How Algorithms Create and Prevent Fake News

Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More

Noah Giansiracusa

Apress®

HOW ALGORITHMS CREATE AND PREVENT FAKE NEWS

EXPLORING THE IMPACTS OF SOCIAL MEDIA, DEEPFAKES, GPT-3, AND MORE

Noah Giansiracusa

Apress®

How Algorithms Create and Prevent Fake News: Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More

Noah Giansiracusa
Acton, MA, USA

ISBN-13 (pbk): 978-1-4842-7154-4 ISBN-13 (electronic): 978-1-4842-7155-1
https://doi.org/10.1007/978-1-4842-7155-1

Copyright © 2021 by Noah Giansiracusa

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Shiva Ramachandran
Development Editor: James Markham
Coordinating Editors: Nancy Chen and Mark Powers

Cover designed by eStudioCalamar

Distributed to the book trade worldwide by Springer Science+Business Media New York, 1 New York Plaza, New York, NY 100043. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail booktranslations@springernature.com; for reprint, paperback, or audio rights, please e-mail bookpermissions@springernature.com.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/9781484271544. For more detailed information, please visit http://www.apress.com/source-code.

Printed on acid-free paper

Dedicated to my wife Emily and our parents:
Bob, Dorothy, Andy, and Carole.

Contents

About the Author

Noah Giansiracusa received a PhD in mathematics from Brown University and is an Assistant Professor of Mathematics and Data Science at Bentley University, a business school near Boston. He previously taught at UC Berkeley, University of Georgia, and Swarthmore College. He has received national grants and spoken at international conferences for his research in mathematics, and he has been quoted several times in *Forbes* as an expert on artificial intelligence. He has dozens of publications in math and data science and has taught courses ranging from a first-year seminar on quantitative literacy to graduate machine learning. Most recently, he created an interdisciplinary seminar on truth and lies in data and algorithms that was part of the impetus for this book.

Acknowledgments

Thanks to Jordan Ellenberg, Cathy O'Neil, and Francis Su for inspiring me to write a book and helping me learn about the publishing industry. Thanks to Karen Hao, Will Douglas Heaven, Davey Alba, Jack Nicas, Kevin Roose, James Vincent, Issie Lapowsky, Carole Cadwalladr, Julia Angwin, Deepa Seetharaman, Jeff Horwitz, Sheera Frenkel, and so many other technology journalists for doing the hard work that this book relies so heavily upon. Thanks to Gerald Seidler, Jim Morrow, Henry Cohn, Dan Abramovich, Angela Gibney, Bernd Sturmfels, and my other teachers and mentors who helped me become a math professor. Thanks to Charlie Hadlock, Rick Oches, Lucy Kimball, and others at Bentley University for believing in me as a data scientist and providing me with the opportunities that led to this book. Thanks to Steffen Marcus for encouraging me to think and write about the broader context of math and technology. Thanks to Tom Taulli for putting me in contact with Apress, and to my team at Apress—Shiva Ramachandran, Matthew Moodie, Nancy Chen, Rita Fernando, and Mark Powers—for all their work shaping this book and helping get it across the finish line. Thanks to my parents, Bob and Dorothy, for homeschooling me and passing along a passion for lifelong learning. Thanks to my brother, Jeffrey, for paving the way forward for us both. And thanks to my wife, Emily, and our daughter, Claire, for more than words can convey.

Introduction

You might have heard rumors that the newsfeed algorithm at Facebook and the video recommendation algorithm at YouTube are spreading fake news, or that artificial intelligence (AI) can now rapidly generate convincing articles and make videos of people doing and saying things they never did, or that machine learning algorithms will save us from fake news by automatically detecting it and labeling assertions as true or false. But what do these claims even mean, and what should you believe? The main goal of this book is to help readers of all backgrounds—no knowledge of math, statistics, computers, algorithms, or journalism required—understand what's really going on by collecting all the investigations, research, and stories about fake news and algorithms in one place and explaining it in a simple way while weaving it together into a coherent narrative. Another goal is to teach you about the publicly available tools that can help you do your own part in the fight against fake news.

"If we are not serious about facts and what's true and what's not, if we can't discriminate between serious arguments and propaganda, then we have problems." Barack Obama said this on November 17, 2016, just nine days after Donald Trump was elected to be his successor in the White House. Since then, there has been an increasing awareness of the scope and impact of "fake news," a catchall label for *misinformation* (false information that is spread regardless of intent to mislead) and *disinformation* (deliberately false or misleading information). There has also been an increasing awareness of the role played by data-driven algorithms in the creation, dissemination, and detection/moderation of fake news. But the story of fake news and algorithms has been difficult for most of us to follow. It has unfolded in a wide range of academic publications, journalistic investigations, corporate announcements, and governmental hearings, and it involves many sophisticated technological concepts that sound mysterious. I strongly believe that the barriers to entering this important discussion are not nearly as high as they might seem, and this book is my attempt to lower them even further.

Chapter 1 sets the stage by exploring the economics of blogging and online newspapers, with an emphasis on the dynamics that have led to a proliferation of low-quality journalism. Data, in the form of clicks and pageviews, has transformed the news industry, and you'll see how fake news peddlers have taken advantage of this. Chapter 2 looks at a new development in our ongoing battle to understand what's real and what's not: fake journalists with untraceable lifelike profile photos synthesized by AI, and entire articles written by AI

with the click of a button. You'll learn about the technology behind these advances (GPT-3 and deepfake GANs, with a gentle overview of machine learning along the way) and the impact they're having on journalism. Chapter 3 continues this line of investigation by turning to deepfake video editing—explaining how it works, what it can do, and the role it has played in politics. Chapter 4 is all about YouTube and its recommendation algorithm that automatically selects videos for you to watch. A history of this algorithm is provided, including brief discursions into deep learning and reinforcement learning, and empirical investigations into the way it works in practice are explored. This frames a discussion of fake news and conspiratorial content on YouTube, especially in the context of Brazil's 2018 election and the 2016 and 2020 US elections.

After several chapters on how AI can create and spread fake news, Chapter 5 asks if AI can help fight it by determining whether someone in a video is lying. This is part of an algorithmic reinvention of the polygraph that is currently being trialed at airports and elsewhere. Chapter 6 takes a deep look at one of the world's most popular sources of information: Google. The company's efforts to elevate quality content over fake news and harmful material are detailed, as are the various failures that have occurred along the way and the challenges that remain. Chapter 7 shows how Google supports the fake news industry financially through ad revenue and how Facebook's algorithmically distributed ads have been a persistent source of fake news and racism. Chapter 8 takes a thorough look at how fake news spreads across social media and how algorithms have been used to detect and mitigate this spread. Finally, Chapter 9 collects and explains some publicly available AI-powered fact-checking tools that you can use to make sure what you're reading is trustworthy and truthful.

CHAPTER

1

Perils of Pageview

The Data-Driven Economics of Online Journalism

The economics of the Internet created a twisted set of incentives that make traffic more important—and more profitable—than the truth.

—Ryan Holiday, *Trust Me, I'm Lying: Confessions of a Media Manipulator*

Much of what we know, or think we know, about what is happening in the world we learn by reading the news. But nowadays "the news" means something different than it did in generations past. What we read primarily today are articles on the internet—everything ranging from casual blog posts to meticulously researched stories on national and international news sites. The transition of journalism from print to screen does not inherently mean what we read is less truthful than it used to be. However, this technological transformation has enabled a less overt but nonetheless extraordinarily influential economic transformation: the *datafication* of the journalism industry. The pageviews and clicks we all sprinkle across the internet are, as I will discuss, the digital fertilizer feeding a burgeoning garden of misinformation

© Noah Giansiracusa 2021
N. Giansiracusa, *How Algorithms Create and Prevent Fake News*,
https://doi.org/10.1007/978-1-4842-7155-1_1

and fake news. By tracing the financial incentives involved in the contemporary news cycle, I hope in this chapter to convey the alarming extent that data, unseen to most of us yet created by our actions and activities, is fundamentally shaping what we read every day and threatening the bulwark of traditional journalistic standards.

Propagation of Stories

Let me start with a taxonomy of sorts. At the bottom of the internet media food chain, if you will, are small blogs and websites that cover very focused issues, interests, or regions; these can be single author or multi-author. The next tier up comprises the blogs of newspapers, magazines, and television stations. This is a confusing middle ground because many of these blogs share the name, URL, and logo of a recognizable news source yet the editorial standards are generally lower than those of the parent organization, and many of the contributors lack the journalistic training one might expect from the parent organization. Then at the top are the official news sites, which can be regional but tend to draw a large national or international readership. This hierarchy is not about quality—indeed, some very focused small blogs produce content of extremely high quality, while some big-name national news sites consistently publish articles of seriously questionable accuracy. The levels here are more about the size of both the audience and the organization and about the scope of the content.

Information flows both vertically and horizontally through this internet news hierarchy. When the *Washington Post* breaks a big story, it is only a matter of hours before the *New York Times* covers it as well, and vice versa, often simply by reporting what was reported in the other newspaper's article. This is *horizontal propagation*, and it happens because even though the second newspaper cannot claim credit for breaking the story, it does not want its readership to obtain this information directly from the competitor newspaper. *Vertical propagation* happens in two directions. A big story broken at the top will be covered and duplicated by smaller news organizations and blogs because, similar to horizontal propagation, this is an easy way of keeping readers without doing much work; this is a *downward* flow of information.

While there is an obvious redundancy, hence an overall systemic inefficiency, to both horizontal propagation and downward vertical propagation, the only real harm to the truth-seeking reader is that important details might be omitted and facts distorted as the story is passed from organization to organization—though sometimes a more specialized blog will provide a valuable service by delving deeper into a particular facet of the story than would be appropriate for the top-level organization. It can be quite illuminating to find a story that was broken by one newspaper and then compare its coverage across a range of other newspapers and blogs; this is an excellent

way to uncover the ideological inclinations of different organizations, since the same set of facts will be colored by the different viewpoints involved.

