Natural Language Engineering (2023), **29**, pp. 483–508
doi:10.1017/S1351324922000481

CAMBRIDGE UNIVERSITY PRESS

EMERGING TRENDS

Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable

Kenneth Church[1,*], Annika Schoene[1], John E. Ortega[1], Raman Chandrasekar[1] and Valia Kordoni[2]

[1]Institute for Experiential AI, Northeastern University, Boston, MA 02115, USA and [2]Humboldt-Universitaet zuBerlin, Berlin, Germany
[*]Corresponding author. E-mail: k.church@northeastern.edu

(Received 15 November 2022; revised 15 November 2022; accepted 15 November 2022;
first published online 19 December 2022)

Abstract

There has been considerable work recently in the natural language community and elsewhere on Responsible AI. Much of this work focuses on fairness and biases (henceforth Risks 1.0), following the 2016 best seller: *Weapons of Math Destruction*. Two books published in 2022, *The Chaos Machine* and *Like, Comment, Subscribe*, raise additional risks to public health/safety/security such as genocide, insurrection, polarized politics, vaccinations (henceforth, Risks 2.0). These books suggest that the use of machine learning to maximize engagement in social media has created a Frankenstein Monster that is exploiting human weaknesses with persuasive technology, the illusory truth effect, Pavlovian conditioning, and Skinner's intermittent variable reinforcement. Just as we cannot expect tobacco companies to sell fewer cigarettes and prioritize public health ahead of profits, so too, it may be asking too much of companies (and countries) to stop trafficking in misinformation given that it is so effective and so insanely profitable (at least in the short term). Eventually, we believe the current chaos will end, like the lawlessness in Wild West, because chaos is bad for business. As computer scientists, this paper will summarize criticisms from other fields and focus on implications for computer science; we will not attempt to contribute to those other fields. There is quite a bit of work in computer science on these risks, especially on Risks 1.0 (bias and fairness), but more work is needed, especially on Risks 2.0 (addictive, dangerous, and deadly).

Keywords: Bias; Responsible AI; Addictive; Social media; Engagement; Risk; Human weaknesses; Misinformation; Truth effects

1. Introduction

We will start with a brief mention of a few examples of risks/trouble. Of course, there is much more than this. Could there be a single root cause for much of this trouble? And are we (partially) responsible?

After mentioning some of this trouble, and summaries thereof in the press (Section 2.2) and academic literature (Section 2.3), Section 3 will then survey some of the work on Risks 1.0 and 2.0 in our field, computer science. Reporters are accusing us of pivoting when they want to hear what we are doing to address Risks 2.0 (addictive, dangerous, and deadly), and we respond with a discussion of our recent progress on Risks 1.0 (bias and fairness).

Section 4 attempts to identify root causes. It has been suggested that the combination of machine learning and social media has created a Frankenstin Monster that takes advantage of human weaknesses. We cannot put our phones down, even though we know it is bad for us (and

bad for society). Our attempts to build toxicity classifiers (Section 3) and moderation (Section 5) are not effective, given incentives (Section 6). We should not blame consumers of misinformation for their gullibility, or suppliers of misinformation (including adversaries) for taking advantage of the opportunities. There would be less toxicity without market makers creating a market for misinformation and fanning the flames.

Finally, after a discussion of history in Section 7, we will end with constructive suggestions in Section 8. In much of the work that we survey, there tends to be more discussion of problems than solutions. While that may be somewhat depressing, we are pleasantly surprised to see so much pushback from so many directions: governments, users, content providers, academics, consumer groups, advertisers, and employees. Given the high stakes, as well as the challenges, we will need all the help we can get from as many perspectives as possible.[a]

We are even more optimistic about the long term. While trafficking in misinformation may have been insanely profitable thus far, as evidenced by stock market caps, the long-term outlook is less bullish. Recent layoffs at Twitter and Facebook suggest the misinformation business may not continue to be as insanely profitable as it has been. At the end of the day, just as the lawlessness of the Wild West did not last long, this too shall pass. Chaos may be insanely profitable in the short term, but in the long term, chaos is bad for business (and many other parties).

As computer scientists, this paper will survey a diverse set of different perspectives and avoid the temptation to editorialize and advocate our own views. We apologize in advance for so many quotes, citations, and footnotes. As computer scientists, we want to make it clear that we are not experts in all these fields, or that we are entitled to a position on these questions. Our goals are more modest than that. We want to survey criticisms that are out there and suggest that our field should work on a response.

That said, we will suggest that we need more work on both Risks 1.0 (bias and fairness) as well as 2.0 (addictive, dangerous, and deadly). Thus far, there has been quite a bit of work in our field on Risks 1.0. We would like to see more work on Risks 2.0.

2. Risks 1.0 and Risks 2.0

2.1. What happened, and was it our fault?

We will start with a brief discussion of trouble around the world. It might seem that these issues are unrelated, but we fear that there may be a common root cause, and we may have contributed to the problem. Machine learning and social media have been implicated in much of this trouble. Correlated risks are more dangerous than uncorrelated risks. It may even be possible to strengthen claims for correlation to causality, as will be discussed in Section 4.2.

Much has been written about big data and Responsible AI (Zook *et al.* 2017; McNamee 2020). The term, *net neutrality*, was introduced in Wu (2003). Tim Wu has also written about many related issues such as *attention theft*.[b]

Cathy O'Neil (2016) warned us that machine learning is a risk to democracy. Machine learning algorithms are being used to make lots of important decisions like who gets a loan and who gets out of jail. Many of these algorithms are biased and unfair (though perhaps not intentionally so by design). We will refer to these risks as Risks 1.0.

[a]We want to make room for a broad interdisciplinary coalition: *The crossword of nature can only be solved by integration and relentless interaction across disciplines* (Christiansen and Chater 2017).

[b]https://www.wired.com/2017/04/forcing-ads-captive-audience-attention-theft-crime/.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

After Hillary Clinton lost the election to Donald Trump, she asked, "What happened?" (Clinton 2017). There has been considerable discussion of the usual suspects: her emails,[c] Wikileaks,[d] the Russians[e,f] (Aral 2020) and the director of the FBI[g] (Comey 2018). Section 4 will suggest the root cause is actually less malicious, but more insidious.

Shortly after Clinton's book, Madeleine Albright (1937–2022), secretary of state of the United States from 1997 to 2001, warned us about the rise of fascism around the world (Albright and Woodward 2018). A review of her book[h] starts with: *A seasoned US diplomat is not someone you'd expect to write a book with the ominous title Fascism: A Warning*. This review continues with her response to a question about authoritarian leaders creating an anti-democratic spiral [underlining added]:

> *But are we witnessing an anti-democratic spiral? I think so. Some people have said my book is alarmist, and my response is always, "It's supposed to be." We ought to be alarmed by what's happening. Demagogic leaders are taking advantage of all these various factors and using it to divide people further. We should absolutely be alarmed by that.*

2.2. What is the press saying?

Two new books, *The Chaos Machine* (Fisher 2022) and *Like, Comment, Subscribe* (Bergen 2022), raise additional risks to public health/safety/security (henceforth, Risks 2.0) and suggest a connection to social media.

Bergen's book is more about YouTube,[i] with more emphasis on domestic issues in America, especially from a perspective inside YouTube/Google. Fisher's book is more about Facebook[j] than YouTube, with more emphasis on international trouble, from the perspective of a journalist that has covered trouble around the world with his colleague, Amanda Taub. Fisher's book covers much of their reporting in the New York Times,[k,l,m,n,o,p,q,r] a newspaper in the United States.

A review of Bergen's book[s] emphasizes the reference to Frankenstein, as well as misinformation [underlining added]:

> *Bergen, a writer for Bloomberg Businessweek, begins the book with a quote from Mary Shelley's "Frankenstein" and it's easy to see why. From outsized YouTube personalities to misinformation campaigns, YouTube oftentimes comes across as the creature whose makers have lost control.*

[c] https://youtu.be/aOOfwN0iYxM.
[d] https://wikileaks.org/.
[e] https://www.nytimes.com/2017/10/01/technology/facebook-russia-ads.html.
[f] https://intelligence.house.gov/social-media-content/.
[g] https://www.fbi.gov/news/press-releases/press-releases/statement-by-fbi-director-james-b-comey-on-the-investigation-of-secretary-hillary-clinton2019s-use-of-a-personal-e-mail-system.
[h] https://www.vox.com/world/2019/2/14/18221913/fascism-warning-madeleine-albright-book-trump.
[i] https://www.youtube.com/.
[j] https://www.facebook.com/.
[k] https://www.nytimes.com/by/max-fisher.
[l] https://www.nytimes.com/column/the-interpreter.
[m] https://www.nytimes.com/by/amanda-taub.
[n] https://www.nytimes.com/2018/04/22/insider/facebook-victims-sri-lanka.html.
[o] https://www.nytimes.com/2018/08/21/world/europe/facebook-refugee-attacks-germany.html.
[p] https://www.nytimes.com/2018/11/07/technology/personaltech/social-media-effect-myanmar-germany.html.
[q] https://www.nytimes.com/2019/02/12/world/europe/facebook-germany-hate-speech.html.
[r] https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html.
[s] https://apnews.com/article/technology-entertainment-reviews-book-e8935e1eb8b59112f3c104edb25c69cb

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

Fisher's book mentions trouble around the world: *Myanmar*,[t] *Sri Lanka*,[u] *opposition to vaccines*,[v] *climate change denial*,[w] *mass shootings*, *GamerGate*,[x] *Piazzagate*,[y] *QAnon*,[z] *right wing politics* in Germany (AfD)[aa,ab] and America (MAGA),[ac] *Charlottsville*,[ad] *the January 6th Insurrection*,[ae] and more. Social media has been implicated in much of this trouble, as well as troubles that are not mentioned in Fisher's book.[af]

A review of Fisher's book[ag] points out that Fisher is a careful journalist and does not explicitly "assume causality," though causality is strongly implied (as will be discussed in Section 4.2) [underlining added]:

> *Fisher, a New York Times journalist who has reported on horrific violence in Myanmar and Sri Lanka, offers firsthand accounts from each side of a global conflict, focusing on the role Facebook, WhatsApp and YouTube play in fomenting genocidal hate. Alongside descriptions of stomach-churning brutality, he details the viral disinformation that feeds it, the invented accusations, often against minorities, of espionage, murder, rape and pedophilia. But he's careful not to assume causality where there may be mere correlation.*

There is considerable discussion of these topics on the internet in videos, blogs, podcasts, and more.[ah,ai,aj,ak] These topics are also discussed in a popular movie on Netflix, *The Social Dilemma*.[al,am] *Frontline*[an] has a documentary with a similar title, *The Facebook Dilemma*.[ao]

2.3. Academic literature

Fisher and Bergen are journalists. Academics provide a different perspective. There are many academic papers suggesting connections between social media and:

1. *addiction* (Young 1998; Griffiths 2000; Kuss and Griffiths 2011a; Kuss and Griffiths 2011b; Andreassen *et al*. 2012; Pontes and Griffiths 2015; Andreassen 2015; van den Eijnden, Lemmens, and Valkenburg 2016; Andreassen *et al*. 2016; Andreassen, Pallesen, and Griffiths 2017; Courtwright 2019),
2. *misinformation* (Lazer *et al*. 2018; Broniatowski *et al*. 2018; Schackmuth 2018; Vosoughi, Roy, and Aral 2018; Johnson *et al*. 2020; Suarez-Lledo *et al*. 2021),

