *J. Advert* **46**, 363–376 (2017).
5. Ruths, D. & Pfeffer, J. Social media for large studies of behavior. *Science* **346**, 1063–1064 (2014).
6. Tufekci, Z. Engineering the public: big data, surveillance and computational politics. *First Monday* https://doi.org/10.5210/fm.v19i7.4901 (2014).
7. Harris, T. How technology is hijacking your mind—from a magician and Google design ethicist. *Thrive Global* https://thriveglobal.com/stories/how-technology-is-hijacking-your-mind-from-a-magician-and-google-design-ethicist/ (18 May 2016).
8. Persily, N. The 2016 US election: can democracy survive the internet? *J. Democracy* **28**, 63–76 (2017).
9. Habermas, J. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. (MIT Press, 1991).
10. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
11. Mocanu, D., Rossi, L., Zhang, Q., Karsai, M. & Quattrociocchi, W. Collective attention in the age of (mis) information. *Comput. Human Behav.* **51**, 1198–1204 (2015).
12. Rich, M.D. *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life*. (RAND Corporation, 2018).
13. Vargo, C. J., Guo, L. & Amazeen, M. A. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media Soc.* **20**, 2028–2049 (2018).
14. Lazer, D. M. J. et al. The science of fake news. *Science* **359**, 1094–1096 (2018).
15. Baldassarri, D. & Gelman, A. Partisans without constraint: political polarization and trends in American public opinion. *Am. J. Sociol.* **114**, 408–446 (2008).
16. Abramowitz, A. I. & Saunders, K. L. Is polarization a myth? *J. Polit.* **70**, 542–555 (2008).
17. McCarty, N., Poole, K.T. & Rosenthal, H. *Polarized America: the Dance of Ideology and Unequal Riches*. (MIT Press, 2006).
18. Fiorina, M. P. & Abrams, S. J. Political polarization in the American public. *Annu. Rev. Polit. Sci.* **11**, 563–588 (2008).
19. McCright, A. M. & Dunlap, R. E. The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *Sociol. Q.* **52**, 155–194 (2011).
20. Cota, W. et al. Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Sci.* **8**, 35 (2019).
21. DiMaggio, P., Evans, J. & Bryson, B. Have American's social attitudes become more polarized? *Am. J. Sociol.* **102**, 690–755 (1996).
22. Fletcher, R., Cornia, A., Graves, L., & Nielsen, R. K., Measuring the reach of "fake news" and online disinformation in Europe. *Reuters Institute Digital News Publication*. http://www.digitalnewsreport.org/publications/2018/measuring-reach-fake-news-online-disinformation-europe/ (2018).