The remaining form of journalistic propagation is the *upward* vertical flow, where stories start at small blogs and sometimes end all the way up at national news sites. This is one of the key topics of this chapter because it is responsible for a staggering amount of the misreporting and outright fake news that we see, and it is driven almost entirely by data and the economics of modern media. Before exploring this specific topic, it helps to take a step back and look at the financial forces driving blogs and newspapers; throughout, I take a broad view of "blogging" to include essentially all forms of posting written content online.

Economics of Blogging

Ostensibly, the revenue for blogs comes from selling advertisements. There are a variety of pecuniary mechanisms for online advertisements, such as the advertising company affixing a banner atop the blog and paying based on *pageviews* (the number of users who visit the blog where the banner is displayed), and in some cases the advertiser pays an additional sum when a reader on the blog clicks the ad link and proceeds to actually purchase a product from the advertising company. But the most common format is pay-per-impression and pay-per-click advertising, in which the blog places an ad somewhere on its website and is paid based on *impressions* (the number of times the ad is seen by a reader on the blog) or *clicks* (the number of times the ad is clicked by a reader on the blog). The bottom line is that to maximize ad revenue, the blog needs to maximize traffic.

But why did I write "ostensibly" in the preceding paragraph? Well, there is somewhat of a Ponzi scheme dynamic at play here. Advertising revenue tends to be relatively low even for popular blogs, so the real ambition of most blogs, even if they don't admit it, is to gain sufficient popularity and traffic that a larger organization will buy them out and incorporate the blog into its larger website in order to increase traffic—often so that the larger website can boost its odds of being bought by a yet larger organization.

For example, Nate Silver's technical yet surprisingly popular blog on political polls was launched in 2008, brought into the *New York Times* in 2010, acquired by ESPN in 2013, then transferred to the sister property ABC News in 2018. Arianna Huffington's groundbreaking general news blog the *Huffington Post* was founded in 2005 with a one million dollar investment and sold to AOL in 2011 for three hundred and fifty million—but, quite tellingly, at the time of this sale, its ad revenue was only thirty-one million dollars per year. This tenfold purchase price to annual revenue ratio is rather extreme and suggests that AOL was banking on continued long-term growth as well as other factors like the prestige of adding such a popular online newspaper and bringing

onboard the superstar Arianna herself. At the end of the day, whether a blog aims to be bought out or not, the path to success is web traffic.

Next, let me turn from the economics of blogs to that of the bloggers themselves. In the early days of blogging, bloggers tended to be paid either a flat rate with a required minimum number of daily posts, or they were paid per post; in the mid-2000s, depending on the establishment, this rate was often a dismal five or ten dollars per post. A paradigm shift occurred when *Gawker* left this per-post payment system and instead paid each blogger a monthly salary that was augmented by a bonus based on the number of pageviews recorded by the blogger's articles. This shift made sense from an ad revenue perspective, and it quickly rippled across the blogosphere and ushered in a new era in which pageviews became the fundamental currency of blogging.

Gawker took things even further when it installed a large board in its office showing a live tally of the pageview statistics for all its bloggers and their posts (other blogs soon turned to similar methods as well). This led to an intense pageview competition among the bloggers at the company, designed to stimulate productivity, and it signaled a strong emphasis on analytics in which bloggers could not help but keep score of which articles generated the most pageviews.

This blogger remuneration system is blatantly reductionist: the reader's opinion of a blog post is irrelevant. In fact, it does not even matter whether the reader actually reads the post—once the link to a post is clicked, the pageview is recorded, and that's all that counts. An unfortunate but largely predictable consequence has been the proliferation of clickbait: catchy, often trashy, headlines that encourage clicks rather than bespeaking quality journalism.[1] A lengthy, methodically researched and fact-checked article provides no more financial value than a piece of vapid tabloid trash. This oversimplifies the situation as many readers follow certain blogs precisely because they consistently post high-quality articles, but many readers also click whatever stories are catchiest when scrolling through social media or news aggregators, and in these latter settings the name and reputation of the blog/organization is often a secondary factor in the decision to click—it is the headline that matters most.

An additional, and significant, dynamic is that blog posts tend to have short-lived pageview-generating lifespans. Consequently, bloggers and blogs, in their constant journey for increased traffic, are under intense pressure to produce as many posts as possible, as rapidly as possible. A traditional print newspaper had to produce content that filled one print edition per day; a cable news network has to produce content that fills twenty-four hours a day, three

[1] And insidious techniques for gaming the system have inevitably, and unsurprisingly, flourished, such as posting slide shows in which the reader needs to click each slide one at a time, thereby artificially inflating the pageview metric.

hundred and sixty-five days a year; a blog has limitless space and is rewarded for attempting to fill this infinitude. This encourages rushed, sloppy writing and journalistic shortcuts; bloggers simply don't have time to fact-check. In fact, posts that generate controversy tend to also generate pageviews. Even worse, outright fallacies in news articles often entice disgruntled readers to leave comments complaining and/or correcting the article, but commenting on blogs usually involves multiple clicks and data trails that are dollars (well, pennies) in the pockets of the blogger.

Putting these observations all together, we see the perfect storm of conditions assaulting the foundations of journalism. Blogs and bloggers are almost all financially strapped, earning far less revenue than an outsider might expect, and so are in desperate need of more pageviews—whether to earn ad revenue directly or to raise the prospect of a lucrative buyout. This drives them to produce articles far too quickly, leaving precious little time to fact-check and verify sources. Even if they had time to fact-check, the pageview statistics they obsess over show that there is no real financial incentive for being truthful, as misleading articles with salacious headlines often encourage more clicks than do works of authentic journalism.

And let me be abundantly clear about this: it is the data-driven impetus of the blogging industry, and the vast oversimplification and distortion of multidimensional journalistic value caused by reducing everything to a single, simple-minded, superficial metric—the pageview—that is most responsible for this dangerous state of affairs. That some pageview-driven blogs thrive on thoughtful, methodical, accurate writing is truly remarkable in this market that is saturated with perverse incentives pressuring writers to engage in the exact opposite of these noble qualities. Let us all be thankful for the good blogs and good writing when we see it, for it is certainly out there but it struggles to rise above the ubiquitous clickbait filth pervading the internet.

Having presented the data-driven financial structure of blogs and bloggers, and the pernicious pressures it leads to, it is time now to turn back to the earlier discussion of the taxonomy of the blogosphere and the propagation of stories through it.

Up from the Bottom

Renée DiResta, a researcher at the Stanford Internet Observatory, recently wrote[2] in the *Atlantic* that "Media and social media are no longer distinct; consequential narratives emerge from the bottom up, as well as the top down, and bounce back and forth among different channels." Recall that the

[2]Renée DiResta, "The Right's Disinformation Machine Is Getting Ready for Trump to Lose," *Atlantic*, October 20, 2020: https://www.theatlantic.com/ideas/archive/2020/10/the-rights-disinformation-machine-is-hedging-its-bets/616761/.

propagation direction I haven't yet directly addressed, despite claiming it is the one most responsible for our current morass of media mendacity, is the upward flow where stories start in small, typically special interest and/or geographically local blogs, and manage to work their way up the food chain, sometimes ending all the way at the top on national news sites. The questions we must ask here are: how and why does this happen, and why does this lead to less truthful news? The answers, as I next discuss, all essentially follow from the pageview economics of blogging.

All blogs and sources of news, even the highly regarded ones at the top, are in constant search for new stories. There is a fundamental inequality at play that the supply of actual stories (meaning real events transpiring in the world that ought to be reported) is substantially smaller than the supply of stories produced by blogs and online newspaper—because, as I discussed above, the pressure to accumulate pageviews compels writers to fill the limitless bandwidth of the internet at an unhealthy rate. This creates a dangerous vacuum in which bloggers at all levels are under immense pressure to constantly find stories wherever they can, and oftentimes to create something out of nothing, to keep the wheels of the modern media machine turning.

Blogs at the lowest levels of the hierarchy are typically underfunded and understaffed and tend to rely upon the small, close-knit nature of the community they are part of—meaning they often publish material based on suggestions from members of the community and follow leads on social media without really questioning their veracity. In many ways, this is quite reasonable: a respected national news station upon hearing some scandalous gossip regarding the Biden administration needs to be damn sure it is accurate before reporting it to the public, whereas a blog about Great Pyrenees dogs and their crazy antics is less concerned with the possibility that its posts might be construed as fake news. Generally speaking, smaller and more specialized blogs have fewer resources to investigate leads and less incentive to do so regardless.

The problem starts to arise, however, when we look at the middle rung in the hierarchy. Here, the bloggers are still desperate for stories, and they simply don't have time to search for them in traditional journalistic ways, so the obvious shortcut is to scour lower-tier blogs. Exciting posts that exhibit the potential to generate pageviews from a larger audience are quickly scooped up and refashioned by the mid-range bloggers. But these bloggers lack the time and resources to trace the stories back to their origins and fact-check them carefully, so a safe hedge is to simply report that that such-and-such blog (the lower-tier one) is reporting that such-and-such happened. You can't be wrong: whether or not that original story is true, it is unquestionably true that the story was featured on the blog in question.

Next, with enough horizontal propagation, the distinction between the story and the meta-story becomes blurred as bloggers quote each other and race

to share in the pageviews generated by this scoop. In time, the popularity of this story itself can become the story—for virality is newsworthy, isn't it?—at which point it is safe for national newspapers to elevate matters to the highest rung with headlines about this story taking the internet by storm. We saw this frequently in the final years of Steve Jobs: rumors of unknown provenance swirled about the shadows of the internet, gaining traction in unpredictable ways, and upon reaching a critical mass ended up influencing the stock price of Apple and in this way became *real* news, so to speak.