[t] https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence.
[u] https://www.aljazeera.com/news/2020/5/13/sri-lanka-facebook-apologises-for-role-in-2018-anti-muslim-riots.
[v] https://www.science.org/content/article/vaccine-opponents-are-gaining-facebook-battle-hearts-and-minds-new-map-shows.
[w] https://www.reuters.com/business/cop/facebook-climate-change-can-falsehoods-be-reined-2022-02-23/.
[x] https://en.wikipedia.org/wiki/Gamergate_(harassment_campaign).
[y] https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory.
[z] https://en.wikipedia.org/wiki/QAnon.
[aa] https://www.afd.de/.
[ab] https://blogs.lse.ac.uk/lsereviewofbooks/2013/04/19/book-review-mobilizing-on-the-extreme-right-germany-italy-and-the-united-states/.
[ac] https://moveme.berkeley.edu/project/maga/.
[ad] https://en.wikipedia.org/wiki/Unite_the_Right_rally.
[ae] https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html.
[af] https://issafrica.org/iss-today/social-media-riots-and-consequences.
[ag] https://www.nytimes.com/2022/09/01/books/review/max-fisher-chaos-machine.html.
[ah] https://www.nytimes.com/column/rabbit-hole.
[ai] https://www.youtube.com/c/JordanHarrod.
[aj] https://www.sph.umn.edu/podcast/series-1/episode-8-a-misinformation-pandemic/.
[ak] https://hbr.org/2021/01/how-to-hold-social-media-accountable-for-undermining-democracy.
[al] https://en.wikipedia.org/wiki/The_Social_Dilemma.
[am] https://www.youtube.com/watch?v=yGi2YKZZNFg.
[an] https://www.pbs.org/wgbh/frontline/.
[ao] https://www.youtube.com/watch?v=T48KFiHwexM.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

3. *polarization/homophily*[ap,aq] (McPherson, Smith-Lovin, and Cook 2001; De Koster and Houtman 2008; Colleoni, Rozza, and Arvidsson 2014; Bakshy, Messing, and Adamic 2015; Barberá 2015; Kurka, Godoy, and Zuben 2016; Allcott and Gentzkow 2017; Fourney *et al*. 2017; Ferrara 2017; Bail *et al*. 2018; Grinberg *et al*. 2019; Rauchfleisch and Kaiser 2020; Kaiser and Rauchfleisch 2020; Baptista and Gradim 2020; Zhuravskaya, Petrova, and Enikolopov 2020),
4. *riots/genocide* (Zeitzoff 2017; Hakim 2020),
5. *cyberbullying* (Zych, Ortega-Ruiz, and Rey 2015; Hamm *et al*. 2015; Paluck, Shepherd, and Aronow 2016),
6. *suicide, depression, eating disorders, etc.* (Luxton, June, and Fairall 2012; O'Dea *et al*. 2015; Choudhury *et al*. 2016; Primack *et al*. 2017; Robinson *et al*. 2016),
7. and *insane profits* (Oates 2020).

Many of these topics are discussed in many other places, as well (Ihle 2019; Aral 2020).

2.4. Summary of risks/trouble

Much of the trouble above is associated with misinformation. It is natural to try to fix the problem by going after bias and misinformation with debiasing, fact-checking,[ar] and machine learning (see footnote bd), but that may not work if misinformation is a consequence of some other underlying root cause and/or unfortunate incentives. Debiasing runs into the criticism from the NLP community: removing bias may not reduce inequality (Senthil Kumar *et al.* 2021), and *Awareness is better than blindness* (Caliskan, Bryson, and Narayanan 2017).

To make matters worse, misinformation is creating correlated risks. Correlated risks are worse than uncorrelated risks.[as] Social media helps various small groups find one another; tweets are not i.i.d[at] (Himelboim, McCreery, and Smith 2013; Barberá 2015; Kurka *et al*. 2016). YouTube recommendations are also not i.i.d. (Kaiser and Rauchfleisch 2020). With these new social media technologies, it is no longer necessary for conspirators to conspire with one another explicitly the way they used to do in face-to-face meetings, and over the phone.

3. What are we doing about Risks 1.0 and Risks 2.0?

There is considerable work on Trustworthy computing,[au] Responsible AI[av,aw,ax] and ethics[ay,az] (Blodgett *et al*. 2020; Rogers, Baldwin, and Leins 2021; Church and Kordoni 2021). There is a documentary on PBS, *Coded Bias Documentary — Facial Recognition and A.I. Bias*, directed by Shalini Kantayya.[ba] The following quotes are from a summary of this documentatary:[bb]

> *Over 117 million people in the US has their face in a facial-recognition network that can be searched by the police.*

[ap] Papers on social media and homophily follow an earlier tradition in sociology that predates social media (Fischer 1982).
[aq] https://www.vice.com/de/article/59d98n/youtubes-algorithmen-sorgen-dafur-dass-afd-fans-unter-sich-bleiben.
[ar] https://onlinemasters.ohio.edu/masters-public-administration/guide-to-misinformation-and-fact-checking/.
[as] https://www.guggenheiminvestments.com/mutual-funds/resources/interactive-tools/asset-class-correlation-map.
[at] https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables.
[au] https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/.
[av] https://ai.northeastern.edu/responsible-ai-services/.
[aw] https://orcaarisk.com/.
[ax] https://www.credo.ai/.
[ay] https://lighthouse3.com/our-blog/100-brilliant-women-in-ai-ethics-you-should-follow-in-2019-and-beyond/.
[az] https://aclweb.org/aclwiki/Ethics_in_NLP.
[ba] https://www.pbs.org/independentlens/documentaries/coded-bias/.
[bb] https://lanredahunsi.com/coded-bias-documentary/.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

Table 1. The top 10 papers by citations. The fan-out (right) are more interdisciplinary (and more cited)

Fan-in (citations)		Fan-out (references)	
Citations	Field of study	Citations	Field of study
5512	CS (Brown *et al*. 2020)	10,663	Soc (Crenshaw 1989)
700	CS (Bender *et al*. 2021)	10,018	Psy, Med (Greenwald, McGhee, and Schwartz 1998)
282	CS (Chen *et al*. 2021)	7740	Phil (Davis 1993)
183	CS (Chen *et al*. 2020)	2027	Med (Shprintzen 1990)
173	CS (Gehman *et al*. 2020)	1750	CS, Math (Bolukbasi *et al*. 2016)
162	CS (Chowdhery *et al*. 2022)	1443	CS, Med (Caliskan *et al*. 2017)
100	CS (Dinan *et al*. 2020)	1069	Psy, Med (Heilman *et al*. 2004)
100	CS (Nangia *et al*. 2020)	989	Eng (Fordyce 2019)
89	CS (Delobelle, Winters, and Berendt 2020)	962	Eng (DiSalvo, Clement, and Pipek 2012)
87	CS (Ouyang *et al*. 2022)	788	Soc (Simonsen and Robertson 2013)

Racism is becoming mechanized and robotized.

Power is being wielded through data collection, through algorithms, through surveillance.

Some of the leading figures in this field participated on an ACM Panel Discussion: "From Coded Bias to Algorithmic Fairness: How do we get there?"[bc]

Many of these concerns are relevant to our field. There is a considerable body of work in the ACL community on *bias* (Mitchell *et al*. 2019; Blodgett *et al*. 2020; Bender *et al*. 2021), *fake news detection*,[bd] *hate speech*[be] (Schmidt and Wiegand 2017; Davidson *et al*. 2017), *offensive language* (Zampieri *et al*. 2020), *abusive language* (Waseem *et al*. 2017), and more.

Many classifiers can be found on HuggingFace[bf] and elsewhere.[bg] Unfortunately, as will be discussed in Section 6.1, these classifiers are unlikely to reduce toxicity given current incentives in the social media business to maximize shareholder value.

Table 1 was computed from Blodgett *et al*. (2020), a highly critical survey of work on bias in our field. They categorized 146 papers and concluded: *the vast majority of these papers do not engage with the relevant literature outside of NLP*.

Table 1 uses Semantic Scholar[bh] to provide additional evidence supporting this conclusion. The table shows 10 papers that cite the survey (fan-in) and 10 papers that are cited by the survey (fan-out). Both columns are limited to just the top 10 papers by citations, since there are too many papers to show (according to Semantic Scholar, 446 papers cite the survey and 236 are cited by the survey).

The Semantic Scholar API was also used to estimate fields of study. Note the differences between fan-out (right) and fan-in (left), both in terms of citation counts as well as fields of study. One of the points of the survey is that work in our field should engage more with the relevant

[bc] https://www.youtube.com/watch?v=ji0gMjKFfmI.
[bd] https://paperswithcode.com/task/fake-news-detection.
[be] https://paperswithcode.com/task/hate-speech-detection.
[bf] https://huggingface.co/Hate-speech-CNERG.
[bg] https://www.perspectiveapi.com/research/.
[bh] https://www.semanticscholar.org/product/api.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

(high-impact) literature in other fields. While there is considerable work in our field on bias, our work has relatively little impact beyond computer science.

3.1. Reporters are accusing our field of pivoting

In addition to concerns about engaging with other fields, we would also like to see more work addressing Risks 2.0. Some journalists are accusing the tech community of pivoting when they want to talk about Risks 2.0 (addictive, dangerous, and deadly), and we respond with a discussion of recent progress on work addressing Risks 1.0 (bias and fairness) [underlining added]:

> *But Entin and Quiñonero had a different agenda. Each time I tried to bring up these topics, my requests to speak about them were dropped or redirected. They only wanted to discuss the Responsible AI team's plan to tackle one specific kind of problem: AI bias, in which algorithms discriminate against particular user groups. An example would be an ad-targeting algorithm that shows certain job or housing opportunities to white people but not to minorities.*[bi]

This criticism was followed by an assertion that they (and we) should be prioritizing Risks 2.0 [underlining added]:

> *By the time thousands of rioters stormed the US Capitol in January, organized in part on Facebook... it was clear... the Responsible AI team had failed to make headway against misinformation and hate speech because it had never made those problems its main focus...*

That criticism was followed by an assertion that they were not prioritizing Risks 2.0 because of incentives [underlining added]:

> *The reason is simple. Everything the company does and chooses not to do flows from a single motivation: Zuckerberg's relentless desire for growth.*

There is a growing concern that much of this criticism could also be applied to the NLP community. There is a real danger that the court of public opinion may not view our work on Risks 1.0 as part of the solution and might even see our work as part of the problem. We need to make progress on both Risks 1.0 as well as Risks 2.0.

The books mentioned above (Fisher 2022; Bergen 2022) have quite a bit to say about Risks 2.0. Many academics are mentioned (e.g., Chaslot, DiResta, Farid, Kaiser, Müller, Rauchfleisch, Schwarz), but there is relatively little discussion of our toxicity classifiers. Our classifiers may not have the impact we would hope because there are few incentives for social media companies to reduce toxicity, as will be discussed in Section 6.

4. Root causes

How does fake news spread? It is often suggested that fake news is spread by malicious bots (Bessi and Ferrara 2016; Ferrara 2017) and malicious adversaries[bj] (Aral 2020), perhaps via APIs (Ng and Taeihagh 2021), but the books mentioned above suggest an alternative mechanism. It is suggested that the use of machine learning to maximize engagement may have (accidentally) created a Frankenstein Monster that spreads fake news more effectively than real news[bk] (Vosoughi *et al.* 2018).

[bi] https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

[bj] https://www.theguardian.com/world/2020/jul/21/russia-report-reveals-uk-government-failed-to-address-kremlin-interference-scottish-referendum-brexit.

[bk] https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

4.1. A Frankenstein monster

What do the social media companies have to do with all this trouble? The suggestion is that a number of companies have been working over a number of years on machine learning algorithms for recommending content (Davidson *et al*. 2010; Covington, Adams, and Sargin 2016). They may or may not have intended to create a malicious Frankenstein monster, but either way, they eventually stumbled on a remarkably effective use of persuasive technology (Fogg 2002), Pavlovian conditioning (Rescorla 1988; Bitterman 2006) and Skinner's intermittent variable reinforcement (Skinner 1953; Skinner 1965; Skinner 1986) to take advantage of human weaknesses. Just as casinos take advantage of addicted gamblers, recommender algorithms (Gillespie 2014) know that it is impossible for us to satisfy our cravings for likes. We cannot put our phones down, and stop taking dozens of dopamine hits every day, even though we know it is bad for us (and bad for society)[bl,bm] (Jacobsen and Forste 2011; Junco 2012; Junco and Cotten 2012).