23. Cinelli, M., Cresci, S., Galeazzi, A., Quattrociocchi, W. & Tesconi, M. The limited reach of fake news on Twitter during 2019 European elections. Preprint at *arXiv* https://arxiv.org/abs/1911.12039 (2020).
24. Guess, A. M., Nyhan, B. & Reifler, J. Exposure to untrustworthy websites in the 2016 US election. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-020-0833-x (2020).
25. Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A. & Bonneau, R. Tweeting from left to right: is online political communication more than an echo chamber? *Psychol. Sci.* **26**, 1531–1542 (2015).
26. Evans, J. H. Have Americans' attitudes become more polarized?—An update. *Soc. Sci. Q.* **84**, 71–90 (2003).
27. Lelkes, Y. Mass polarization: manifestations and measurements. *Public Opin. Q.* **80**, 392–410 (2016).
28. Del Vicario, M. et al. The spreading of misinformation online. *Proc. Natl Acad. Sci. USA* **113**, 554–559 (2016).
29. Watts, D. J. Should social science be more solution-oriented? *Nat. Hum. Behav.* **1**, 15 (2017).
30. Larson, H. J. The biggest pandemic risk? Viral misinformation. *Nature* **562**, 309–310 (2018).
31. Sundar, S. The MAIN model: a heuristic approach to understanding technology effects on credibility. in *Digital Media, Youth, and Credibility* (eds Metzger, M. J. & Flanagin, A. J.) 73–100 (MIT Press, 2007).
32. Gigerenzer, G., Hertwig, R. & Pachur, T. *Heuristics: The Foundations of Adaptive Behavior* (Oxford University Press, 2011).
33. de Freitas Melo, P., Vieira, C.C., Garimella, K., de Melo, P.O.V. & Benevenuto, F. Can WhatsApp counter misinformation by limiting message forwarding? in *International Conference on Complex Networks and Their Applications* 372–384 (2019).
34. Baron-Cohen, S. Keynote address at ADL's 2019 Never Is Now Summit on anti-Semitism and hate. *Anti-Defamation League* https://www.adl.org/news/article/sacha-baron-cohens-keynote-address-at-adls-2019-never-is-now-summit-on-anti-semitism (Accessed 7 December 2019).
35. Kozyreva, A., Herzog, S., Lorenz-Spreen, P., Hertwig, R. & Lewandowsky, S. *Artificial Intelligence in Online Environments: Representative Survey of Public Attitudes in Germany* (Max Planck Institute for Human Development, 2020).
36. Smith, A. *Public Attitudes Toward Computer Algorithms* (Pew Research Center, 2018).
37. Pennycook, G. et al. Understanding and reducing the spread of misinformation online. Preprint at *PsyArXiv* https://psyarxiv.com/3n9u8/ (2019).
38. Zuboff, S. Surveillance capitalism and the challenge of collective action. *New Labor Forum* **28**, 10–29 (2019).
39. Klein, D., & Wueller, J. Fake news: a legal perspective. *J. Internet Law* https://ssrn.com/abstract=2958790 (2017).
40. Assemblée Nationale. Proposition de loi relative à la lutte contre la manipulation de l'information, No. 799 [Proposed Bill on the Fight Against the Manipulation of Information, No. 799] http://www.assemblee-nationale.fr/15/ta/tap0190.pdf (Accessed 26 June 2019).
41. van Ooijen, I. & Vrabec, H. U. Does the GDPR enhance consumers' control over personal data? An analysis from a behavioural perspective. *J. Consum. Policy* **42**, 91–107 (2019).
42. Nouwens, M., Liccardi, I., Veale, M., Karger, D. & Kagal, L. Dark patterns after the GDPR: scraping consent pop-ups and demonstrating their influence. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13 https://doi.org/10.1145/3313831.3376321 (2020).
43. Hertwig, R. When to consider boosting: some rules for policy-makers. *Behav. Public Policy* **1**, 143–161 (2017).
44. Epstein, Z., Pennycook, G. & Rand, D. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on socialmedia by downranking distrusted sources. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–11 https://doi.org/10.1145/3313831.3376232 (2020).
45. Britt, M. A., Rouet, J. F., Blaum, D. & Millis, K. A reasoned approach to dealing with fake news. *Policy Insights Behav. Brain Sci.* **6**, 94–101 (2019).
46. Thaler, R.H. & Sunstein. C. R. *Nudge: Improving Decisions about Health, Wealth, and Happiness* (Yale University Press, 2008)
47. Hertwig, R. & Grüne-Yanoff, T. Nudging and boosting: steering or empowering good decisions. *Perspect. Psychol. Sci.* **12**, 973–986 (2017).
48. Griffiths, K. M. & Christensen, H. Website quality indicators for consumers. *J. Med. Internet Res.* **7**, e55 (2005).
49. Nickel, M., Murphy, K., Tresp, V. & Gabrilovich, E. A review of relational machine learning for knowledge graphs. *Proc. IEEE* **104**, 11–33 (2015).
50. Dong, X. et al. Knowledge Vault: a web-scale approach to probabilistic knowledge fusion. in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 601–610 (2014).