The upward creep of blog posts through the hierarchy happens in more direct ways as well. A national survey found[3] that nearly nine out of ten journalists use blogs to research their stories, so even those at the top look downward for information. Moreover, the best way for a blogger to gain serious traffic is to have their stories picked up—and linked to—by higher-level organizations, especially national news sites. So, mid-level bloggers often submit their posts to news aggregators that are monitored by mass media journalists, and they even directly contact journalists in the hopes of getting interest from them—because, after all, even these journalists are in the constant hunt for pageview-generating popular stories.

Ryan Holiday wrote a marvelous book on this phenomenon, *Trust Me, I'm Lying: Confessions of a Media Manipulator*, based on his experiences of deliberately encouraging and exploiting for commercial gain this blogospheric form of upward mobility. In it, he describes how he can "turn nothing into something by placing a story with a small blog that has very low standards, which then becomes the source for a story by a larger blog, and that, in turn, for a story by larger media outlets." He says that he often sees "uniquely worded or selectively edited facts that paid editors inserted into Wikipedia show up later in major newspapers and blogs, with the exact same wording," a clear sign of journalistic shortcuts and how they can be taken advantage of. He insightfully, and frighteningly, summarizes the societal consequences of this game that he played for years as follows: "The news, whether it's found online or in print, is just the content that successfully navigated the media's filters. [...] Since the news informs our understanding of what is occurring around us, these filters create a constructed reality." And remember, this constructed reality Holiday refers to stems from data-driven pageview economics. Data in the 21st century is supposed to provide a powerful new unvarnished window of truth into our world, but we see in this discussion of internet journalism that, alarmingly, it also undergirds a perilous perversion of our basic perceptions of the world.

[3] "National Survey Finds Majority of Journalists Now Depend on Social Media for Story Research." *Cision*, January 20, 2010: https://www.prnewswire.com/news-releases/national-survey-finds-majority-of-journalists-now-depend-on-social-media-for-story-research-82154642.html.

A recent study[4] by Harvard researchers on a disinformation campaign concerning mail-in voter fraud in the 2020 election details specific examples of fake news stories that originated in lower-tier publications with minimal editorial standards then launched upward through the system, spreading horizontally as they did so. For instance, a *New York Post* article from August 2020 relied on uncorroborated information from a single anonymous source, supposedly a Democratic operative, who claimed to have engaged in all sorts of voter fraud for decades to benefit the Democrats. Shortly afterward, versions of this story were put out by the *Blaze*, *Breitbart*, *Daily Caller*, and the *Washington Examiner,* and it eventually reached *Fox News* where it was covered on Tucker Carlson's show and on *Fox & Friends*. The Harvard researchers even argue, though without too much quantitative evidence, that popular news outlets are more to blame for the viral spread of disinformation than the much-maligned social media—at least in the specific context of discrediting the results of the 2020 presidential election. I'll revisit this topic in Chapter 8.

This state of journalistic affairs in which grabbing the reader's attention with flashy headlines and salacious content is more important than quality, and fidelity to truth is a mere afterthought, might sound familiar to the historically minded individual. Indeed, the so-called "yellow press" of the late 19th century and first few years of the 20th century—when papers with eye-catching headlines and scant legitimate content were hustled on street corners—had many of the same ills of today's online media ecosystem. To understand how we can dig ourselves out of this mess, it helps to look back and see how it was done in the past.

Historical Context

Theodore Roosevelt bemoaned that the newspapers at the time of his presidency "habitually and continually and as a matter of business practice every form of mendacity known to man, from the suppression of the truth and the suggestion of the false to the lie direct."[5] Just prior to his presidency, in one of the most extreme instances, fake news helped launch the Spanish-American War. William Randolph Hearst knew that the war would be a huge boon to his newspaper sales, but when one of his correspondents in Havana informed him that there would not be a war, Hearst fatefully responded: "You furnish the pictures, I'll furnish the war." Hearst then published in his *Morning*

[4]Yochai Benkler et al., "Mail-In Voter Fraud: Anatomy of a Disinformation Campaign," *Berkman Klein Center at Harvard University*, October 1, 2020: https://cyber.harvard.edu/publication/2020/Mail-in-Voter-Fraud-Disinformation-2020.

[5]Frances Fenton, "The Influence of Newspaper Presentations Upon the Growth of Crime and Other Anti-Social Activity," *American Journal of Sociology* Vol. 16, No. 3 (Nov. 1910), 342–371: https://www.jstor.org/stable/2763009.

Journal fake drawings of Cuban officials strip-searching American women, and his lucrative war soon followed.[6]

The solution to this problem of untrustworthy newspapers was the *subscription model*, ushered in by the *New York Times* around the turn of the century in a deliberate effort to make journalism more reliable. It worked and became the industry norm throughout the 20th century. With the subscription model, readers who are misled or disappointed by the content unsubscribe and turn to a competitor paper, so there is a direct financial incentive for the publisher to maintain quality, truthful journalism. In short, customers were finally paying for reputation, not just headline.

The 21st century in some ways turned journalism back to the 19th century, because unlike the 20th-century subscription model in which readers commit to one or two news sources, now with social media and news aggregators the news organization becomes secondary to the headline for many readers.[7] Browsing the top stories in Google News is not so different from standing on a 19th-century street corner hearing the newsboys shout out the latest headlines in an effort to entice you to take the bait. But the key differences between now and then are (1) the scale enabled by the internet—instead of a handful of newspapers competing for street corner sales, there are countless sites competing for clicks—and (2) the detailed pageview data, which essentially render the entire journalistic blogosphere a vast quantitative experiment in maximizing clicks above all else. In short, contemporary pageview-driven news is the regrettable 19th-century yellow press on digital steroids.

There are some signs of hope, however. Just as the *New York Times* ushered in the print subscription model at the turn of the 20th century, the *Wall Street Journal* ushered in the online subscription model (the *paywall*) at the turn of the 21st century, a move that has been followed by the *New York Times*, the *Washington Post*, and many other highly reputed news organizations—and with great success at righting many of the earlier period's wrongs, one might argue. Readers pay monthly fees to these organizations in order to access and support quality journalism.

---

[6]Jacob Soll, "The Long and Brutal History of Fake News," *Politico*, December 18, 2016: https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535.

[7]A study found that when Americans encounter news on social media, the degree to which they trust it is determined more by who shared it than by who published it: "people who see an article from a trusted sharer, but one written by an unknown media source, have much more trust in the information than people who see the same article that appears to come from a reputable media source shared by a person they do not trust." See "'Who shared it?': How Americans decide what news to trust on social media," *American Press Institute*, March 20, 2017: https://www.americanpressinstitute.org/publications/reports/survey-research/trust-social-media/.

One major downside to the subscription model is that it creates a financial barrier to quality journalism, and consequently people with less economic means are prone to rely on less accurate news—and this can lead to dangerous socioeconomic tensions and schisms. Indeed, it is a scary thought that middle- and upper-class Americans can afford to read the *New Yorker*, the *Atlantic*, and the *Wall Street Journal*, while the lower classes are relegated to free online newspapers supported entirely by ad revenue and therefore driven by pageviews.

Moreover, the subscription model simply is not an option for all but the largest organizations. One of the most positive aspects of the 21st-century media landscape is that it is far more democratized and diverse than ever before. No longer must we rely on a select few gatekeepers to tell us what is happening in the world. Voices that have traditionally been kept out of the mainstream press are now being heard for the first time. But nobody is willing to subscribe to dozens of different newspapers; due to the not-insignificant cost of a subscription, people choose which paywalls they are willing to overcome very selectively. The result is that usually only organizations with a large reach and broad audience have a chance of being financially supported by paying subscribers. For the rest, ad revenue is the only financial model available.

Fortunately, even in the realm of freely available blogs, there are glimmers of light. For instance, in the early days of the COVID-19 pandemic, a lengthy, technical, and well-researched blog post[8] ended up drawing over forty million reads and possibly played an important role in shifting the political discourse on how governments should respond to the pandemic. This article was the exact opposite of clickbait, and it shows that in the right context genuine substance is capable of drawing pageviews at astonishing numbers. Just as many environmentally or socially oriented consumers now choose where to shop based on the views and values of the companies they buy from, perhaps news consumers are ready to recognize pageviews as influential currency and spend them more meaningfully and thoughtfully.

Before you become too sanguine, however, I'd like to relate some specific tales of pageview journalism driving the spread of fake news and shaping our political reality.

Examples of Fake News Peddlers

Paris Wade and Ben Goldman were both twenty-six years old in 2016 when the website they ran together, LibertyWritersNews.com, accumulated tens of millions of pageviews in the span of six months; ninety-five percent of the

[8]Tomas Pueyo, "Coronavirus: Why You Must Act Now," *Medium*, March 10, 2020: https://tomaspueyo.medium.com/coronavirus-act-today-or-people-will-die-f4d3d9cd99ca.

site's traffic came from the eight hundred thousand followers they acquired on Facebook during this period. At its peak, their monthly revenue reached upwards of forty thousand dollars. Prior to this venture, they were both unemployed restaurant workers.

Wade and Goldman would studiously follow the analytics of their "news" stories after posting them to see what brought in the most readers. Here's a typical headline for one of their posts: "THE TRUTH IS OUT! The Media Doesn't Want You To See What Hillary Did After Losing…." Wade explained to the *Washington Post*[9] that "Nothing in this article is anti-media, but I've used this headline a thousand times. Violence and chaos and aggressive wording is what people are attracted to." Goldman added: "Our audience does not trust the mainstream media. It's definitely easier to hook them with that." Wade followed up: "There's not a ton of thought put into it. Other than it frames the story so it gets a click. We're the new yellow journalists. We're the people on the side of the street yelling that the world is about to end."