Sean Parker, who had become Facebook's first president at the age of 24 years, put it this way [underlining added]:

> *we are unconsciously chasing the approval of an automated system designed to turn our needs against us* (Fisher 2022) (p. 31)

Fisher's book (Fisher 2022) (pp. 24–25) suggests a connection between Napster's strategy of exploiting a weakness in the music industry and Facebook's strategy of exploiting a weakness in human nature, which they refer to as *the social-validation feedback loop* [underlining added]:

> *Parker had cofounded Napster, a file-sharing program . . . that. . . damaged the music business. . . Facebook's strategy, as he described it, was not so different from Napster's. But rather than exploiting weaknesses in the music industry, it would do so for the human mind. . . "How do we consume as much of your time and conscious attention as possible?" . . . To do that, he said, "We need to sort of give you a little dopamine hit every once in a while, because someone liked or commented on a photo or a post or whatever. And that's going to get you to contribute more content, and that's going to get you more likes and comments." He termed this the "social-validation feedback loop. . ." exploiting a vulnerability in human psychology." He and Zuckerberg "understood this" from the beginning, he said, and "we did it anyway."*

Maximizing engagement brings out the worst in people, with significant risks for public health, public safety, and national security [underlining added]:

> *Either unable or unwilling to consider that its product might be dangerous, Facebook continued expanding its reach in Myanmar and other developing and under-monitored countries. It moored itself entirely to a self-enriching Silicon Valley credo that* [Google's Eric] *Schmidt had recited on that early visit to Yangon: "The answer to bad speech is more speech. More communication, more voices."* (Fisher 2022) (p. 38)

Guillaume Chaslot worked on YouTube's algorithm but was fired because he wanted to make the algorithm less toxic (and less profitable). He has since become an outspoken critic of maximizing engagement [underlining added]:[bn,bo]

> *YouTube was exploiting a cognitive loophole known as the illusory truth effect.* (Fisher 2022) (p. 125)

[bl] https://magenta.as/how-facebook-twitter-and-pinterest-hook-users-5c0eb134992f.
[bm] https://mindmatters.ai/2022/04/how-social-media-are-ruining-our-lives/.
[bn] https://twitter.com/gchaslot.
[bo] https://www.youtube.com/watch?v=Et2n0J0OeQ8.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

Chaslot's mention of "the illusory truth effect" is a reference to the literature on truth effects[bp] (Dechêne *et al.* 2010; Fazio *et al.* 2015; Unkelbach *et al.* 2019). There is a well-known tendency to believe false information to be correct after repeated exposure. The illusory truth effect plays a significant role in such fields as election campaigns, advertising, news media, and political propaganda.[bq]

Eli Pariser coined the term "filter bubble" circa 2010.[br] The Frankenstein monster is creating polarization by giving each of us a personalized view that we are likely to agree with, leading to confirmation bias. He gave a TED Talk on filter bubbles in 2011 [underlining added]:

> *As web companies strive to tailor their services (including news and search results) to our personal tastes, there's a dangerous unintended consequence: We get trapped in a "filter bubble" and don't get exposed to information that could challenge or broaden our worldview. . . this will ultimately prove to be bad for us and bad for democracy.*[bs]

In summary, the root cause of much of the trouble mentioned above is the market maker and the business case, not the suppliers and consumers of misinformation (or even malicious adversaries like the Russian Internet Research Agency).[bt] Trafficking in misinformation is so insanely profitable (at least in the short term) that it is in the market-maker's benefit to do so, despite risks to public health/safety/security. Suppliers and consumers of misinformation (and adversaries) would not do what they are doing if the market-makers did not create the market and fan the flames.

4.2. Causality

Although Fisher does not assume causality, as discussed in Section 2.2, he provides considerable evidence connecting the dots between social media and violence [underlining added]:

> *The country's leaders* [in Sri Lanka], *desperate to stem the violence, blocked all access to social media. It was a lever they had resisted pulling, reluctant to block platforms that some still credited with their country's only recent transition to democracy, and fearful of appearing to reinstate the authoritarian abuses of earlier decades. Two things happened almost immediately. The violence stopped; without Facebook or WhatsApp driving them, the mobs simply went home. And Facebook representatives, after months of ignoring government ministers, finally returned their calls. But not to ask about the violence. They wanted to know why traffic had zeroed out.* (Fisher 2022) (p. 175)

Fisher and Taub provide a second example of causality in footnote o, where they refer to Müller and Schwarz (2021) as "a landmark study," providing strong evidence for causality [underlining added]:

> *This may be more than speculation. Little Altena exemplifies a phenomenon long suspected by researchers who study Facebook: that the platform makes communities more prone to racial violence. And, now, the town is one of 3,000-plus data points in a landmark study* that claims to prove it.

The abstract of Müller and Schwarz (2021) does not mince words: there is a strong assertion of causality [underlining added]:

[bp] https://davidepstein.bulletin.com/why-propaganda-works-the-illusory-truth-effect/.
[bq] https://en.wikipedia.org/wiki/Illusory_truth_effect.
[br] https://en.wikipedia.org/wiki/Filter_bubble.
[bs] https://www.youtube.com/watch?v=B8ofWFx525s.
[bt] https://en.wikipedia.org/wiki/Internet_Research_Agency.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

> *We show that anti-refugee sentiment on Facebook predicts crimes against refugees. . . To establish causality, we exploit exogenous variation in major Facebook and internet outages, which fully undo the correlation between social media and hate crime. . . Our results suggest that social media can act as a propagation mechanism between online hate speech and violent crime.* (Müller and Schwarz 2021)

A third argument for causality involves misinformation in Sri Lanka. This misinformation infected all but the elderly, who are immune to the problem because they are less exposed to Facebook [underlining added]:

> *When I asked Lal and the rest of his family if they believed the posts were true, all but the elderly, who seemed not to follow, nodded.* (Fisher 2022) (p. 167)

4.3. An example of trouble: Vaccines and minority rule

Much has been written about misinformation and vaccines in Nature (Johnson *et al.* 2020) and Chapter 1 of Fisher's book, and elsewhere (Suarez-Lledo *et al.* 2021). Apparently, large majorities support vaccines, and yet, there are schools with low vaccination rates. Misinformation is creating a serious (correlated) risk to public health [underlining added]:[bu,bv]

> *It was 2014. . . and DiResta had only recently arrived in Silicon Valley. . . she began to investigate whether the anti-vaccine anger she'd seen online reflected something broader. Buried in the files of California's public-health department, she realized, were student vaccination rates for nearly every school in the state. . . What she found shocked her. Some of the schools were vaccinated at only 30 percent. . ." She called her state senator's office to ask if anything could be done to improve vaccination rates. It wasn't going to happen, she was told. Were vaccines really so hated? she asked. No, the staffer said. Their polling showed 85 percent support for a bill that would tighten vaccine mandates in schools. But lawmakers feared the extraordinarily vocal anti-vaccine movement. . . seemed to be emerging from Twitter, YouTube, and Facebook. . . Hoping to organize some of those 85 percent of Californians who supported the vaccination bill, she started a group—where else? — on Facebook. When she bought Facebook ads to solicit recruits, she noticed something curious. Whenever she typed "vaccine," or anything tangentially connected to the topic, into the platform's ad-targeting tool, it returned groups and topics that were overwhelmingly opposed to vaccines.* (Fisher 2022) (pp. 13–14)

Similar mechanisms may explain why the minority has such a good chance to control all three branches of the US government (presidency, congress, and the courts) in the near future. Is it possible that the anti-vaccination movement is strong, not because of the facts/science, or the number of supporters, but because of social media? There are some scary precedents where minority rule ended badly.[bw]

4.4. Summary of root causes and precedents

Many of these risks are not new. There has been a long tradition of misinformation, propaganda[bx] and hype (Aral 2020). The "Big Lie" used to refer to Goebbels.[by] Mark Twain is credited with the

[bu] https://news.stanford.edu/2022/02/24/curbing-spread-covid-19-vaccine-related-mis-disinformation/.
[bv] https://cyber.fsi.stanford.edu/io/news/virality-project-final-report.
[bw] https://www.shankerinstitute.org/blog/can-it-happen-here-donald-trump-and-fracturing-americas-constitutional-order-0.
[bx] https://yalereview.org/article/computational-propaganda.
[by] https://www.jewishvirtuallibrary.org/joseph-goebbels-on-the-quot-big-lie-quot.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

aphorism that a lie can travel halfway around the world while the truth is putting on its shoes (Jin *et al*. 2014). There are examples of protest movements that went viral long before Facebook and the Arab spring.[bz]

What is new is the speed and connectivity. With modern technology, a lie can travel faster than ever before.

5. Moderation: An expensive nonsolution

Facebook and YouTube have expensive cost centers that attempt to clean up the mess, but they cannot be expected to keep up with better-resourced profit centers that are pumping out toxic sludge as fast as they can [underlining added]:

> *Some had joined the company thinking they could do more good by improving Facebook from within than by criticizing from without. And they had been stuck with the impossible job of serving as janitors for the messes made by the company's better-resourced, more-celebrated growth teams. As they fretted over problems like anti-refugee hate speech or disinformation in sensitive elections, the engineers across the hall were redlining user engagement in ways that, almost inevitably, made those problems worse.* (Fisher 2022) (p. 261)

Facebook outsources much of the clean-up effort to under-resourced third parties [underlining added]:

> *After a few weeks had passed* [after a violent incident in Sri Lanka], *we asked Facebook how many Sinhalese-speaking moderators they'd hired. The company said only that they'd made progress. Skeptical, Amanda scoured employment websites in nearby countries. She found a listing, in India, for work moderating an unnamed platform in Sinhalese. She called the outsourcing firm through a translator, asking if the job was for Facebook. The recruiter said that it was. They had twenty-five Sinhalese openings, every one unfilled since June 2017 — nine long months earlier. Facebook's "progress" had been a lie.* (Fisher 2022) (p. 177)

Much has been written about moderation in Bergen (2022), Fisher (2022), and elsewhere.[ca,cb] Moderation is unlikely to work, given the lack of incentives [underlining added]:

> *With little incentive for the social media giants to confront the human cost to their empires—a cost borne by everyone else, like a town downstream from a factory pumping toxic sludge into its communal well—it would be up to dozens of alarmed outsiders and Silicon Valley defectors to do it for them.* (Fisher 2022) (pp. 11–12)

Zuckerberg posted a blog on moderation[cc] and reward curves. Engagement (and profits) increase as content comes closer and closer to the line of acceptability, but if content crosses the line, then there will be no engagement after it is censored.

Interestingly, there is less toxicity in China, perhaps because of differences in reward curves. The penalties can be severe in China for coming close to the line, and there is more uncertainty about where the line is. In California, liability lawsuits have been effective in convincing electric companies to prevent forest fires.[cd] Similar methods might convince social media companies to address toxicity. Zuckerberg's blog mentions many suggestions, but not increases in penalties and/or liabilities.