51. Shu, K., Sliva, A., Wang, S., Tang, J. & Liu, H. Fake news detection on social media: A data mining perspective. *SIGKDD Explor.* **19**, 22–36 (2017).
52. Klašnja, M., Barberá, P., Beauchamp, N., Nagler, J. & Tucker, J. Measuring public opinion with social media data. in *The Oxford Handbook of Polling and Survey* Methods (eds Atkeson, L. R. & Alvarez, R. M.) https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190213299.001.0001/oxfordhb-9780190213299-e-3 (2017).
53. Dong, X. L. et al. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings VLDB Endowment* **8**, 938–949 (2015).
54. Hull, J. Google Hummingbird: where no search has gone before. *Wired* https://www.wired.com/insights/2013/10/google-hummingbird-where-no-search- has-gone-before/ (accessed: 9 July 2019).
55. Luo, H., Liu, Z., Luan, H. & Sun, M. Online learning of interpretable word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1687–1692 (2015).
56. Schmidt, A. & Wiegand, M. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10 (2017).
57. Schmitt, J. B., Rieger, D., Rutkowski, O. & Ernst, J. Counter-messages as prevention or promotion of extremism?! The potential role of YouTube: recommendation algorithms. *J. Commun.* **68**, 780–808 (2018).
58. Arno, A. & Thomas, S. The efficacy of nudge theory strategies in influencing adult dietary behaviour: a systematic review and meta-analysis. *BMC Public Health* **16**, 676 (2016).
59. Kurvers, R. H. et al. Boosting medical diagnostics by pooling independent judgments. *Proc. Natl Acad. Sci. USA* **113**, 8777–8782 (2016).
60. Lusardi, A. & Mitchell, O. S. The economic importance of financial literacy: theory and evidence. *J. Econ. Lit.* **52**, 5–44 (2014).
61. Roozenbeek, J. & van der Linden, S. Fake news game confers psychological resistance against online misinformation. *Palgrave Commun.* **5**, 65 (2019).
62. Pennycook, G. & Rand, D. G. Lazy, not biased: susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).
63. Hilbert, M. & López, P. *The world's technological capacity to store, communicate, and compute information. Science* **332**, 60–65 (2011).
64. Rosa, H. *Social Acceleration: A New Theory of Modernity*. (Columbia University Press, 2013).
65. Lorenz-Spreen, P., Mønsted, B. M., Hövel, P. & Lehmann, S. Accelerating dynamics of collective attention. *Nat. Commun.* **10**, 1759 (2019).
66. Wu, F. & Huberman, B. A. Novelty and collective attention. *Proc. Natl Acad. Sci. USA* **104**, 17599–17601 (2007).
67. Hills, T. T., Noguchi, T. & Gibbert, M. Information overload or search-amplified risk? Set size and order effects on decisions from experience. *Psychon. Bull. Rev.* **20**, 1023–1031 (2013).
68. Hills, T. T. The dark side of information proliferation. *Perspect. Psychol. Sci.* **14**, 323–330 (2019).
69. American Society of News Editors (ASNE). ASNE statement of principles. *ASNE.org* https://www.asne.org/content.asp?pl=24& sl=171&contentid=171 (accessed 27 May 2019).
70. Epstein, R. & Robertson, R. E. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proc. Natl Acad. Sci. USA* **112**, E4512–E4521 (2015).
71. Lazer, D. The rise of the social algorithm. *Science* **348**, 1090–1091 (2015).
72. Resnick, P. & Varian, H. R. Recommender systems. *Commun. ACM* **40**, 56–58 (1997).
73. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
74. Martens, Be., Aguiar, L., Gomez-Herrera, E. & Mueller-Langer, F. The digital transformation of news media and the rise of disinformation and fake news. Digital Economy Working Paper 2018–02, *Joint Research Centre Technical Reports*. https://ssrn.com/abstract=3164170 (2018).
75. Cosley, D., Lam, S.K., Albert, I., Konstan, J.A. & Riedl, J. Is seeing believing? How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems* 585–592 (2003).
76. Pan, B. et al. In Google we trust: users' decisions on rank, position, and relevance. *J. Comput. Mediat. Commun.* **12**, 801–823 (2007).
77. Bozdag, E. Bias in algorithmic filtering and personalization. *Ethics Inf. Technol.* **15**, 209–227 (2013).
78. Sunstein, C.R. *Republic.com*. (Princeton University Press, 2002).
79. Chakraborty, A., Ghosh, S., Ganguly, N. & Gummadi, K.P. Optimizing the recency-relevancy trade-off in online news recommendations. In *Proceedings of the 26th International Conference on World Wide Web* 837–846 (2017).
80. Zuboff, S. Big other: surveillance capitalism and the prospects of an information civilization. *J. Inf. Technol.* **30**, 75–89 (2015).
81. Matz, S. C., Kosinski, M., Nave, G. & Stillwell, D. J. Psychological targeting as an effective approach to digital mass persuasion. *Proc. Natl Acad. Sci. USA* **114**, 12714–12719 (2017).