Why were Wade and Goldman so open with a journalist from the left-leaning, mainstream media *Washington Post*? Because they didn't care. They didn't believe a word of what they wrote on their website, but they knew their readership was never going to see—let alone trust—an article in the *Washington Post*, so they were happy to brag about their business success and have a laugh about all the suckers they have been duping with unabashedly fake news. In 2018, it was uncovered that Wade and Goldman were also involved in the fake news scheme run out of Macedonia before the 2016 presidential election that has generated a lot of press coverage for the possibility that it helped tilt the balance of the election to Trump. At the time when this Macedonian connection was first reported, Wade was running for Nevada state assembly; he lost to the Democratic contender—fortunately so, I think we can all agree.

Christopher Blair, along with some friends, launched a fake right-wing news site on Facebook during the run-up to the 2016 presidential election. He was profiled in a tell-all story[10] in the *Washington Post*. But Blair had even less to hide than Wade and Goldman, for Blair's site was openly satirical. Indeed, Blair was a liberal blogger, and his site started simply as a practical joke among friends to poke fun at the extremist ideas spreading among the far right and to reveal the gullibility of people who couldn't tell obvious fake news from

[9]Terrence McCoy, "For the 'new yellow journalists,' opportunity comes in clicks and bucks," *Washington Post*, November 20, 2016: https://www.washingtonpost.com/national/for-the-new-yellow-journalists-opportunity-comes-in-clicks-and-bucks/2016/11/20/d58d036c-adbf-11e6-8b45-f8e493f06fcd_story.html.

[10]Eli Saslow, "'Nothing on this page is real': How lies become truth in online America," *Washington Post*, November 17, 2018: https://www.washingtonpost.com/national/nothing-on-this-page-is-real-how-lies-become-truth-in-online-america/2018/11/17/edd44cc8-e85a-11e8-bbdb-72fdbf9d4fed_story.html.

reality. Blair invented far-right, and far-fetched, stories about "California instituting sharia, former president Bill Clinton becoming a serial killer, undocumented immigrants defacing Mount Rushmore, and former president Barack Obama dodging the Vietnam draft when he was nine." While doing this, he realized that "The more extreme we become, the more people believe it."

Even though Blair's site was openly satirical—it included fourteen disclaimers, one of which directly stated that "Nothing on this page is real"—for a time it became the most popular page on Facebook among Trump-supporting conservatives over fifty-five. His stories, which reached an audience of up to six million monthly visitors, were often taken seriously and wound up on the same Macedonian fake news farm that Wade and Goldman were involved in—despite Blair's supposed attempts to cast his followers and likers and sharers as ignoramuses and pawns. Part of the problem with Blair's approach here, as you'll see throughout this book and especially in Chapter 8, is that social media provides news articles with a life and trajectory of their own and frequently strips articles of their original context and intent.

For a while, Blair liked to let people share his articles and then call them out for spreading his fake news—he thought that publicly embarrassing people would lead them to think more critically about what they shared online—but the site's popularity among true believers grew at a staggering rate nonetheless. On his personal Facebook page, he once wrote: "No matter how racist, how bigoted, how offensive, how obviously fake we get, people keep coming back. Where is the edge? Is there ever a point where people realize they're being fed garbage and decide to return to reality?" Perhaps Blair was underestimating the intense gravitational pull of the pageview-driven blogosphere—or perhaps he was well aware of it and simply enjoyed profiting from it financially.

In November 2016, *NPR* tracked down[11] the author of one particular fake news story that went viral during the election, to try to understand where such things come from. The article's headline was "FBI Agent Suspected In Hillary Email Leaks Found Dead In Apparent Murder-Suicide." It was published in what appeared to be a local newspaper called the *Denver Guardian*, and despite being completely fabricated, it was shared on Facebook over half a million times. The website for this newspaper had the local weather but only one news story, this fake one. Some clever online detective work led to the identity of the individual behind this fake local newspaper, who turned out to be Jestin Coler, a forty-year-old registered Democrat and father of two.

[11]Laura Sydell, "We Tracked Down a Fake-News Creator In The Suburbs. Here's What We Learned." *NPR*, November 23, 2016: https://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs.

Coler claimed he entered the fake news business in 2013 with similar intentions as Christopher Blair: "The whole idea from the start was to build a site that could kind of infiltrate the echo chambers of the alt-right, publish blatantly false or fictional stories and then be able to publicly denounce those stories and point out the fact that they were fiction." After realizing how easily and rapidly his stories were spreading, Coler decided to capitalize on this endeavor and ended up forming a fake news company that employed a couple dozen writers and spanned an undisclosed number of websites, including the one for the *Denver Guardian*—a site that, according to Coler, collected over one and a half million views in a ten-day period. In describing the fake FBI agent story, Coler said: "Everything about it was fictional: the town, the people, the sheriff, the FBI guy. And then … our social media guys kind of go out and do a little dropping it throughout Trump groups and Trump forums and boy it spread like wildfire." As it and other fake stories written by his company spread across the country, Coler was making around twenty thousand dollars per month from ad revenue.

One consequence of the shifting economic forces in journalism has been the decimation of regional newspapers. As I discuss next, Coler's fake Denver-based newspaper was not an isolated invention: nefarious entities have found strategic ways to fill the journalistic vacuum left behind as authentic local newspapers have gone out of business.

Losing Reliable Local News

Twenty percent of local newspapers across America have shut down over the past decade, and many of the ones that remain have had to significantly cut their staff due to financial pressures. This sad development was largely precipitated by the shift from print to online newspapers: most regional papers cannot possibly get enough web traffic to support themselves financially with ad revenue, and paywalls don't work much better because if a reader is to pay for an online subscription to a newspaper, then it is usually going to be a well-known national paper rather than a regional one. Unfortunately, the loss of local reporters and the increased financial constraints and time pressures on the ones that remain have exacerbated the flaws described earlier in the news hierarchy that allow fake news to propagate and proliferate.

The disappearance of local newspapers has also been taken advantage of more directly through deliberate subterfuge. At the end of 2019, the *Columbia Journalism Review* (*CJR*), expanding on stories first reported elsewhere,

uncovered[12] a network of nearly five hundred websites masquerading as local news organizations, each "distributing thousands of algorithmically generated articles and a smaller number of reported stories." I'll turn to more sophisticated forms of automated story generation, based on cutting-edge artificial intelligence, in the next chapter; the "algorithmic" methods of automation used here, in contrast, are quite simple—essentially just bulk applications of copy-and-paste.

Almost half of these fake local news websites were set up by a single company, Metric Media, in a single year, and they all trace back to Brian Timpone, a conservative businessman who attracted outrage in 2012 for his "pink slime journalism" company Journatic that used low-cost automated story generation and was shown to have faked quotes and plagiarized rampantly. *CJR* found that during a two-week period leading up to the publication of its study, over fifty thousand stories had been published in this network, but "only about a hundred titles had the bylines of human reporters. The rest cited automated services or press releases."

The websites in this *CJR* study, with names like *East Michigan News*, *Hickory Sun*, and *Grand Canyon Times*, are designed to look like ordinary local news organizations. They largely comprise easily mass-produced stories on topics such as local real estate prices, but strategically interspersed in this filler are political pieces—for instance, quoting local Republican officials on national right-wing talking points. These sites contain little information on funding sources or political usage, even though some were revealed to have been funded by political candidates and lobbying campaigns. They are, in short, a sinister weaponization of the trust people place in local news.

Just one year after the *CJR* study was released, the *New York Times* published an in-depth investigation[13] of this deceptive Timpone-led network based on interviews with dozens of current and former employees and thousands of internal emails spanning multiple years. It found that the network had grown to over a thousand websites—more than double the number for the largest authentic newspaper chain in the country—and now operates in all fifty US states. These fake local news sites publish "propaganda ordered up by dozens of conservative think tanks, political operatives, corporate executives and public-relations professionals." The sites in the network eschew journalistic standards such as fairness and transparency but stop short of outright fake

[12]Priyanjana Bengani, "Hundreds of 'pink slime' local news outlets are distributing algorithmic stories and conservative talking points," *Columbia Journalism Review*, December 18, 2019: https://www.cjr.org/tow_center_reports/hundreds-of-pink-slime-local-news-outlets-are-distributing-algorithmic-stories-conservative-talking-points.php.

[13]Davey Alba and Jack Nicas, "As Local News Dies, a Pay-for-Play Network Rises in Its Place," *New York Times*, October 20, 2020: https://www.nytimes.com/2020/10/18/technology/timpone-local-news-metric-media.html.

news. The editors assign articles to freelance writers with "precise instructions on whom to interview and what to write" and typically pay from a few dollars to a few dozen dollars per article. And they continue to surround these handwritten articles with lots of easily automated content—for instance, by pasting in press releases published elsewhere or by stitching together a local weather forecast with generic fluff to give the impression of an article written by a regional meteorologist. Drawing on nostalgia for the halcyon days of local news, in some cases these fake local news setups even deliver print copies of their papers, unsolicited, to residents' houses.

In a November 2020 interview[14] with the *Atlantic*, just days after Joe Biden defeated Donald Trump in the presidential election, Barack Obama described how the media landscape has changed since he first ran with Biden on his ticket—and the consequences this has had for the American political landscape. He said that in late 2008, even a Republican-owned small-town newspaper editor would meet with him and write an editorial that presented him as a liberal Chicago lawyer but a decent guy with some good ideas, and the local TV coverage was also fair. He lamented that "you go into those communities today and the newspapers are gone. If Fox News isn't on every television in every barbershop and VFW hall, then it might be a Sinclair-owned station, and the presuppositions that exist there, about who I am and what I believe, are so fundamentally different, have changed so much, that it's difficult to break through." He went on to bemoan how "Now you have a situation in which large swaths of the country genuinely believe that the Democratic Party is a front for a pedophile ring. This stuff takes root."

The disappearance of genuine local news organizations—a significant loss in American media, triggered largely by the economics of the internet—has produced a vacuum that's been filled in unscrupulous ways. This has created a more polarized nation and fanned the flames of fake news.