[bz] https://www.economist.com/christmas-specials/2011/12/17/how-luther-went-viral.
[ca] https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation.
[cb] https://www.newyorker.com/tech/annals-of-technology/the-fight-for-the-future-of-youtube.
[cc] https://www.facebook.com/notes/751449002072082/.
[cd] https://www.walkuplawoffice.com/2021/10/06/when-could-an-electric-company-be-liable-for-a-wildfire/.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

5.1. Jacob, a whistleblower

Even if Facebook had been able to hire enough moderators to keep up with better-resourced profit centers pumping out toxic sludge, the moderating task is an impossible task (Fisher 2022) (pp. 4–6) [underlining added]:

> *At the other end of the world, a young man I'll call Jacob, a contractor. . ., had formed much the same suspicions as my own. He had raised every alarm he could. His bosses had listened with concern, he said, even sympathy. They'd seen the same things he had. Something in the product they oversaw was going dangerously wrong. . .*
>
> *Jacob recorded his team's findings and concerns to send up the chain. Months passed. The rise in online extremism only worsened.*
>
> *Jacob first reached me in early 2018. . . Facebook, on learning what I'd acquired, invited me to their sleek headquarters, offering. . . corporate policymakers available to talk.*

5.2. I know it when I see it

Moderators are supposed to follow written rules. It is inevitable that rules become more and more complicated over time.[ce] It might be an impossible task to define rules to cover all imaginable cases across languages, countries, and cultures. The US Supreme Court tried to define obscenity, but eventually, ended up with the famous non-definition: *I know it when I see it*.[cf]

Context often matters. Innocent videos can become not-so-innocent when seen by a different audience from a different perspective. For example, footnote r describes some examples of pedophiles taking advantage of innocent videos of children. Given this reality, it may not be possible for moderators to know it when they see it.[cg]

5.3. Twitter is perhaps more open to moderation

According to Fisher (2022) pp. 219–220 and other sources,[ch] there was a time when Twitter may have been relatively open to addressing toxicity, even if doing so could have led to a reduction in engagement/profits [underlining added]:

> *At Twitter, Dorsey. . . was shifting toward. . . deeper changes. . . that the Valley had long resisted. . . instead of turbocharging its algorithms or retooling the platform to surface argument and emotion, as YouTube and Facebook had done. . ., Dorsey announced. . . social media was toxic. . . The company. . . would reengineer its systems to promote "healthy" conversations rather than engaging ones.*

Unfortunately, the effort failed.[ci] Twitter is smaller than Facebook and YouTube and less profitable, perhaps because Twitter is less committed to the business plan of maximizing engagement

[ce] *the worldwide guides had sprawled to hundreds of confusing and often contradictory pages* (Fisher 2022) (p. 5)

[cf] https://supreme.justia.com/cases/federal/us/378/184/.

[cg] It should be possible to find this kind of misuse using a method like "Communities of Interest" (Cortes, Pregibon, and Volinsky 2001). AT&T used "Communities of Interest" to find fraud remarkably quickly by following the customers of iffy businesses. It should be easy to track demographics of the audience and quickly discover innocent videos with less innocent audiences.

[ch] https://www.businessinsider.com/twitter-ceo-jack-dorsey-on-fixing-fake-news-abuse-2018-3.

[ci] *It was unclear whether Dorsey's experiment in reimagining Twitter had fallen through because his attention drifted, because increasingly rebellious investors pressured Twitter to boost growth instead, or because the solutions proved unpalatable to a company still locked in the Silicon Valley mindset. Accounts from Twitter employees suggest it was likely a combination of all three.* (Fisher 2022) (pp. 219–220)

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

(and toxicity). There is less discussion of Twitter in Fisher (2022): there are 500 mentions of Facebook, 465 mentions of YouTube, and just 169 mentions of Twitter.

It is unclear what will happen to Twitter after the recent acquisition. There have been suggestions that it needs to think more about its long-term strategy:

> *I don't think Elon* [Musk] *has a plan for: this should be a nicer place to be. Because what you need to do is to go from 229M monetizable daily users to 2B.*[cj]

It would be nice if Twitter was a nice place to be. That seems unlikely to happen, especially given recent layoffs (including people working on moderation),[ck] and changes to user verification policies.[cl,cm,cn]

6. Incentives

The problem is that trafficking in misinformation is so insanely profitable.[co,cp] We cannot expect social media companies to regulate themselves.[cq] Most social media companies have an obligation to maximize shareholder value, under normal assumptions.[cr] It is easier for nonprofits such as Wikipedia and Scratch[cs] to address toxicity because nonprofits are not expected to maximize shareholder value.

Competition is forcing a race to the bottom, where everyone has to do the wrong thing. If one company decides to be generous and do the right thing, they will lose out to a competitor that is less generous.[ct]

> *In an unintended 2015 test of this* [race to the bottom], *Ellen Pao, still Reddit's chief, tried something unprecedented: rather than promote superusers, Reddit would ban the most toxic of them. Out of tens of millions of users, her team concluded, only about 15,000, all hyperactive, drove much of the hateful content. Expelling them, Pao reasoned, might change Reddit as a whole. She was right, an outside analysis found. With the elimination of this minuscule percentage of users, hate speech overall dropped an astounding 80 percent among those who remained. Millions of people's behavior had shifted overnight. It was a rare success in combating a problem that would only deepen on other, larger platforms, which did not follow Reddit's lead. They had no interest in suppressing their most active users, much less in acknowledging that there might be such a thing as too much time online.* Fisher (2022) p. 189. [underlining added]

This race to the bottom is described in a segment on the CBS television show, *60 Minutes*, titled "Brain Hacking."[cu] This segment leads with Tristan Harris,[cv] who makes similar points in a TED Talk,[cw] at Stanford[cx] and elsewhere.[cy]

[cj] https://youtu.be/YozxsyHfStg?t=2382.
[ck] https://www.washingtonpost.com/technology/2022/10/20/musk-twitter-acquisition-staff-cuts/.
[cl] https://www.msn.com/en-us/news/technology/key-senator-raises-the-heat-on-musk-after-he-s-impersonated-on-twitter-fix-your-companies-or-congress-will/ar-AA144uWb.
[cm] https://slate.com/technology/2022/11/parody-accounts-of-twitter-blue.html.
[cn] https://www.washingtonpost.com/technology/2022/11/11/twitter-fake-verified-accounts/.
[co] https://theconsciousvibe.com/how-do-social-media-companies-make-money/.
[cp] https://scalar.usc.edu/works/everything-you-always-wanted-to-know-about-social-media-but-were-too-afraid-to-ask/money-and-social-media.
[cq] https://www.nytimes.com/2017/11/19/opinion/facebook-regulation-incentive.html.
[cr] https://www.nytimes.com/2019/08/19/business/business-roundtable-ceos-corporations.html.
[cs] https://scratch.mit.edu/.
[ct] https://www.youtube.com/watch?v=omPSRUmsKZ0.
[cu] https://www.youtube.com/watch?v=awAMTQZmvPE.
[cv] https://www.tristanharris.com/.
[cw] https://www.youtube.com/watch?v=C74amJRp730.
[cx] https://www.youtube.com/watch?v=anEykhlBd-Q.
[cy] http://minimizedistraction.com/.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

If social media companies do not want to reduce toxicity, then it is unlikely to happen. Asking social media companies to reduce toxicity is like the joke about about therapists and light bulbs:

Question: *How many therapists does it take to change a light bulb?*
Answer: *Just one — but the light bulb has to really want to change.*[cz]

The profits are so large that going cold turkey could have serious consequences not only for the companies but also for the national (and international) economy. Of the top 10 stocks by market cap, more than half are technology stocks, and some of their core businesses involve trafficking in misinformation.

6.1. Non-solutions

A number of solutions are unlikely to work given these incentives:

1. Toxicity classifiers (as discussed in Section 3)
2. Just say no[da]

Suppose we were given a magic toxicity classifier that just worked. Given these incentives, the social media company should use the classifier in the reverse direction. That is, rather than use the classifier to minimize toxicity, the social media company should maximize toxicity (in order to maximize profits).

It is also unreasonable to ask companies to cut off their main source of revenue just as we cannot expect tobacco companies to sell fewer cigarettes. The CBS television show, *60 Minutes*, ran similar stories on whistle-blowers in tobacco companies[db] and social media.[dc] In both cases, the companies appeared to know more than they were willing to share about risks to public health and public safety.

6.2. It is easier to say no to noncore businesses

It is hard for a company to shut down its core business. Thus, it may be difficult for Facebook and YouTube to shutdown their core business in social media. Microsoft, on the other hand, is different, because Microsoft is not a social media company. Hany Farid[dd] explained the difference this way:

> *"YouTube is the worst," he said. Of what he considered the four leading web companies—Google/YouTube, Facebook, Twitter, and Microsoft—the best at managing what he'd called "the poison" was, he believed, Microsoft. "And it makes sense, right? It's not a social media company," he said. "But YouTube is the worst on these issues"* Fisher (2022), p. 198.

China also differentiates Microsoft from the others. Of the four companies Farid called out, Microsoft is the only one that is not blocked in China. Many apps are blocked in many countries,[de,df] though there are some interesting exceptions.[dg]

[cz] http://www.takechargecounseling.org/yahoo_site_admin/assets/docs/How_Many_Therapists_Does_It_Take_To_Change_A_Lightbulb.331122841.htm.
[da] https://www.youtube.com/watch?v=lQXgVM30mIY.
[db] https://www.youtube.com/watch?v=1_-Vu8LrUDk.
[dc] https://www.youtube.com/watch?v=_Lx5VmAdZSI.
[dd] https://www.ischool.berkeley.edu/people/hany-farid.
[de] https://www.top10vpn.com/tools/blocked-in-china/.
[df] https://www.reuters.com/technology/china-expresses-serious-concerns-india-banning-chinese-apps-2022-02-17/.
[dg] https://www.reuters.com/markets/europe/minister-says-russia-not-planning-block-youtube-interfax-2022-05-17/.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

7. History

7.1. Precedents and unsuccessful attempts to just say no

There is a long tradition of prioritizing profits ahead of public health and public safety. Consider the Opium Wars and the role of the East India Company and the British Empire in this conflict. According to *Imperial twilight: The opium war and the end of China's last golden age* (Platt 2018) (p. 393 and note 11 on p. 503), the term "opium wars" was coined by *The Times*, the conservative paper in England, in a strongly worded editorial. The conservatives were opposed to the opium trade because of the risk to their core businesses in tea and textiles. History remembers the conservative's sarcastic name for the conflict, even though the conservatives lost the debate in parliament.

When the first author worked at AT&T, they attempted to say no to 976 numbers when they realized that these numbers were being used by iffy businesses in pornography and various scams. AT&T viewed those businesses like *The Times* viewed opium: not profitable enough to justify risks to more important core businesses. AT&T also valued its brand and would not risk it for short-term gains.

AT&T's attempt to end 976 numbers involved a change of numbers, as well as a change in tariffs. The new 900 numbers were tariffed as a joint venture, where AT&T was responsible for billing and transport, and the other company was responsible for content. As a joint venture, AT&T could opt out if it did not approve of the business. The old 976 numbers were tariffed like like sealed box cars, where AT&T was prohibited from breaking the seal. Even if AT&T knew what was inside those box cars, they were required by the tariff to ship the unpleasant cargo.

Unfortunately, the effort was ineffective. While 900 numbers provided AT&T with a legal right to opt out of iffy businesses, there were so many iffy businesses that AT&T was unable to keep up with the problem. The problem eventually became someone else's problem when the internet came along and proved to be a superior technology for iffy businesses.[dh]

7.2. What happened to Our Idealism?

It is hard to remember these days, but there was a time about a decade ago when most of us thought social media technology would make the world a better place.

> *I want to remind us of the awe-inspiring power of the Hype Machine to create positive change in our world. But I have to temper that optimism by noting that its sources of positivity are also the sources of the very ills we are trying to avoid. . . This dual nature makes managing social media difficult. Without a nuanced approach, as we turn up the value, we will unleash the darkness. And as we counter the darkness, we will diminish the value.* (Aral 2020) (pp. 356–357)

What happened to our optimism?