82. Youyou, W., Kosinski, M. & Stillwell, D. Computer-based personality judgments are more accurate than those made by humans. *Proc. Natl Acad. Sci. USA* **112**, 1036–1040 (2015).
83. Ortiz-Ospina, E. The rise of social media. *Our World in Data* https://ourworldindata.org/ rise-of-social-media (accessed: 5 December 2019).
84. Porten-Cheé, P. & Eilders, C. The effects of likes on public opinion perception and personal opinion. *Communications* https://doi.org/10.1515/commun-2019-2030 (2019).
85. Dandekar, P., Goel, A. & Lee, D. T. Biased assimilation, homophily, and the dynamics of polarization. *Proc. Natl Acad. Sci. USA* **110**, 5791–5796 (2013).
86. Lee, E. et al. Homophily and minority-group size explain perception biases in social networks. *Nat. Hum. Behav.* **3**, 1078–1087 (2019).
87. Stewart, A. J. et al. Information gerrymandering and undemocratic decisions. *Nature* **573**, 117–121 (2019).
88. Ross, L., Greene, D. & House, P. The "false consensus effect": an egocentric bias in social perception and attribution processes. *J. Exp. Soc. Psychol.* **13**, 279–301 (1977).
89. Colleoni, E., Rozza, A. & Arvidsson, A. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *J. Commun.* **64**, 317–332 (2014).
90. Leviston, Z., Walker, I. & Morwinski, S. Your opinion on climate change might not be as common as you think. *Nat. Clim. Chang.* **3**, 334–337 (2013).
91. Baumann, F., Lorenz-Spreen, P., Sokolov, I., Starnini, M., Modeling echo chambers and polarization dynamics in social networks. *Phys. Rev. Letters* (in the press).
92. Sunstein, C. R. The law of group polarization. *J. Polit. Philos.* **10**, 175–195 (2002).
93. Sunstein, C.R. *Conspiracy Theories and Other Dangerous Ideas*. (Simon and Schuster, 2014).
94. Van der Linden, S. The conspiracy-effect: exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance. Pers. Individ. Dif. **87**, 171–173 (2015).
95. Lewandowsky, S., Oberauer, K. & Gignac, G. E. NASA faked the moon landing–therefore, (climate) science is a hoax: an anatomy of the motivated rejection of science. *Psychol. Sci.* **24**, 622–633 (2013).
96. Scheufele, D. A. & Krause, N. M. Science audiences, misinformation, and fake news. *Proc. Natl Acad. Sci. USA* **116**, 7662–7669 (2019).
97. Lewandowsky, S., Cook, J., Fay, N. & Gignac, G. E. Science by social media: attitudes towards climate change are mediated by perceived social consensus. *Mem. Cognit.* **47**, 1445–1456 (2019).
98. Muchnik, L., Aral, S. & Taylor, S. J. Social influence bias: a randomized experiment. *Science* **341**, 647–651 (2013).
99. Alipourfard, N., Nettasinghe, B., Abeliuk, A., Krishnamurthy, V. & Lerman, K. Friendship paradox biases perceptions in directed networks. *Nat. Commun.* **11**, 707 (2020).
100. Pennycook, G. & Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl Acad. Sci. USA* **116**, 2521–2526 (2019).
101. Ecker, U. K., Lewandowsky, S. & Tang, D. T. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Mem. Cognit.* **38**, 1087–1100 (2010).
102. Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N. & Cook, J. Misinformation and its correction: continued influence and successful debiasing. *Psychol. Sci. Public Interest* **13**, 106–131 (2012).
103. Rosen, G., Harbath, K., Gleicher, N. & Leathern, R. Helping to protect the 2020 US elections. *Facebook* https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/ (accessed 22 January 2020).