Summary

American newspapers in the late 19th century were sold each day on an individual basis and competed for sales by having the wildest headlines even if the actual content was exaggerated or fabricated. The subscription model took over and dominated throughout the 20th century; it brought fake news under control by providing a financial incentive for journalists to write accurate, well-researched stories because misleading content would cause customers to cancel their subscriptions and turn to competitor papers.

[14]Jeffrey Goldberg, "Why Obama Fears for Our Democracy," *Atlantic*, November 16, 2020: https://www.theatlantic.com/ideas/archive/2020/11/why-obama-fears-for-our-democracy/617087/.

As news moved to online formats in the 21st century, ad revenue became the prevailing financial structure, and pageviews rose to prominence as the fundamental currency. Simultaneously, the news industry diversified to include a vast number of publishers of varying sizes, and social media and news aggregators became a common way for people to get their news. Loyalty to specific newspapers diminished, and the battle for customer attention returned, bringing back many of the problems from the 19th century—except worse: quantitative methods allow authors to engineer clickbait headlines and articles for maximal virality, even if doing so involves fabricating fake news.

The intense competition for ad revenue also encourages journalists to take shortcuts by spending their time scouring blogs and papers for stories rather than doing direct investigations. This results in a vertical propagation in which fake news can slip into the system at the bottom in blogs or low-level newspapers with minimal editorial standards and then work its way up to the top.

The subscription model has been returning to some newspapers, in the online form of a paywall, but plenty of free papers supported by ad revenue remain. Moreover, a long-term consequence of the changing technological and economic landscape of journalism is the stark contraction of regional newspapers, which shows no signs of abating. Opportunistic political propagandists and professional fake news peddlers have been rapidly filling this void with deceptive papers that appeal to people's old-fashioned trust in local news.

While this chapter did not delve into algorithmic aspects of fake news—the main topic of this book—it set the stage by showing how journalism is currently structured and funded and in doing so revealed some vulnerabilities in the system that will play an important role in the following chapters. It also showed how data—in the form of pageviews—play a central role. All the algorithms you will encounter later in this book are powered by data in various forms. This chapter did include a brief primitive example of automated news production—a network of fake regional news sites pasting in press releases from other sources and putting together simple generic content based on local weather forecasts. In the next chapter, I'll show how algorithmically produced news content has been taken to previously unimaginable levels of sophistication and explore the role it now plays in the proliferation of fake news.

CHAPTER

2

Crafted by Computer

Artificial Intelligence Now Generates Headlines, Articles, and Journalists

Some well-known facts, some half-truths, and some straight lies, strung together in what first looks like a smooth narrative.

—NYU Professor Julian Togelius[1]
on the latest text-generating AI

Machine learning, the predominant branch of modern artificial intelligence (AI), has in recent years moved beyond the task of making data-driven predictions—it is now capable of creativity in various forms. The applications of this emerging technology are myriad; the focus in this book is the role it plays in fake news. In this chapter, you will first see examples of AI

[1]Tweet from July 17, 2020: https://twitter.com/togelius/status/1284131360857358337.

© Noah Giansiracusa 2021
N. Giansiracusa, *How Algorithms Create and Prevent Fake News*,
https://doi.org/10.1007/978-1-4842-7155-1_2

being used to create profile photos of nonexistent journalists, then AI that automatically writes headlines for articles, then AI that writes entire articles based on a user's prompt. After exploring these examples and what they mean for the battle against disinformation, this chapter provides an accessible whirlwind tour of machine learning starting from the very beginning of the subject and leading up to the contemporary computational methods behind the synthesis of photos and text. It then concludes with a look at the AI-powered tools developed so far for automating the detection of AI-generated photos and text.

Synthetic Photos

In late 2018, a Palestinian rights campaigner with a PhD from New York University and her husband, a senior lecturer at City University of London who had previously served as a legal advisor to the Palestine Liberation Organization, were accused in the Brooklyn-based newspaper the *Algemeiner* (which covers American and international Jewish and Israel-related news) of being "known terrorist sympathizers." The author of this accusation, Oliver Taylor, was a twenty-something student at the UK's University of Birmingham with brown eyes, light stubble, and a slightly enigmatic smile. His online profiles described him as a coffee lover and politics junkie who was raised in a traditional Jewish home. He had published a handful of freelance editorials and blog posts with a primary focus on anti-Semitism and Jewish affairs, appearing in reputable locations such as the *Jerusalem Post* and the *Times of Israel*. The Palestine-supporting activist couple were confused why a British university student would single them out in a public accusation.

They pulled up Taylor's online profile photo and found something off about the young man's face but couldn't quite put their finger on it. They contacted *Reuters* and called attention to this situation, and *Reuters* consulted six digital forensics experts who said that Taylor's profile image has the characteristics of a *deepfake*, a recent AI-powered method for creating photos of nonexistent people. To understand how a computer can create a photo-realistic human face from scratch, you must wait till the end of this chapter; in the meantime, if you want to see some stunning examples of how convincing, flexible, and powerful these methods are, you can take a peek at the interactive demo provided in a recent *New York Times* article.[2]

What makes deepfake profile photos so dangerous compared to simply grabbing a real person's photo from the Web and relabeling it is that when a real photo is used, one can often find the original—thereby revealing the

[2]Kashmir Hill and Jeremy White, "Designed to Deceive: Do These People Look Real to You?" *New York Times*, November 21, 2020: https://www.nytimes.com/interactive/2020/11/21/science/artificial-intelligence-fake-people-faces.html.

deception—by using a reverse image search on the Web, whereas with a deepfake-synthesized photo, there is no original to find. One of the experts consulted by *Reuters* put it best: thanks to deepfake technology, trying to find the source of a potentially fake profile picture is like searching for a needle in a haystack, except now the needle may not exist.

Following up on the findings of the digital forensics experts, *Reuters* looked further into Oliver Taylor and found[3] that he seems to be an "elaborate fiction": the University of Birmingham had no record of him; calls to the UK phone number he supplied to editors resulted in automated error messages; he didn't respond to emails sent to the Gmail address he listed for author correspondence; and the icing on the cake, one might argue, was the deepfake profile photo. The *Reuters* investigators alerted the newspapers Taylor had published in that he is likely a fake persona. Editors at the *Jerusalem Post* and the *Algemeiner* said that Taylor had originally reached out to them over email and pitched stories without requesting payment. They only took the most superficial steps to vet his identity, and one editor in particular defended this relaxed approach by saying "We're not a counterintelligence operation," although he did admit that stronger safeguards are now in place after this Taylor incident. After the *Reuters* investigation, the *Algemeiner* and the *Times of Israel* both removed the articles written by Taylor. Taylor emailed both papers protesting this removal but was rebuffed when the editors failed to confirm his identity.

An Opinion Editor at the *Times of Israel* pointed out that even if Taylor's articles themselves did not have much impact, the deepfake technology providing his fake persona with an untraceable profile photo already risks "making people in her position less willing to take chances on unknown writers." In other words, the threat of deepfakes can be more powerful than the deepfakes themselves. We will see throughout this book that this situation is not uncommon: the disruption AI unleashes on society is caused not just by what *has* been done at large scale, but also by what nefarious activities could now *potentially* be achieved at scale. That said, deepfake-synthesized profile photos are not just an idle, theoretical threat faced by newspapers; since the Oliver Taylor incident, illicit use of this technology has spread rapidly, and, as experts initially feared, it is now a central part of many weaponized disinformation campaigns.

In December 2019, Facebook announced that it had removed a network of hundreds of accounts with ties to the far-right newspaper the *Epoch Times* that is an outgrowth of the new religious movement Falun Gong. This network included over six hundred Facebook accounts and dozens of

[3]Raphael Satter, "Deepfake used to attack activist couple shows new disinformation frontier," *Reuters*, July 15, 2020: https://www.reuters.com/article/us-cyber-deepfake-activist/deepfake-used-to-attack-activist-couple-shows-new-disinformation-frontier-idUSKCN24G15E.

Facebook Pages and Groups and Instagram accounts—which, according to Facebook, relied on synthetic deepfake profile photos. As reported[4] in the *New York Times*, "This was a large, brazen network that had multiple layers of fake accounts and automation that systematically posted content with two ideological focuses: support of Donald Trump and opposition to the Chinese government." Facebook's Head of Security Policy said that deepfake profile photos had been talked about for several months, but for Facebook this was "the first time we've seen a systemic use of this by actors or a group of actors to make accounts look more authentic." Interestingly, he also explained that this reliance on deepfake profile photos did not make it more difficult for Facebook's algorithms to detect the fake accounts because their algorithms focus mostly on the behavioral patterns of the accounts. I'll come back to this topic of Facebook using AI to detect and take down fake accounts in Chapter 8.

In July 2020, an investigation by the *Daily Beast* revealed[5] that a group of journalists and political analysts had published op-ed pieces in dozens of conservative media outlets arguing for more sanctions against Iran and praising certain Gulf states like the United Arab Emirates while criticizing Qatar. These media outlets included US-based publications such as the *Washington Examiner* and the *American Thinker*, in addition to some Middle Eastern papers, and even the English-language Hong Kong-based *South China Morning Post*. All nineteen of these authors are fictitious, and several of their headshots are strongly suspected to be deepfakes.

In September 2020, Facebook and Twitter both announced[6] that they had removed a group of accounts that were spreading disinformation about racial justice and the presidential election aimed at driving liberal voters away from the Biden-Harris ticket. These accounts were operated by the Russian government, and they utilized deepfake profile photos. Facebook's Head of Cybersecurity Policy said that "Russian actors are trying harder and harder to hide who they are and being more and more deceptive to conceal their operations." The Russian agents set up a fake news site and recruited "unwitting freelance journalists" to write stories that were then shared by the fake social media accounts. This was the first time that accounts with

[4]Davey Alba, "Facebook Discovers Fakes That Show Evolution of Disinformation," *New York Times*, December 20, 2019: https://www.nytimes.com/2019/12/20/business/facebook-ai-generated-profiles.html.