1. The Arab Spring[di] (Howard *et al*. 2011; Fuchs 2012; Smidi and Shahin 2017) was followed by the Arab Winter.[dj,dk]

[dh] https://priceonomics.com/the-rise-and-fall-of-the-1-900-number/.

[di] *Though it's easy to forget now, events like the Arab Spring uprisings of 2011 had been, at the time, viewed as proof of social media's liberating potential.* (Fisher 2022) (pp. 164–165)

[dj] https://en.wikipedia.org/wiki/Arab_Winter.

[dk] *Eventually, the sunny view of the Arab Spring came to be revised. "This revolution started on Facebook," Wael Ghonim, an Egyptian programmer who'd left his desk at Google to join his country's popular uprising, had said in 2011. "I want to meet Mark Zuckerberg someday and thank him personally." Years later, however, as Egypt collapsed into dictatorship, Ghonim warned, "The same tool that united us to topple dictators eventually tore us apart." The revolution had given way to social and religious distrust, which social networks widened by "amplifying the spread of misinformation, rumors, echo chambers, and hate speech," Ghonim said, rendering society "purely toxic."* (Fisher 2022) (pp. 164–165)

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

2. What happened to "Hope and change?" Technology helped elect Obama in 2008 and Trump in 2016.[dl] Why was 2016 different than 2008?
3. "Don't be evil"[dm] became less idealistic ("Move fast and break things"),[dn] more profit driven ("Tech Rules our Economy)"[do] and chaotic (Taplin 2017).

The details are different in each case. Consider Obama's use of technology in 2008. In that case, it is useful to appreciate how long it took to deploy broadband. We tend to think that the roll-out happened quickly, but actually, it took decades. They were connecting about 7M households per year in the United States. About half of the 100M households had broadband in 2008. Since it is cheaper to wire up houses in urban areas, the half with broadband in 2008 overlapped with Obama's base. By 2016, the roll-out was largely completed, eliminating that advantage.

In addition, and more seriously, many of the root causes in Fisher's book became important between 2008 and 2016. Trump benefited in 2016 by maximizing engagement and trafficking in misinformation (McNamee 2020). Many of the details behind Trump's victory involve Cambridge Analytica and data scraped from Facebook[dp,dq,dr] (Wylie 2019). At first, we thought social media technology would benefit positions we agreed with, but more realistically, these forces favor polarization and extremism (O'Callaghan *et al.* 2015; Allcott and Gentzkow 2017) (Fisher 2022) (p. 152).

More generally, when people (and companies) are young, there are more possibilities for growth (and optimism for the future). But as people and companies grow up, there are fewer opportunities for growth, and more downsides. It is natural for the youth to be anti-establishment ("move fast and break things"), and for the establishment to be more risk averse and more realistic and less idealistic.

The word, corporation, is a legal fiction, where companies are treated like people. But there is more to the analogy than that. Start-up companies are like teenagers. After a while, they become middle-aged, and wish they were young again. Companies eventually become senior citizens. Seniors are not as agile as they used to be. Growth stocks eventually become value stocks. Social media will eventually grow up and become a utility.

8. Constructive suggestions

What can we do about this nightmare? We view the current chaos like the Wild West. Just as that lawlessness did not last long because it was bad for business, so too, in the long run, the current chaos will be displaced by more legitimate online businesses.

What can we do in the short term? Many of the books mentioned above (O'Neil 2016; Aral 2020; Fisher 2022; Bergen 2022) have more to say about the problem than the solution. For an example of how we can be part of the solution,[ds] read Chapter 5 of *Zucked* (McNamee 2020), *Mr. Harris and Mr. McNamee Go to Washington*. (There is a condensed version of McNamee's book on *Democracy Now!*)[dt,du]

[dl] https://www.newyorker.com/magazine/2020/03/09/the-man-behind-trumps-facebook-juggernaut.
[dm] https://gizmodo.com/google-removes-nearly-all-mentions-of-dont-be-evil-from-1826153393.
[dn] https://hbr.org/2019/01/the-era-of-move-fast-and-break-things-is-over.
[do] https://youtu.be/E0iRuULJr7g?t=262.
[dp] https://www.npr.org/2019/10/08/768216311/whistleblower-explains-how-cambridge-analytica-helped-fuel-u-s-insurgency.
[dq] https://www.youtube.com/watch?v=wqxx_Ixo1bo.
[dr] https://www.npr.org/2019/10/08/767293251/in-new-book-cambridge-analytica-whistleblower-stops-short-of-a-full-mea-culpa.
[ds] https://www.youtube.com/watch?v=V4Hdn8cgtCU
[dt] https://www.youtube.com/watch?v=ndOsqevLOME.
[du] https://www.youtube.com/watch?v=I-VheZFinX8.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

As an early investor in Facebook, Roger McNamee has connections to the Facebook leadership. He tried to use those connections to raise awareness within Facebook. When that failed, he published an op-ed[dv] and worked with Tristan Harris on the TED Talk mentioned in footnote cw. When those efforts failed to raise enough awareness to make meaningful progress, Harris and McNamee went to Washington, and were more successful there, as described in Chapter 6 of his book, *Congress Gets Serious*.

McNamee seems to be having more success in government than with the Facebook leadership. His efforts may or may not succeed, but either way, we respect his persistence, as well as his emphasis on constructive solutions. He wrote a piece for law-makers with a title that emphasizes fixes: *How to Fix Facebook—Before It Fixes Us*.[dw] This piece was not only effective with law-makers, but it also reached Soros, who gave a speech at Davos along similar lines.[dx,dy] The text of Soros's remarks can be found in Appendix 2 of McNamee (2020). Facebook tends to ignore such criticisms, but they are not ignoring Soros.[dz]

McNamee (p. 231) describes a simple project involving word associations. This project could be a good exercise for students in our classes. He suggests that Facebook's brand suffered since the 2016 election and provides evidence involving associations with pejorative words such as: *scandal, breach, investigation, fake, Russian, alleged, critical, false, leaked, racist*. It should be relatively easy for students in our classes to use word associations and deep nets (BERT) to track sentiment toward various brands as a function of time.

8.1. Pushback from many perspectives

As mentioned above, we are pleasantly surprised to see so much pushback from so many directions: governments, users, content providers, academics, consumer groups, advertisers and employees. Given the high stakes, as well as the challenges, we will need all the help we can get from so many different perspectives:

1. Pushback from government(s): (See Section 8.3) Regulation, anti-trust, bans, taxes, fines, liability, data privacy, education[ea]
2. Pushback from users and their friends and family (including parents, children and peers)[eb,ec] (Allcott *et al*. 2020)
3. Pushback from investors[ed] (McNamee 2020)
4. Pushback from content providers[ee]
5. Pushback from media[ef] (Bergen 2022; Fisher 2022)
6. Pushback from academics (Aral 2020)

[dv] https://www.usatoday.com/story/opinion/2017/08/08/my-google-and-facebook-investments-made-fortune-but-now-they-menace/543755001/.

[dw] https://washingtonmonthly.com/2018/01/07/how-to-fix-facebook-before-it-fixes-us/.

[dx] https://www.youtube.com/watch?v=WaHzUlR2MUg.

[dy] https://www.fastcompany.com/90296585/roger-mcnamee-bet-on-zuckerberg-helped-write-the-soros-speech-that-skewered-facebook.

[dz] https://www.youtube.com/watch?v=0RqgMYWwh2A.

[ea] https://scoop.upworthy.com/students-learn-empathy-in-denmark-schools

[eb] https://www.theguardian.com/technology/2019/feb/01/facebook-mental-health-study-happiness-delete-account.

[ec] https://www.builtinla.com/2018/01/11/dopamine-labs-boost-user-engagement.

[ed] https://www.gsb.stanford.edu/insights/roger-mcnamee-facebook-terrible-america.

[ee] https://www.nbcnews.com/news/us-news/youtube-shooter-nasim-aghdam-was-vegan-who-had-complained-about-n862586.

[ef] https://www.theguardian.com/commentisfree/2021/jun/27/case-for-brexit-built-on-lies-five-years-later-deceit-is-routine-in-our-politics.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

7. Pushback from consumer groups,[eg,eh] activists,[ei,ej] and consultants.[ek,el]
8. Pushback from advertisers[em] (Fisher 2022) (p. 311)
9. Pushback from employees (see Section 8.2)
10. Economic auctions: Google's Ad Auction finds an equilibrium satisfying the needs of three parties:[en] readers, writers, and advertisers. YouTube's "audience first" strategy[eo] prioritizes the audience ahead of other parties. Perhaps they would have more success with an auction that addresses the needs of more parties. The market maker should not favor one party over the others: *You're Not the Customer; You're the Product*.[ep]

There is considerable discussion of many of the suggestions above, though so far, there are relatively few examples that are as successful as face recognition. Following concerns in Buolamwini and Gebru (2018), there are limits on the use of face recognition technology involving a combination of legislation[eq,er] and voluntary actions.[es,et]

8.2. Pushback from employees

Pushback from employees is already happening and may be more effective than most of the suggestions above.

Employees are writing books (Martinez 2018) and participating in documentaries such as *The Social Dilemma*, mentioned in Section 2.2. Tristan Harris, for example, has been mentioned several times above.

There are a number of quotes from employees in Fisher (2022):

> *"Can we get some courage and actual action from leadership in response to this behavior?" a Facebook employee wrote on the company's internal message board as the riot unfolded. "Your silence is disappointing at the least and criminal at worst."* (Fisher 2022) (p. 325)

Polls of employees confirm this sentiment:[eu]

> *"When I joined Facebook in 2016, my mom was so proud of me," a former Facebook product manager told Wired magazine. "I could walk around with my Facebook backpack all over the world and people would stop and say, 'It's so cool that you worked for Facebook.' That's not the case anymore." She added, "It made it hard to go home for Thanksgiving."*[ev] (Fisher 2022) (p. 248)

[eg] https://adage.com/article/privacy-and-regulation/youtubes-pushback-kids-privacy-criticized-consumer-groups/2222061.
[eh] https://techcrunch.com/2019/02/12/jim-steyer-runs-the-powerful-nonprofits-common-sense-media-and-hes-increasingly-using-his-influence-around-tech-consumption/.
[ei] https://www.ajl.org/.
[j] https://www.humanetech.com/who-we-are.
[ek] https://orcaarisk.com/.
[el] https://www.credo.ai/.
[em] https://www.nytimes.com/2020/08/01/business/media/facebook-boycott.html.
[en] https://www.youtube.com/watch?v=a8qQXLby4PY.
[eo] https://www.thinkwithgoogle.com/intl/en-154/marketing-strategies/video/audience-first-strategy-youtube/.
[ep] https://quoteinvestigator.com/2017/07/16/product/.
[eq] https://www.seattletimes.com/seattle-news/politics/washington-senate-passes-bill-to-regulate-governments-use-of-facial-recognition-technology/.
[er] https://www.nytimes.com/2021/02/27/technology/Massachusetts-facial-recognition-rules.html.
[es] https://learn.microsoft.com/en-us/legal/cognitive-services/computer-vision/limited-access-identity.
[et] https://www.nytimes.com/2022/06/21/technology/microsoft-facial-recognition.html.
[eu] *. . . the share of employees who said they were proud to work at Facebook had declined from 87 to 70 percent in just a year. The share who felt their company made the world a better place had dropped from 72 to 53 percent, and on whether they felt optimistic about Facebook's future, from the mid-80s to just over 50 percent.* (Fisher 2022) (p. 248)
[ev] https://www.wired.com/story/facebook-mark-zuckerberg-15-months-of-fresh-hell/.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

There have been a number of other examples of pushback from employees in the news recently, starting with Uber.[ew] More recently, Facebook employees wrote an open letter to Zuckerberg.[ex] Even more recently, Google has been in the news.[ey,ez,fa,fb,fc,fd]

8.3. Regulation

There must be a way to make it less insanely profitable to traffic in misinformation. Regulators should "Follow the money"[fe] and "take away the punch bowl."[ff] The risks to public health, public safety, and national security are too great (Oates 2020).