104. Wineburg, S. & McGrew, S. Lateral reading: reading less and learning more when evaluating digital information. Working Paper No 2017.A1/Stanford History Education Group https://ssrn.com/abstract=3048994 (2017).
105. Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J. & Griskevicius, V. The constructive, destructive, and reconstructive power of social norms. *Psychol. Sci.* **18**, 429–434 (2007).
106. Hoffrage, U., Lindsey, S., Hertwig, R. & Gigerenzer, G. Communicating statistical information. *Science* **290**, 2261–2262 (2000).
107. Tucker, J. A., Theocharis, Y., Roberts, M. E. & Barberá, P. From liberation to turmoil: social media and democracy. *J. Democracy* **28**, 46–59 (2017).
108. Facebook for Business. Capturing attention in feed: the science behind effective video creative. https://www.facebook.com/business/news/insights/capturing-attention-feed-video-creative (accessed 8 December 2019).
109. Kozyreva, A., Lewandowsky, S. & Hertwig, R. Citizens versus the internet: confronting digital challenges with cognitive tools. Preprint at *PsyArXiv* https://psyarxiv.com/ky4x8/ (2019).
110. Reijula, S. & Hertwig, R. Self-nudging and the citizen choice architect. *Behav. Publ. Policy* https://doi.org/10.1017/bpp.2020.5 (2020).
111. Noriega-Campero, A. et al. Adaptive social networks promote the wisdom of crowds. *Proc. Natl Acad. Sci. USA* **117**, 11379–11386 (2020).
112. Vosoughi, S. Automatic detection and verification of rumors on Twitter. Doctoral dissertation, Massachusetts Institute of Technology (2015).
113. Zhou, X. & Zafarani, R. Fake news: a survey of research, detection methods, and opportunities. Preprint at *arXiv* https://arxiv.org/abs/1812.00315 (2018).
114. Martignon, L., Katsikopoulos, K. V. & Woike, J. K. Categorization with limited resources: A family of simple heuristics. *J. Math. Psychol.* **52**, 352–361 (2008).
115. Phillips, N. D., Neth, H., Woike, J. K. & Gaissmaier, W. FFTrees: a toolbox to create, visualize, and evaluate fast-and-frugal decision trees. *Judgm. Decis. Mak.* **12**, 344–368 (2017).
116. Banerjee, S., Chua, A. Y. & Kim, J. J. Don't be deceived: using linguistic analysis to learn how to discern online review authenticity. *J. Assoc. Inf. Sci. Technol.* **68**, 1525–1538 (2017).
117. Cook, J., Lewandowsky, S. & Ecker, U. K. H. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS ONE* **12**, e0175799 (2017).
118. Roozenbeek, J. & van der Linden, S. The fake news game: actively inoculating against the risk of misinformation. *J. Risk Res.* **22**, 570–580 (2018).
119. Basol, M., Roozenbeek, J. & van der Linden, S. Good news about bad news: gamified inoculation boosts confidence and cognitive immunity against fake news. *J. Cognition* **3**, 2 (2020).
120. Dias, N., Pennycook, G. & Rand, D. G. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review* https://doi.org/10.37016/mr-2020-001 (2020).

Acknowledgements

We thank A. Kozyreva and S. Herzog for their helpful comments and D. Ain for editing the manuscript. R.H. and S.L. acknowledge support from the Volkswagen Foundation. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

P.L.S., S.L. and R.H. conceptualized the project; P.L.S., S.L., C.R.S. and R.H. wrote the manuscript.

Competing interests

C.R.S. has served as a paid consultant on a few occasions for Facebook.

Additional information

Correspondence should be addressed to P.L.-S.

Peer review information Primary handling editors: Mary Elizabeth Sutherland and Stavroula Kousta

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.