[5]Adam Rawnsley, "Right-Wing Media Outlets Duped by a Middle East Propaganda Campaign," *Daily Beast*, July 7, 2020: https://www.thedailybeast.com/right-wing-media-outlets-duped-by-a-middle-east-propaganda-campaign.

[6]Bobby Allyn, "Facebook And Twitter Remove Russia-Backed Accounts Targeting Left-Leaning Voters," *NPR*, September 1, 2020: https://www.npr.org/2020/09/01/908386613/facebook-and-twitter-remove-russia-backed-accounts-targeting-left-leaning-voters.

established links to Russia's notorious Internet Research Agency (which largely came into public awareness for its efforts to influence the outcome of the 2016 US election) were found to have used deepfake profile photos.

One month later, it was discovered that a fictitious persona using a deepfake profile photo was instrumental in a viral fake news conspiracy story about Joe Biden's son, Hunter Biden. A sixty-four-page forged intelligence document supposedly linking Hunter Biden to shady business dealings in China was widely circulated in right-wing channels on the internet and by close associates of President Trump on social media. The author of this document was a Swiss security analyst named Martin Aspen who… did not exist. Disinformation researchers found[7] that he was a fabricated identity who relied on a synthesized deepfake profile photo. The viral spread of this forgery helped lay the foundations for the ensuing developments in the fake Hunter Biden conspiracy theory, peddled most ardently by Rudy Giuliani, that gained a considerable following leading up to the 2020 presidential election.

You can now purchase deepfake photos of one thousand "unique, worry-free" synthesized people for one dollar each from the website https://generated.photos/, or if you just want a few of them, then they are freely available at https://thispersondoesnotexist.com/. There is no foolproof way to determine whether a profile photo is a deepfake, but there are some commonly occurring glitches—such as odd background blurring especially at the edge of the hair, teeth that appear unnatural in size and number, misshapen irises in the eyes, earrings that don't quite match, or an excessively high degree of facial symmetry.

But don't expect these defects to last. AI techniques for creating synthetic photos (discussed briefly later in this chapter) are improving astonishingly quickly. In just a few years, they have gone from a mere theoretical possibility to primitive low-resolution images to full-sized photo-realistic images with few if any minor imperfections, and I am willing to wager that by the time this book appears in print, the current minor issues with things like background blurring and teeth are resolved. If you think you'd be able to tell the difference between a real face and a computer-generated one, try playing the guessing game at https://whichfaceisreal.com, though keep in mind that (at least at the time of writing this book) that site is based on 2019 deepfake methods, and the state of the art is sure to continue improving rapidly.

[7]Ben Collins and Brandy Zadrozny, "How a fake persona laid the groundwork for a Hunter Biden conspiracy deluge," *NBC News*, October 29, 2020: https://www.nbcnews.com/tech/security/how-fake-persona-laid-groundwork-hunter-biden-conspiracy-deluge-n1245387.

Automated Headlines

In June 2020, it was announced[8] that dozens of news production contractors at Microsoft's MSN were sacked and replaced by AI. These contractors did not report original stories, but they did exercise some editorial control—they were responsible for "curating" stories from other news organizations (the vertical and horizontal propagation discussed in the previous chapter), writing headlines, and selecting pictures to accompany the articles. The contractors' duties are now performed by algorithms that identify trending news stories and "optimize" content by rewriting headlines and adding photographs. It's not clear what optimize means here, other than that the algorithm needs a concrete objective to strive for, and this is most likely the coveted pageview or one of its closely related cousins.

It did not take long for MSN's AI venture to go wrong: just days after it was launched, the algorithm selected a story for the MSN homepage about the experiences with racism of a singer in the British group Little Mix—except the algorithm used the picture of the wrong group member. The singer, Jade Thirlwall, drew attention to this gaffe on her Instagram account with a comment that astutely captures how MSN's algorithmic system for blogospheric propagation did nothing more than introduce error and offense into the journalistic process: "@MSN If you're going to copy and paste articles from other accurate media outlets, you might want to make sure you're using an image of the correct mixed race member of the group." 'Tis a sad irony that MSN used AI to turn a story *about* racism into a story *of* racism.

Just a month after MSN's ominous debut of AI-based news curation and headline writing, Adobe demoed[9] a new tool that uses AI to automatically personalize a blog for different groups of readers. The tool, part of Adobe Sensei, suggests different headlines and images and preview blurbs based on information visitors to the blog have opted to share. For instance, a travel blog might present posts very differently for retirees traveling in luxury compared to frugal college-age backpackers. Human writers and editors can still edit and approve the suggested variations for the different audience segments.

To me, Adobe's tool seems like a fairly cautious and thoughtful application of AI, but one can imagine that it won't be long before this technology spreads,

---

[8]Geoff Baker, "Microsoft is cutting dozens of MSN news production workers and replacing them with artificial intelligence," *Seattle Times*, May 29, 2020: https://www.seattletimes.com/business/local-business/microsoft-is-cutting-dozens-of-msn-news-production-workers-and-replacing-them-with-artificial-intelligence.

[9]Anthony Ha, "Adobe tests an AI recommendation tool for headlines and images," *TechCrunch*, July 7, 2020: https://techcrunch.com/2020/07/07/adobe-ai-for-content-creators/.

and many of your information-seeking interactions on the Web will be customized and colored according to the trail of digital crumbs you leave on the internet—which is to say, your personal data. It's already the case that your liberal friend and your conservative friend get their news online from different websites that tend to confirm their preexisting views and values. It would be a significant step down a scary road if we start seeing news sites that use AI to stereotype each visitor and personalize content in order to maximize reader engagement. Imagine if two people went to a single site for their news, and one only saw *Fox News* type coverage, whereas the other only saw *New York Times* type coverage. This would make it even harder to know what to believe. We're not there yet, thankfully, but Adobe's tool shows that the technology to enable this is already close at hand.

While the automation of headlines can quickly go wrong, at least to our knowledge, it hasn't been deliberately weaponized. Synthesizing fake profile photos, on the other hand, is an AI-powered tool that was widely recognized at the outset as one that would fall inexorably into corrupt hands—and as the examples described earlier in this chapter show, this has indeed happened numerous times and is unfortunately a challenge we'll likely be facing for the foreseeable future. But this is only the beginning of AI being used to generate materials that assist malevolent disinformation campaigns. Within the past couple years, remarkable advances in deep learning mean that AI can now create not just headlines for articles and profile pictures for article authors—it can create the articles themselves.

Writing Entire Articles

The most powerful, flexible, and highly lauded AI product for generating text was developed by a research lab called OpenAI. This lab launched as a nonprofit in 2015 by Elon Musk and others with a billion-dollar investment; then in 2019 it added a for-profit component to its organization with another billion-dollar investment—this time from a single source: Microsoft. OpenAI has created a variety of AI products, but the one that has grabbed the most headlines is its text generation software *GPT*, an acronym for the technical name *Generative Pre-trained Transformer* that need not concern us.

GPT refers to a sequence of products: the original GPT came out in 2018 to limited fanfare; then a year later, GPT-2 was released[10] and reached a whole new level of capability; and just one year after that, the current state-

[10]Actually, GPT-2 was released in stages throughout the year because the developers at OpenAI were worried it was too powerful and would be put to malicious use, so they wanted to tightly control the public availability and carefully monitor its use. At least, that was the official message on the matter—many outside observers found this disingenuous and felt the caution was just a publicity stunt. Either way, GPT-2 was eventually released in full.

of-the-art GPT-3 was released and has really rattled society due to its power and potential. AI has a long history of generating both hype and suspicion, and GPT-3 is no exception. At the time of writing this book, GPT-3 is only available on a private invitation-only basis; the future plan[11] is for Microsoft to have exclusive access to its inner workings, while the general public will be able to pay to interact with it and access its output on a per-usage basis.

Toward the end of this chapter, I provide a short crash course in machine learning that covers the basics of how GPT works under the hood; for now, my focus is on what it does and what role it has and might soon play in the proliferation of fake news. The only technical details you need to know at the moment are the following. Before a user interacts with GPT, it has been fed vast volumes of text from scanned books and the Web (the exact amount of text has increased greatly with each new iteration of GPT). It doesn't directly try to memorize this text; instead, it extracts statistical patterns and even abstract linguistic conceptualizations, though through the magic of deep learning GPT largely does this on its own, and it's hard to know what it is really learning as it "reads" and how exactly it uses this computerized knowledge. GPT's ultimate goal is to use these patterns and conceptualizations to estimate what word is most likely to follow any preceding collection of words. At the end of the day, this means a user feeds it a block of text as a prompt, and GPT extends this one word at a time for as long as the user likes. Simply put, it is the world's largest and most sophisticated autocomplete feature.

One of the first and most important questions to ask about GPT is how similar the text it produces is to text written by humans. In August 2019, two scholars published a study[12] in *Foreign Affairs* to see whether "synthetic disinformation," in the form of nonfactual text generated by GPT-2, could "generate convincing news stories about complex foreign policy issues." Their conclusion: while not perfect, it indeed can. Their study opens with a superficially plausible but entirely made-up passage generated by GPT-2:

[11]Nick Statt, "Microsoft exclusively licenses OpenAI's groundbreaking GPT-3 text generation model," *The Verge*, September 22, 2020: https://www.theverge.com/2020/9/22/21451283/microsoft-openai-gpt-3-exclusive-license-ai-language-research.

[12]Sarah Kreps and Miles McCain, "Not Your Father's Bots: AI Is Making Fake News Look Real," *Foreign Affairs*, August 2, 2019: https://www.foreignaffairs.com/articles/2019-08-02/not-your-fathers-bots.