Regulation can come in many forms: anti-trust, censorship, bans, rules about data privacy, taxes, and liabilities. Europe has been leading the way on regulation (Section 8.3.1), especially when compared to America (Section 8.3.2).

8.3.1. Regulation in the European Union

Regulation is taken very seriously in Europe. As mentioned above, there are strong data privacy laws such as the GDPR,[fg] and companies have been fined.[fh] There are also rules to combat fake news, hate speech, and misinformation such as the Network Enforcement Act (Netzwerkdurchsetzungsgesetz), known colloquially as the Facebook Act.[fi,fj] Even stronger regulation is under discussion.[fk] Many people are involved in these discussions, including members of our field.

8.3.2. Less regulation in the United States

There is less regulation in the United States than in Europe [underlining added]:

> *Some agencies, such as the Food and Drug Administration or the Department of Transportation, have been working for years to incorporate AI considerations into their regulatory regimes. In late 2020, the Trump Administration's Office of Management and Budget encouraged agencies to consider what regulatory steps might be necessary for AI, although it generally urged a light touch.*[fl]

There is a link from the final phrase, *light touch*, to a strong criticism of the lack of regulation of AI in the United States under the Trump administration. This criticism ends with:

[ew] https://www.susanjfowler.com/blog/2017/2/19/reflecting-on-one-very-strange-year-at-uber.
[ex] https://www.nytimes.com/2019/10/28/technology/facebook-mark-zuckerberg-letter.html.
[ey] https://www.theguardian.com/technology/2021/feb/19/google-fires-margaret-mitchell-ai-ethics-team.
[ez] https://www.theguardian.com/technology/2021/feb/04/google-timnit-gebru-ai-engineers-quit.
[fa] https://www.theguardian.com/technology/2018/nov/01/google-walkout-global-protests-employees-sexual-harassment-scandals.
[fb] https://googlewalkout.medium.com/standing-with-dr-timnit-gebru-isupporttimnit-believeblackwomen-6dadc300d382.
[fc] https://www.theguardian.com/technology/2020/dec/02/google-labor-laws-nlrb-surveillance-worker-firing.
[fd] https://www.theguardian.com/technology/2018/nov/01/google-walkout-global-protests-employees-sexual-harassment-scandals.
[fe] https://en.wikipedia.org/wiki/Follow_the_money.
[ff] https://thehill.com/opinion/finance/564743-time-for-the-fed-to-take-away-the-punchbowl/.
[fg] https://gdpr.eu/.
[fh] https://www.theguardian.com/technology/2022/jan/06/france-fines-google-and-facebook-210m-over-user-tracking-cookies.
[fi] https://en.wikipedia.org/wiki/Network_Enforcement_Act.
[fj] https://www.br.de/puls/themen/netz/hate-speech-maas-gesetz-100.html.
[fk] https://artificialintelligenceact.eu/.
[fl] https://www.brookings.edu/blog/techtank/2022/02/01/the-eu-and-u-s-are-starting-to-align-on-ai-regulation/.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

> *Yet there is a real risk that this document becomes a force for maintaining the status quo, as opposed to addressing serious AI harms.*[fm]

In the United States, there has been more regulation of health care and energy than Artificial Intelligence (Goralski and Górniak-Kocikowska 2022; Munoz and Maurya 2022), though there is an effort to increase regulation of AI under the Biden administration.[fn] There are also efforts at the state level to regulate AI,[fo] privacy,[fp] self-driving cars,[fq] and facial recognition and biometrics.[fr]

Congress may take action based on anti-trust considerations:[fs]

> *The effects of this significant and durable market power are costly. The Subcommittee's series of hearings produced significant evidence that these firms wield their dominance in ways that erode entrepreneurship, degrade Americans' privacy online, and undermine the vibrancy of the free and diverse press. The result is less innovation, fewer choices for consumers, and a weakened democracy.*

There are some strong advocates of anti-trust.[ft] That said, even the threat of anti-trust action can be effective.[fu] One the other hand, anti-trust will take time, as pointed out in Chapter 12 of Aral (2020).

Realistically, it is unlikely that regulation will succeed in America as long as one party or the other believes that the status quo is in their best interest. As discussed in Section 4.3, trafficking in misinformation enables minority rule.[fv] With help from social media, it is likely that all three branches of the US government (executive branch, congress, and courts) will be captured by less than 50% of the voters.[fw]

8.3.3. Data privacy

There is considerable discussion of data privacy in McNamee (2020). Doctors and lawyers are not allowed to sell personal data (p. 226). So too, social media companies should be liable for inappropriate disclosures, as would be expected in many industries: medicine, banking, etc. McNamee advocates for a fiduciary rule; companies would more careful if consumers had the right to sue.

McNamee also advises social media companies to cooperate with privacy laws, but that is unlikely to happen [underlining added]:

[fm] https://www.brookings.edu/blog/techtank/2020/12/08/new-white-house-guidance-downplays-important-ai-harms/.

[fn] https://www.whitehouse.gov/ostp/news-updates/2022/10/04/fact-sheet-biden-harris-administration-announces-key-actions-to-advance-tech-accountability-and-protect-the-rights-of-the-american-public/.

[fo] https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx.

[fp] https://www.ncsl.org/research/telecommunications-and-information-technology/2022-consumer-privacy-legislation.aspx.

[fq] https://www.ncsl.org/research/transportation/autonomous-vehicles.aspx.

[fr] https://www.ncsl.org/research/civil-and-criminal-justice/facial-recognition-and-biometrics.aspx.

[fs] https://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf.

[ft] https://www.warren.senate.gov/newsroom/press-releases/warren-delivers-remarks-at-freedom-from-facebook-and-google-break-up-big-tech.

[fu] *The ruling was thrown out on appeal—the judge had discussed the case with reporters as it proceeded, tainting his impartiality—and the incoming Bush administration dropped the case. Still, Microsoft's stock price had been halved; its entry to internet services cut short, never to recover; and its standing with the public and regulators so weakened that Bill Gates, its founder, stepped down. Years later, he would counsel Zuckerberg not to repeat what he saw as his mistake: antagonizing Washington and ignoring lawmakers he saw as wrongheaded. "I said, 'Get an office there, now,'" Gates recalled, referring to Washington, where Facebook and Google began spending millions on lobbying. "And Mark did, and he owes me."* Fisher (2022) (pp. 255–256)

[fv] https://www.politifact.com/article/2022/jun/14/most-republicans-falsely-believe-trumps-stolen-ele/.

[fw] https://www.npr.org/2021/06/09/1002593823/how-democratic-is-american-democracy-key-pillars-face-stress-tests.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

> *the target industry is usually smart to embrace the process early, cooperate, and try to satisfy the political needs of policy makers before the price gets too high. For Facebook and Google, the first "offer" was Europe's General Data Protection Regulation (GDPR). Had they embraced it fully, their political and reputational problems in Europe would have been reduced dramatically, if not eliminated altogether. For reasons I cannot understand, both companies have done the bare minimum to comply with the letter of the regulation, while blatantly violating the spirit of it.* (McNamee 2020) (p. 221)

The American companies should show more respect to the regulators in Europe. As discussed in Section 8.3.1, there have already been a few fines, and there will be more.

Much has been written about data privacy in different parts of the world. It is said that there is relatively little privacy in China, but the PIPL[fx] in China is similar to the GDPR in Europe and the CCPA[fy] in California. Some companies and some countries are more careful with data than others. The first author has worked for a number of companies. In his experience, his employer in China (Baidu) is more careful than his employers in America. The penalties for inappropriate disclosures can be severe in China. The American government seems to be able to get what it wants.[fz] Other governments may be similar. It is not clear how users can defend themselves from a government or a sophisticated adversary.[ga] One could avoid the use of cell phones, and connections to the internet, as is standard practice in a SCIF,[gb] but it is hard to imagine that most of us would be willing to do that.

9. Conclusions: Trafficking in misinformation is insanely profitable in the short term, but bad for business in the long term

We have discussed Risks 1.0 (fairness and bias) and Risks 2.0 (addictive, dangerous, and deadly). The combination of machine learning and social media has created a Frankenstein Monster that uses persuasive technology, the illusory truth effect, Pavlovian conditioning, and Skinner's intermittent variable reinforcement to take advantage of human weaknesses and biases. We cannot put our phones down, even though we know it is bad for us and bad for society. The result is insanely profitable, at least in the short term.

Much of the trouble mentioned above is caused by the market maker and the business strategy, not the suppliers and consumers of misinformation, or even malicious adversaries. We should not blame consumers of misinformation for their gullibility, or suppliers of misinformation (including adversaries) for taking advantage of the opportunities. Without the market makers creating a market for misinformation, and fanning the flames, there would be much less toxicity. Regulators can help by making it less insanely profitable to traffic in misinformation. "Follow the money" and "take away the punch bowl."

There has been considerable discussion of these issues in our community. There are a number of toxicity classifiers on HuggingFace. Unfortunately, given short-term incentives to maximize shareholder value (and maximize engagement), it is unlikely that such classifiers could be effective without first convincing the social media companies that it is in their interest to reduce toxicity. Moreover, as discussed in Section 3.1, there is a risk that work on Risks 1.0 (toxicity classifiers) could be seen as an attempt to pivot away from Risks 2.0 (addictive, dangerous, and deadly). We need to address both Risks 1.0 and Risks 2.0.

[fx] https://www.cooley.com/news/insight/2021/2021-11-30-china-new-national-privacy-law.
[fy] https://oag.ca.gov/privacy/ccpa.
[fz] https://www.nytimes.com/2015/08/16/us/politics/att-helped-nsa-spy-on-an-array-of-internet-traffic.html.
[ga] https://www.aljazeera.com/news/2022/2/8/what-you-need-to-know-about-israeli-spyware-pegasus.
[gb] https://csrc.nist.gov/glossary/term/sensitive_compartmented_information_facility.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

We are more optimistic about the long term. Assuming that markets are efficient, rational, and sane, at least in the long term at steady state, then insane profits cannot continue for long. There are already hints that the short-term business case may be faltering at Twitter (as discussed in Section 5.3) and at Facebook.[gc,gd,gegf,gg,gh,gi,gj,gk] You know it must be bad for social media companies when *The Late Show with Stephen Colbert* is making jokes at their expense.[gl] The telephone monopoly was broken up soon after national television made jokes at their expense.[gm] As discussed in Section 7.1, AT&T valued its brand and would not risk the brand for short-term gains. Social media companies should be more risk-averse with their brand.

Just as the lawlessness of the Wild West did not last long, this too shall pass. The current chaos is not good for business (and many other parties). We anticipate a sequel to "How the West Was Won"[gn] entitled "How the Web Was Won," giving a whole new meaning to: WWW.

References

Albright M. and **Woodward B.** (2018). *Fascism*. New York, NY, USA: HarperCollins.

Allcott H., **Braghieri L.**, **Eichmeyer S.** and **Gentzkow M.** (2020). The welfare effects of social media. *American Economic Review* **110**(3), 629–676.

Allcott H. and **Gentzkow M.** (2017). Social media and fake news in the 2016 election. *CSN: Politics (Topic)*.

Andreassen C.S. (2015). Online social network site addiction: A comprehensive review. *Current Addiction Reports* **2**(2), 175–184.

Andreassen C.S., **Billieux J.**, **Griffiths M.D.**, **Kuss D.J.**, **Demetrovics Z.**, **Mazzoni E.** and **Pallesen S.** (2016). The relationship between addictive use of social media and video games and symptoms of psychiatric disorders: A large-scale cross-sectional study. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors* **30**(2), 252–262.