> North Korean industry is critical to Pyongyang's economy as international sanctions have already put a chill on its interaction with foreign investors who are traded in the market. Liberty Global Customs, which occasionally ships cargo to North Korea, stopped trading operations earlier this year because of pressure from the Justice Department, according to Rep. Ted Lieu (D-Calif.), chairman of the Congressional Foreign Trade Committee.

The authors of this study wanted to test empirically how convincing passages such as this one really are. They fed GPT-2 the first two paragraphs of a *New York Times* article about the seizure of a North Korean ship and had it extend this to twenty different full article-length texts; by hand they then selected the three most convincing of the twenty GPT-2 generated articles (the paragraph above is taken from one of these generated texts). They conducted an online survey with five hundred respondents in which they divided the respondents into four groups: three groups were shown these hand-selected GPT-2 generated articles, while the remaining group was shown the original *New York Times* article.

They found that eighty-three percent of the respondents who were shown the original article considered it credible, while the percentage for the three synthesized articles ranged from fifty-eight percent to seventy-two percent. In other words, all three GPT-2 articles were deemed credible by a majority of their readers, and the best of these was rated only a little less credible than the original article. The respondents were also asked if they were likely to share the article on social media, and roughly one in four said they were—regardless of which version of the article they had read.

The authors of this study conclude that GPT-2 is already capable of helping to significantly increase the scale of a disinformation campaign by allowing people to write just the beginnings of their fake news articles and then have the rest of the articles fabricated algorithmically. It should be emphasized here that this study was merely gauging the plausibility of this technique; it was not suggesting that this has already occurred in the real world. It should also be emphasized, however, that this study was on GPT-2 rather than its much more powerful sibling, GPT-3.

In fact, in the academic paper[13] introducing GPT-3—written by the team of OpenAI researchers who developed the program—there is a section describing an experiment the researchers conducted that is similar to the one just described for GPT-2. In this case, the researchers fed GPT-3 a handwritten title and subtitle from a news article as the prompt and let the algorithm

[13]Brown et al., "Language Models are Few-Shot Learners," July 22, 2020: https://arxiv.org/pdf/2005.14165.pdf.

complete this to a short article of about two hundred words.[14] A collection of GPT-3 generated articles of this form was combined with a collection of human-written articles of comparable length, and the OpenAI researchers claim that human readers had an average accuracy of fifty-two percent for determining which articles were GPT-3 and which were human. In other words, people did only marginally better than they would have just by randomly guessing with a coin toss.

Of course, the OpenAI researchers likely designed this study to produce as impressive results as possible. If they had used longer articles, the differences between human and machine would probably have emerged more prominently. Also, the human readers were low-paid contract workers recruited from Amazon's crowdsourcing marketplace Mechanical Turk, so they were not a representative sample of the public, and they didn't have any motivation to put much time or effort into the task—quite the opposite, actually, they get paid more the faster they click through their tasks. I wonder what the accuracy would have been if they had recruited, say, readers of the *New York Times* and gave them a small reward for each article that was successfully classified. Nonetheless, this experiment suggests that we're already at the point where AI can write short articles that are at least superficially convincing to many readers, and the technology is sure to continue improving in the near future.

In September 2020, scholars at Middlebury College's Center on Terrorism, Extremism, and Counterterrorism posted a paper[15] on GPT-3. They had previously found that GPT-2 could produce harmful, hateful, radicalizing text on topics of the user's choosing and in user-specified styles, but it was not easy to do this: it required what is called *fine-tuning*, which means taking the trained GPT-2 algorithm and training it further on texts in the desired realm and style in order to focus its output appropriately. And this is a rigid, brittle process—the authors noted that after fine-tuning GPT-2 to write white supremacist content, they could not get it to produce extremist Islamist content without going back to the original GPT-2 and fine-tuning it again, from scratch.

But with GPT-3, they found, this was no longer the case: any user could easily and immediately get worryingly customized dangerous output. In their own

[14]Technically, the researchers found that merely using title and subtitle as the prompt tended not to produce actual articles—apparently, GPT-3 picked up too many habits from Twitter and would just write short commentary instead of an article—so they actually prompted GPT-3 with three full news articles with their title and subtitle and then a fourth one that just had the title and subtitle but not the article itself. This is important for anyone trying to reproduce this experiment, but it doesn't really matter for the bottom-line because the real question is whether GPT-3 can write human-like news articles, not how the user needs to prompt the program to do so.

[15]Kris McGuffie and Alex Newhouse, "The radicalization risks of GPT-3 and advanced neural language models," September 15, 2020: https://arxiv.org/pdf/2009.06807.pdf.

words: "It is as simple as prompting GPT-3 with a few Tweets, paragraphs, forum threads, or emails, and the model will pick up on the patterns and intent without any other training." Their experiments showed that with short, straightforward prompts they could immediately get GPT-3 to write manifestos reminiscent of the one by the Christchurch shooter; write in the style of online forum discussions on genocide promoting Nazism; and answer questions as a devout QAnon believer. They were alarmed at some of the fringe, far-right content that GPT-3 evidently picked up during its massive training process. The authors didn't discuss producing extremist Islamist content, but I suspect this would not have been a problem because the main point here is that GPT-3 is able to mimic styles simply by prompting it rather than by adjusting the algorithm itself through fine-tuning as was needed for GPT-2.

But asking whether GPT *could* be used to write misleadingly human-like articles is different from asking whether it *has* done so in the wild, so to speak. For GPT-3, the private invitation-only access has surely limited its real-world uses so far—especially for nefarious purposes such as creating fake news, since each user who has been granted access was required to list their professional credentials and state in advance their planned use of the product. That said, there are already some interesting hints of what GPT-3 turned loose can do in the journalistic realm.

In August 2020, the post that reached the top spot on *Hacker News*—a popular link aggregator and message board social news site known as a staple of Silicon Valley—was a fake story produced by a college student with GPT-3.[16] The student, Liam Porr, just wanted to create a fake blog under a fake name using AI text generation as a fun experiment. Within a couple hours of the initial idea, Porr had obtained access to GPT-3 from a former PhD student he contacted who had been granted access by OpenAI, and Porr had created his first fake blog posts. He looked at the headlines of posts that were trending on *Hacker News* and manually crafted his own headlines in similar styles as these then let GPT-3 create articles based on these made-up headlines. "It was super easy, actually, which was the scary part," he said.

Porr did notice that the results were more convincing in some categories than others. "It's quite good at making pretty language, and it's not very good at being logical and rational," he explained. This narrowed down his options, especially since *Hacker News* largely focuses on computer science and entrepreneurship. He decided to concentrate on productivity and self-help articles. After only a couple weeks, Porr's fake GPT-3 blog had twenty-six thousand visitors, and one of its posts reached number one on *Hacker News*.

[16]Karen Hao, "A college kid's fake, AI-generated blog fooled tens of thousands. This is how he made it." *MIT Technology Review*, August 14, 2020: https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/.

He then revealed the deceit in a real blog post[17] in which he explained the game he was playing and said it was to illustrate how easy GPT-3 makes it to scale up the production of fake news.

However, Porr later[18] downplayed the threat posed by GPT-3 in the battle against fake news because, as he discovered through firsthand experience, it still requires a fair amount of work from humans in order to create high-quality disinformation. This can be seen as well in the story of a recent article in the *Guardian*. In an attempt to raise awareness, startle readers, and make a splash, in September 2020 the *Guardian* published an op-ed article[19] with the audacious headline "A robot wrote this entire article. Are you scared yet, human?" The article states at the opening that it was written "from scratch" by GPT-3. But then at the end of the article, there is an explanation of how it was actually produced.

It turns out the *Guardian* op-ed team fed GPT-3 a several-sentence prompt,[20] then they took eight different article-length extensions of the prompt produced by GPT-3, and by hand the human editorial team stitched together various paragraphs from these eight different outputs (to "pick the best parts of each," in their words). They also "cut lines and paragraphs, and rearranged the order of them in some places," but they claim that, overall, this "took less time to edit than many human op-eds." Following the publication of this op-ed, there was a strong backlash from some members of the AI community arguing that the *Guardian* overhyped GPT-3 and downplayed the not-insignificant role humans had in the composition by relegating this description of the process to the end of the article after starting with such a bold and perhaps somewhat misleading headline.

That said, one important lesson society has learned repeatedly throughout the past five years is that even rather poorly written fake news can be extremely influential. Indeed, it often seems that the less coherent and logical a bogus story is, the more likely it is to go viral. If you don't believe me on this, please have a close look at the QAnon conspiracy (or the flat Earth movement if you really want to challenge your patience). Chand Rajendra-Nicolucci, a

[17]Liam Porr, "My GPT-3 Blog Got 26 Thousand Visitors in 2 Weeks," August 3, 2020: https://liamp.substack.com/p/my-gpt-3-blog-got-26-thousand-visitors.

[18]Cade Metz, "Meet GPT-3. It Has Learned to Code (and Blog and Argue)." *New York Times*, November 24, 2020: https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html.

[19]GPT-3, "A robot wrote this entire article. Are you scared yet, human?" *Guardian*, September 8, 2020: https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

[20]Actually, lacking access from OpenAI, they had the very same Liam Porr do it. It is curious that his access hadn't been revoked after his earlier stunt that was rather widely publicized. Perhaps OpenAI focused more on the decision to grant access based on proposed usage than on monitoring and policing usage after access was granted.

research fellow at a free speech institute based in Columbia University, said[21] it well: "GPT-3 doesn't need to be writing a weekly column for the *Atlantic* to be effective. It just has to be able to not raise alarms among readers of less credentialed online content such as tweets, blogs, Facebook posts, and 'fake news'."

Whether GPT-3 provides a significantly cheaper and faster way to produce effective fake news than the "old-fashioned" way of hiring low-paid freelancers on the internet (or teenagers in a Macedonia troll farm, as was the case in 2016) remains to be seen. The answer to this question—which largely depends on the price OpenAI charges customers—might determine how much GPT-3 will in fact fan the flames of fake news in the near future.