Andreassen C.S., **Pallesen S.** and **Griffiths M.D.** (2017). The relationship between addictive use of social media, narcissism, and self-esteem: Findings from a large national survey. *Addictive Behaviors* **64**, 287–293.

Andreassen C.S., **Torsheim T.**, **Brunborg G.S.** and **Pallesen S.** (2012). Development of a facebook addiction scale. *Psychological Reports* **110**, 5010–5517.

Aral S. (2020). *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health–and How We Must Adapt*. New York, NY, USA: Currency, Penguin Random House LLC.

Bail C.A., **Argyle L.P.**, **Brown T.W.**, **Bumpus J.P.**, **Chen H.**, **Hunzaker M.F.**, **Lee J.**, **Mann M.**, **Merhout F.** and **Volfovsky A.** (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115(37), 9216–9221.

Bakshy E., **Messing S.** and **Adamic L.A.** (2015). Exposure to ideologically diverse news and opinion on facebook. *Science* **348**, 1130–1132.

Baptista J.P. and **Gradim A.** (2020). Online disinformation on facebook: The spread of fake news during the portuguese 2019 election. *Journal of Contemporary European Studies* **30**, 297–312.

Barberá P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis* **23**, 76–91.

Bender E.M., **Gebru T.**, **McMillan-Major A.** and **Shmitchell S.** (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.

Bergen M. (2022). *Like, Comment, Subscribe: Inside YouTube's Chaotic Rise to World Domination*. New York: Viking.

Bessi A. and **Ferrara E.** (2016). Social bots distort the 2016 US presidential election online discussion. *First Monday* **21**(11–7).

[gc] https://www.washingtonpost.com/technology/2022/11/09/facebook-layoffs/.

[gd] https://medium.com/@alt.cap/time-to-get-fit-an-open-letter-from-altimeter-to-mark-zuckerberg-and-the-meta-board-of-392d94e80a18.

[ge] https://reason.com/2022/11/01/twitter-was-toxic-long-before-musk-took-over/.

[gf] https://www.youtube.com/watch?v=r-mAyu5RruU.

[gg] https://www.forbes.com/sites/kateoflahertyuk/2021/01/10/facebook-users-have-3-superb-reasons-to-quit-in-2021/.

[gh] https://www.forbes.com/sites/niallmccarthy/2019/03/08/is-facebook-becoming-social-medias-retirement-home-infographic/.

[gi] https://www.nasdaq.com/articles/every-time-mark-zuckerberg-mentioned-tiktok-on-metas-earnings-call.

[gj] https://www.youtube.com/watch?v=zuqBRN5SJZ8.

[gk] https://www.youtube.com/watch?v=_0h7R8q0Ibk.

[gl] https://youtu.be/70z6FQ5Cn80?t=445.

[gm] https://www.youtube.com/watch?v=U0Gw9IUmjwM.

[gn] https://www.rottentomatoes.com/m/how_the_west_was_won.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

Bitterman M.E. (2006). Classical conditioning since pavlov. *Review of General Psychology* **10**(4), 365–376.

Blodgett S.L., **Barocas S.**, **Daumé III H.** and **Wallach H.** (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 5454–5476.

Bolukbasi T., **Chang K.-W.**, **Zou J.Y.**, **Saligrama V.** and **Kalai A.T.** (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*.

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

Brown T.B., **Mann B.**, **Ryder N.**, **Subbiah M.**, **Kaplan J.**, **Dhariwal P.**, **Neelakantan A.**, **Shyam P.**, **Sastry G.**, **Askell A.**, **Agarwal S.**, **Herbert-Voss A.**, **Krueger G.**, **Henighan T.**, **Child R.**, **Ramesh A.**, **Ziegler D.M.**, **Wu J.**, **Winter C.**, **Hesse C.**, **Chen M.**, **Sigler E.**, **Litwin M.**, **Gray S.**, **Chess B.**, **Clark J.**, **Berner C.**, **McCandlish S.**, **Radford A.**, **Sutskever I.** and **Amodei D.** (2020). Language models are few-shot learners. *NeurIPS*.

Buolamwini J. and **Gebru T.** (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In **Friedler S.A.** and **Wilson C.** (eds), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Proceedings of Machine Learning Research, vol. 81, PMLR, pp. 77–91.

Caliskan A., **Bryson J.J.** and **Narayanan A.** (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186.

Chen J., **Dong H.**, **Wang X.**, **Feng F.**, **Wang M.** and **He X.** (2020). Bias and debias in recommender system: A survey and future directions. ArXiv, abs/2010.03240.

Chen M., **Tworek J.**, **Jun H.**, **Yuan Q.**, **Ponde H.**, **Kaplan J.**, **Edwards H.**, **Burda Y.**, **Joseph N.**, **Brockman G.**, **Ray A.**, **Puri R.**, **Krueger G.**, **Petrov M.**, **Khlaaf H.**, **Sastry G.**, **Mishkin P.**, **Chan B.**, **Gray S.**, **Ryder N.**, **Pavlov M.**, **Power A.**, **Kaiser L.**, **Bavarian M.**, **Winter C.**, **Tillet P.**, **Such F.P.**, **Cummings D.W.**, **Plappert M.**, **Chantzis F.**, **Barnes E.**, **Herbert-Voss A.**, **Guss W.H.**, **Nichol A.**, **Babuschkin I.**, **Balaji S.A.**, **Jain S.**, **Carr A.**, **Leike J.**, **Achiam J.**, **Misra V.**, **Morikawa E.**, **Radford A.**, **Knight M.M.**, **Brundage M.**, **Murati M.**, **Mayer K.**, **Welinder P.**, **McGrew B.**, **Amodei D.**, **McCandlish S.**, **Sutskever I.** and **Zaremba W.** (2021). Evaluating large language models trained on code. ArXiv, abs/2107.03374.

Choudhury M.D., **Kiciman E.**, **Dredze M.**, **Coppersmith G.A.** and **Kumar M.** (2016). Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

Chowdhery A., **Narang S.**, **Devlin J.**, **Bosma M.**, **Mishra G.**, **Roberts A.**, **Barham P.**, **Chung H.W.**, **Sutton C.**, **Gehrmann S.**, **Schuh P.**, **Shi K.**, **Tsvyashchenko S.**, **Maynez J.**, **Rao A.**, **Barnes P.**, **Tay Y.**, **Shazeer N.**, **Prabhakaran V.**, **Reif E.**, **Du N.**, **Hutchinson B.**, **Pope R.**, **Bradbury J.**, **Austin J.**, **Isard M.**, **Gur-Ari G.**, **Yin P.**, **Duke T.**, **Levskaya A.**, **Ghemawat S.**, **Dev S.**, **Michalewski H.**, **Garcia X.**, **Misra V.**, **Robinson K.**, **Fedus L.**, **Zhou D.**, **Ippolito D.**, **Luan D.**, **Lim H.**, **Zoph B.**, **Spiridonov A.**, **Sepassi R.**, **Dohan D.**, **Agrawal S.**, **Omernick M.**, **Dai A.M.**, **Pillai T.S.**, **Pellat M.**, **Lewkowycz A.**, **Moreira E.**, **Child R.**, **Polozov O.**, **Lee K.**, **Zhou Z.**, **Wang X.**, **Saeta B.**, **Diaz M.**, **Firat O.**, **Catasta M.**, **Wei J.**, **Meier-Hellstern K.**, **Eck D.**, **Dean J.**, **Petrov S.** and **Fiedel N.** (2022). Palm: Scaling language modeling with pathways.

Christiansen M.H. and **Chater N.** (2017). Towards an integrated science of language. *Nature Human Behaviour* **1**(8), 1–3.

Church K.W. and **Kordoni V.** (2021). Emerging trends: Ethics, intimidation, and the cold war. *Natural Language Engineering* **27**, 379–390.

Clinton H.R. (2017). *What Happened*. New York, NY, USA: Simon and Schuster.

Colleoni E., **Rozza A.** and **Arvidsson A.** (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication* **64**, 317–332.

Comey J. (2018). *A Higher Loyalty: Truth, Lies, and Leadership*. New York, NY, USA: Flatiron Books.

Cortes C., **Pregibon D.** and **Volinsky C.** (2001). Communities of interest. In *International Symposium on Intelligent Data Analysis*. Berlin, Heidelberg: Springer, pp. 105–114.

Courtwright D.T. (2019). *The Age of Addiction: How Bad Habits Became Big Business*. Cambridge, Massachusetts, USA: Harvard University Press.

Covington P., **Adams J.** and **Sargin E.** (2016). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 191–198.

Crenshaw K.W. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, **1989**(1).

Davidson J., **Liebald B.**, **Liu J.**, **Nandy P.**, **Van Vleet T.**, **Gargi U.**, **Gupta S.**, **He Y.**, **Lambert M.**, **Livingston B.** and **Sampath D**. (2010). The youtube video recommendation system. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, pp. 293–296.

Davidson T., **Warmsley D.**, **Macy M.W.** and **Weber I.** (2017). Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Davis A.Y. (1993). Black feminist thought: Knowledge, consciousness and the politics of empowerment. *Teaching Philosophy* **16**, 351–353.

De Koster W. and **Houtman D.** (2008). 'stormfront is like a second home to me' on virtual community formation by right-wing extremists. *Information, Communication & Society* **11**(8), 1155–1176.

Dechêne A., **Stahl C.**, **Hansen J.** and **Wänke M.** (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review* **14**(2), 238–257.

Delobelle P., **Winters T.** and **Berendt B.** (2020). RobBERT: a Dutch RoBERTa-based Language Model, Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3255–3265.

Dinan E., **Fan A.**, **Williams A.**, **Urbanek J.**, **Kiela D.** and **Weston J.** (2020). Queens are powerful too: Mitigating gender bias in dialogue generation. In *EMNLP*, pp. 8173–8188.

DiSalvo C., **Clement A.** and **Pipek V.** (2012). Communities: Participatory design for, with and by communities. In *Routledge International Handbook of Participatory Design*. Routledge, pp. 202–230.

Fazio L.K., **Brashier N.M.**, **Payne B.K.** and **Marsh E.J.** (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General* **144**(5), 993.

Ferrara E. (2017). Disinformation and social bot operations in the run up to the 2017 french presidential election. *Interorganizational Networks & Organizational Behavior eJournal* **22**, 993–1002.

Fischer C.S. (1982). *To Dwell Among Friends: Personal Networks in Town and City*. London, UK: University of Chicago Press.

Fisher M. (2022). *THE CHAOS MACHINE: The Inside Story of How Social Media Rewired Our Minds and Our World*. Hachette Book Group, New York, NY, USA: Little, Brown & Company.

Fogg B.J. (2002). Persuasive technology: Using computers to change what we think and do. *Ubiquity* **2002**, 5.

Fordyce S. (2019). Value sensitive design: Shaping technology with moral imagination. *Design and Culture* **12**, 109–111.

Fourney A., **Rácz M.Z.**, **Ranade G.**, **Mobius M.M.** and **Horvitz E.** (2017). Geographic and temporal trends in fake news consumption during the 2016 US presidential election. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Fuchs C. (2012). Social media, riots, and revolutions. *Capital & Class* **36**(3), 383–391.

Gehman S., **Gururangan S.**, **Sap M.**, **Choi Y.** and **Smith N.A.** (2020). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. ArXiv, abs/2009.11462.

Gillespie T. (2014). The relevance of algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, vol. **167**.

Goralski M.A. and **Górniak-Kocikowska K.** (2022). Artificial intelligence in the United States. *International Perspectives on Artificial Intelligence* **5**.

Greenwald A.G., **McGhee D.E.** and **Schwartz J.L.K.** (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* **74**(6), 1464–1480.

Griffiths M.D. (2000). Does internet and computer "addiction" exist? some case study evidence. *Cyberpsychology, Behavior, and Social Networking* **3**, 211–218.

Grinberg N., **Joseph K.**, **Friedland L.**, **Swire-Thompson B.** and **Lazer D.** (2019). Fake news on twitter during the 2016 US presidential election. *Science* **363**, 374–378.

Hakim N. (2020). How social media companies could be complicit in incitement to genocide. *Chicago Journal of International Law* **21**, 83.

Hamm M.P., **Newton A.S.**, **Chisholm A.**, **Shulhan J.**, **Milne A.**, **Sundar P.**, **Ennis H.**, **Scott S.D.** and **Hartling L.** (2015). Prevalence and effect of cyberbullying on children and young people: A scoping review of social media studies. *JAMA Pediatrics* **169**(8), 770–777.

Heilman M.E., **Wallen A.S.**, **Fuchs D.** and **Tamkins M.M.** (2004). Penalties for success: Reactions to women who succeed at male gender-typed tasks. *The Journal of Applied Psychology* **89**(3), 416–427.

Himelboim I., **McCreery S.** and **Smith M.A.** (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *Journal of Computer-Mediated Communication* **18**, 40–60.

Howard P.N., **Duffy A.**, **Freelon D.**, **Hussain M.M.**, **Mari W.** and **Maziad M.** (2011). Opening closed regimes: What was the role of social media during the arab spring? Available at SSRN 2595096.

Ihle L.H. (2019). *The Rise of Viral Epistemology Dissecting a New Knowledge Ideal*. Unpublished Thesis, Sant Anna School of Advanced Studies.

Jacobsen W.C. and **Forste R.** (2011). The wired generation: Academic and social outcomes of electronic media use among university students. *Cyberpsychology, Behavior and Social Networking* **14**(5), 275–280.

Jin F., **Wang W.**, **Zhao L.**, **Dougherty E.R.**, **Cao Y.**, **Lu C.-T.** and **Ramakrishnan N.** (2014). Misinformation propagation in the age of twitter. *Computer* **47**, 90–94.

Johnson N.F., **Velásquez N.**, **Restrepo N.J.**, **Leahy R.**, **Gabriel N.**, **El Oud S.**, **Zheng M.**, **Manrique P.**, **Wuchty S.** and **Lupu Y.** (2020). The online competition between pro-and anti-vaccination views. *Nature* **582**(7811), 230–233.

Junco R. (2012). Too much face and not enough books: The relationship between multiple indices of facebook use and academic performance. *Computers in Human Behavior* **28**(1), 187–198.

Junco R. and **Cotten S.R.** (2012). No A 4 U: The relationship between multitasking and academic performance. *Computers and Education* **59**, 505–514.

Kaiser J. and **Rauchfleisch A.** (2020). Birds of a feather get recommended together: Algorithmic homophily in Youtube's channel recommendations in the United States and Germany. *Social Media + Society* **6**, 1–15.

Kurka D.B., **Godoy A.** and **Zuben F.J.V.** (2016). Birds of a feather tweet together: Computational techniques to understand user communities in social networks. In *#Microposts*.

Kuss D.J. and **Griffiths M.D.** (2011a). Internet gaming addiction: A systematic review of empirical research. *International Journal of Mental Health and Addiction* **10**, 278–296.

Kuss D.J. and **Griffiths M.D.** (2011b). Online social networking and addiction—a review of the psychological literature. *International Journal of Environmental Research and Public Health* **8**, 3528–3552.

Lazer D., **Baum M.A.**, **Benkler Y.**, **Berinsky A.J.**, **Greenhill K.M.**, **Menczer F.**, **Metzger M.J.**, **Nyhan B.**, **Pennycook G.**, **Rothschild D.M.**, **Schudson M.**, **Sloman S.A.**, **Sunstein C.R.**, **Thorson E.A.**, **Watts D.J.** and **Zittrain J.** (2018). The science of fake news. *Science* **359**, 1094–1096.

Luxton D.D., **June J.D.** and **Fairall J.** (2012). Social media and suicide: A public health perspective. *American Journal of Public Health* **102**(Suppl 2), S195–200.

Martinez A.G. (2018). *Chaos Monkeys: Obscene Fortune and Random Failure in Silicon Valley*. New York, NY, USA: HarperCollins Publishers.

McNamee R. (2020). *Zucked: Waking up to the Facebook Catastrophe*. New York, NY, USA: Penguin Random House LLC.

McPherson M., **Smith-Lovin L.** and **Cook J.M.** (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology* **27**, 415–444.

Mitchell M., **Wu S.**, **Zaldivar A.**, **Barnes P.**, **Vasserman L.**, **Hutchinson B.**, **Spitzer E.**, **Raji I.D.** and **Gebru T.** (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229.

Müller K. and **Schwarz C.** (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association* **19**(4), 2131–2167.

Munoz, **J.M.** and **Maurya**, **A.** (2022). *International Perspectives on Artificial Intelligence*. London, UK: Anthem Press.

Nangia N., **Vania C.**, **Bhalerao R.** and **Bowman S.R.** (2020). Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *EMNLP*.

Ng L.H.X. and **Taeihagh**, **A.** (2021). How does fake news spread? understanding pathways of disinformation spread through apis. *Policy & Internet* **13**(4), 560–585.

O'Callaghan D., **Greene D.**, **Conway M.**, **Carthy J.** and **Cunningham P.** (2015). Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review* **33**(4), 459–478.

O'Dea B., **Wan S.**, **Batterham P.J.**, **Calear A.L.**, **Paris C.** and **Christensen H.** (2015). Detecting suicidality on Twitter. *Internet Interventions* **2**, 183–188.

O'Neil C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Random House LLC, New York, NY, USA.

Oates S.L. (2020). The easy weaponization of social media: Why profit has trumped security for US companies. *Digital War* **1**, 117–122.

Ouyang L., **Wu J.**, **Jiang X.**, **Almeida D.**, **Wainwright C.L.**, **Mishkin P.**, **Zhang C.**, **Agarwal S.**, **Slama K.**, **Ray A.**, **Schulman J.**, **Hilton J.**, **Kelton F.**, **Miller L.E.**, **Simens M.**, **Askell A.**, **Welinder P.**, **Christiano P.F.**, **Leike J.** and **Lowe R.J.** (2022). Training language models to follow instructions with human feedback. ArXiv, abs/2203.02155.

Paluck E.L., **Shepherd H.** and **Aronow P.M.** (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences* 113, 566–571.

Platt S.R. (2018). *Imperial Twilight: The Opium War and the End of China's Last Golden Age*. New York, NY, USA: Penguin Random House.

Pontes H.M. and **Griffiths M.D.** (2015). Measuring dsm-5 internet gaming disorder: Development and validation of a short psychometric scale. *Computers in Human Behavior* **45**, 137–143.

Primack B.A., **Shensa A.**, **Escobar-Viera C.G.**, **Barrett E.L.**, **Sidani J.E.**, **Colditz J.B.** and **James A.E.** (2017). Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among u.s. young adults. *Computers in Human Behavior* **69**, 1–9.

Rauchfleisch A. and **Kaiser J.** (2020). The german far-right on youtube: An analysis of user overlap and user comments. *Journal of Broadcasting & Electronic Media* **64**(3), 373–396.

Rescorla R. (1988). Pavlovian conditioning. It's not what you think it is. *The American Psychologist* **43**(3), 151–60.

Robinson J., **Cox G.**, **Bailey E.**, **Hetrick S.E.**, **Rodrigues M.**, **Fisher S.** and **Herrman H.** (2016). Social media and suicide prevention: A systematic review. *Early Intervention in Psychiatry* **10**, 103–121.

Rogers A., **Baldwin T.** and **Leins K.** (2021). 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic. Association for Computational Linguistics, pp. 4821–4833.

Schackmuth A. (2018). *Extremism, Fake News and Hate: Effects of Social Media in the Post-Truth Era*. Unpublished Thesis.

Schmidt A. and **Wiegand M.** (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Valencia, Spain. Association for Computational Linguistics, pp. 1–10.

https://doi.org/10.1017/S1351324922000481 Published online by Cambridge University Press

Senthil Kumar B., **Chandrabose A.** and **Chakravarthi B.R.** (2021). An overview of fairness in data–illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Kyiv: Association for Computational Linguistics, pp. 34–45.

Shprintzen R.J. (1990). What's in a name? *The Cleft Palate Journal* **27**(4), 335–336.

Simonsen J. and **Robertson T.** (2013). Routledge International Handbook of Participatory Design, vol. 711. Routledge New York.

Skinner B.F. (1953). Some contributions of an experimental analysis of behavior to psychology as a whole. *American Psychologist* **8**(2), 69.

Skinner B.F. (1965). *Science and Human Behavior*, vol. **92904**, A Free Press paperback. Psychology on Simon and Schuster.

Skinner B.F. (1986). Is it behaviorism? *Behavioral and Brain Sciences* **9**, 716.

Smidi A. and **Shahin S.** (2017). Social media and social mobilisation in the middle east: A survey of research on the arab spring. *India Quarterly* **73**(2), 196–209.

Suarez-Lledo V. and **Alvarez-Galvez J.** (2021). Prevalence of health misinformation on social media: Systematic review. *Journal of Medical Internet Research* **23**(1), e17187.

Taplin J. (2017). *Move Fast and Break Things: How Facebook, Google, and Amazon Cornered Culture and Undermined Democracy*. New York, NY, USA: Little, Brown and Company, Hachette Book Group.

Unkelbach C., **Koch A.**, **Silva R.R.** and **Garcia-Marques T.** (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science* **28**(3), 247–253.

van den Eijnden R.J.J.M., **Lemmens J.S.** and **Valkenburg P.M.** (2016). The social media disorder scale. *Computers in Human Behavior* **61**, 478–487.

Vosoughi S., **Roy D.K.** and **Aral S.** (2018). The spread of true and false news online. *Science* **359**, 1146–1151.

Waseem Z., **Davidson T.**, **Warmsley D.** and **Weber I.** (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, Vancouver, BC, Canada. Association for Computational Linguistics, pp. 78–84.

Wu T. (2003). Network neutrality, broadband discrimination. *Journal on Telecommunications and High Technology Law* **2**, 141.

Wylie C. (2019). *Mindf∗ck: Cambridge Analytica and the Plot to Break America*. New York, NY, USA: Random House.

Young K.S. (1998). Internet addiction: The emergence of a new clinical disorder. *Cyberpsychology, Behavior, and Social Networking* **1**, 237–244.

Zampieri M., **Nakov P.**, **Rosenthal S.**, **Atanasova P.**, **Karadzhov G.**, **Mubarak H.**, **Derczynski L.**, **Pitenis Z.** and **Çöltekin Ç.** (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona (online). International Committee for Computational Linguistics, pp. 1425–1447.

Zeitzoff T. (2017). How social media is changing conflict. *Journal of Conflict Resolution* **61**, 1970–1991.

Zhuravskaya E., **Petrova M.** and **Enikolopov R.** (2020). Political effects of the internet and social media. *Annual Review of Economics* **12**, 415–438.

Zook M., **Barocas S.**, **Boyd D.**, **Crawford K.**, **Keller E.**, **Gangadharan S.P.**, **Goodman A.**, **Hollander R.**, **Koenig B.A.**, **Metcalf J.**, **Narayanan A.**, **Nelson A.** and **Pasquale F**. (2017). Ten simple rules for responsible big data research. *PLoS Computational Biology* **13**(3), e1005399.

Zych I., **Ortega-Ruiz R.** and **Rey R.D.** (2015). Systematic review of theoretical studies on bullying and cyberbullying: Facts, knowledge, prevention, and intervention. *Aggression and Violent Behavior* **23**, 1–21.

Cite this article: Church K, Schoene A, Ortega JE, Chandrasekar R and Kordoni V (2023). Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable. *Natural Language Engineering* **29**, 483–508. https://doi.org/10.1017/S1351324922000481