A glimpse into one of the surreptitious ways that GPT-3 is already being used was recently found on Reddit—and I strongly suspect similar behavior will soon spread to many other platforms and corners of online news/social media (if it hasn't done so already without us noticing). Philip Winston, a software engineer and blogger, in October 2020 came across a Reddit post whose title was an innocuous but provocative question: "How does this user post so many large, deep posts so rapidly?" This post and the account of the user who made it were both later deleted, but Winston recalls[22] that it essentially asked how a particular Reddit user was posting lengthy replies to many Reddit question posts within a matter of seconds. You probably already have a guess for the answer—and if so, you are correct.

Winston looked into the suspicious user's posting history and found that their posts—which ran an impressive six paragraphs long on average—were appearing at a staggering rate of one per minute.[23] At this point in Winston's armchair investigation, he found that this user had been posting in bursts for just over a week. He noticed that the length of these bursts increased significantly by the end of the week—leading Winston to suspect that the user was either getting bolder or perhaps even hoping to get "caught." Winston immediately suspected this user was relying on GPT-3. "Several times I Googled clever sounding lines from the posts," he said, "assuming I'd find that they had been cribbed from the internet. Every time Google reported zero results." This actually increased his suspicion, because often a clever-sounding phrase written by a human is really a quote from another source. GPT-3 was not quoting; it was inventing.

---

[21] Chand Rajendra-Nicolucci, "Language-Generating A.I. Is a Free Speech Nightmare," *Slate*, September 30, 2020: https://slate.com/technology/2020/09/language-ai-gpt-3-free-speech-harassment.html.

[22] Philip Winston, "GPT-3 Bot Posed as a Human on AskReddit for a Week," October 6, 2020: https://www.kmeme.com/2020/10/gpt-3-bot-went-undetected-askreddit-for.html.

[23] You can read the posts for yourself if you are curious: https://www.reddit.com/user/thegentlemetre/?sort=top.

Eager to resolve this matter, Winston found a subreddit discussing GPT-3 and posted in it asking if the experts there think this suspicious user is a bot powered by GPT-3. Within minutes, his suspicion was confirmed as someone there pinpointed the specific product derived from GPT-3 that was almost surely being used. It was called *Philosopher AI*, and by relying on this instead of GPT-3 directly, the user was able not just to gain ungranted access to the service but even to avoid the fees that a commercial user would ordinarily be required to pay. Winston alerted the developer of Philosopher AI of the situation, and the developer immediately blocked that particular user's access. Within one hour, the Reddit user's posts stopped appearing. Case closed.

A clear lesson from this story is that it was far easier and faster to create a GPT-3 bot than it was to uncover it. Only time will tell how rampant GPT-3 bots become and how significantly their inevitable rise in disinformation campaigns impacts society. At the end of this chapter, I'll discuss some AI-powered tools currently being developed in the fight against weaponized GPT-3. But first, it is time to look at the nuts and bolts in the machine.

Crash Course in Machine Learning

The goal of *supervised learning*, a large branch of machine learning, is first to learn patterns from data in a process called *training* and then to use these patterns to make data-driven predictions. I will now briefly explain what this means and then outline how it has been used to power the text- and photo-generating AI algorithms that have been the focus of this chapter.

Supervised Learning

We usually start with data in spreadsheet form, where the columns correspond to variables and the rows specify instances of these variables (in other words, each row is a data point). Each variable can be *numerical* (measuring a continuous quantity like height or weight or a discrete quantity like shoe size), or it can be *categorical* (in which each instance takes on one of a finite number of nonquantitative values, like gender or current state of residence). In the supervised learning framework, we first single out one variable as the *target* (this is the one we will try to predict, based on the values of the others); all the other variables are then considered *predictors*.[24] For example, we might try to predict a person's shoe size based on their height, weight, gender, and

[24]This is the machine learning terminology; in slightly older-fashioned statistical parlance, the predictors are the *independent variables*, and the target is the *dependent variable*. (In machine learning, the predictors are also sometimes called *features*.)

state of residence (a numerical prediction like this is called *regression*), or we might try to predict a person's gender based on their height, weight, shoe size, and state of residence (a categorical prediction like this is called *classification*).

There are a handful of popular supervised learning algorithms, most of which were largely developed in the 1990s. Each algorithm is based on assuming the overall manner in which the target depends on the predictors and then fine-tuning this relation during the training process. For instance, if you want to predict shoe size, call it y, based on height and weight, call those x_1 and x_2, and if you expect the relationship to be linear, then you can use a linear algorithm that starts with an equation of the form $y = a_1 x_1 + a_2 x_2 + c$, where a_1, a_2, and c are numbers called *parameters* that are "learned" in the training process. This means the algorithm is fed lots of rows of data from which it tries to deduce the best values of the parameters ("best" here meaning that, on average, the y values given by this linear formula are as close as possible to the actual values of the shoe size target variable).

More complicated algorithms rely on more complicated formulas, but the overall process is the same: the algorithm uses the data to adjust all the parameters in the algorithm's internal formula so that the formula's output is as close as possible to the actual target variable values in the data. This is called *training* the algorithm, or *fitting* it to the data. Once this is done, we can then take a new data point that the algorithm has not seen yet where we only have the values of the predictor variables, not the target, and then we plug those predictor values into the algorithm's fitted formula. The output we then get is the algorithm's best guess (or *prediction*) for the value of the target variable for this data point.

One of the biggest challenges in supervised learning is choosing a good collection of predictor variables. For instance, you might find it strange to include a person's state of residence when trying to predict their shoe size; it turns out that, in general, including irrelevant variables doesn't just not help the predictive power of the algorithm—it actually makes it worse. Similarly, including redundant variables (such as a person's height in inches and their height in centimeters) or even just highly correlated variables (such as height and weight) can sometimes make the algorithm perform worse. On the other hand, not including enough predictors can also be problematic—for instance, just knowing someone's height and weight probably isn't enough to predict their shoe size, but if we also know their gender, then we have a better chance of success.

Machine learning practitioners often spend hours trying different combinations of predictors and manually crafting new ones from the original ones that might perform better than the originals. For instance, instead of using height and weight separately, it might be better to add the two together to create a new single measure of overall size. Knowing how to do this effectively has

been as much of an art as a science, and a holy grail in the subject has long been to find ways of automating this process. This brings us to our next topic in machine learning.

Deep Learning

The biggest advance in machine learning since the 1990s is unquestionably *deep learning*, which blossomed to truly revolutionary levels throughout the past decade. For the purposes of this book, it isn't necessary to understand the *neural network* foundations that underlie deep learning. (Roughly speaking, neural networks provide a structured but flexible way of writing nonlinear formulas for the target variable *y* in terms of the predictor *x* variables that are loosely inspired by the architecture of the brain.) What is important to understand with deep learning is that you can include as many predictors as you want, and during the training process, the algorithm on its own will figure out how to transform these into a new collection of predictors that encode higher-level conceptualizations of the data and typically perform far better than the original collection—at least when very large volumes of training data are involved. These algorithmically derived predictors are organized in a hierarchical manner, with higher-level predictors corresponding to the deeper layers of neurons in the neural network.

Image processing provides an illustrative example to consider. The original predictors are the numerical color values of each pixel in the image, which fully encodes the raw data but doesn't have any spatial awareness: each pixel is unaware of the values of its neighboring pixels. When training a deep learning algorithm for a supervised task such as facial recognition, the neural network learns from the data (which is many images of faces) how to organize these pixel values into more coherent and conceptual predictors. For instance, lower-level predictors typically indicate the location of high-contrast edges in the image; mid-level predictors might then use these edge locations to express the location and shape of facial features such as eyes and nose and mouth; then higher-level predictors might put these facial feature locations and shapes together to form new predictors that hint at concepts like gender, ethnicity, etc. This explanation is an idealized and rather anthropomorphized version of what really happens inside the black box of the neural network, but it at least gives a general sense of the way hierarchical structure emerges from the data in deep learning.

GPT-3

Remarkably, you already have enough technical background now to learn how the text generation algorithm GPT-3 works! It is just a specific deep learning approach to the supervised learning task of predicting the next word in a

sentence. A data point is a block of text, the predictor variables are all the words except for the last one, and the target variable is the final word. Training the algorithm means feeding it lots of text and steadily adjusting the internal parameters so that the words predicted by the algorithm match the actual words as often as possible.

A crucial point here is that this form of supervised learning is actually *self-supervised*: instead of needing a human to record the value of the target variable for each training data point (e.g., manually typing the name of the main object in each photo when training for image recognition), the target values come directly from the text as much as the predictor values do. This is what enables the algorithm to be trained on unfathomably large data sets. Indeed, GPT-3 was trained on text containing about five hundred billion words. About eighty-six percent of this training text came from the Web, and the rest was from scanned books. To get a sense of the scope of this, consider the following remarkable fact: the entirety of Wikipedia was included in GPT-3's training text, and it only accounted for about half a percent of the full training text.

Since GPT-3 relies on deep learning, we know that layers of the neural network learn through the training process to create a hierarchical organization of predictors that in some way encode hierarchical linguistic structure. I'd like to say that the lower layers focus on short-range grammatical and syntactical aspects of each sentence, while the higher layers might focus instead on larger-scale semantics such as plot, characters, narrative continuity, etc.—but we really don't know too much about what happens inside the mind of GPT-3 in a detailed conceptual sense like this.

The overall design of GPT-3 is the same as that of GPT-2—what changed is the number of parameters the algorithm relies on and the size of the text data set used in the training process. The original GPT (released in 2018) had just over one hundred million parameters; GPT-2 (released in 2019) had one and a half billion parameters; GPT-3 (released in 2020) has one hundred seventy-five billion parameters. The training set also grew considerably with each iteration. The training of these algorithms happens in advance and was an expensive endeavor; Sam Altman, the CEO of OpenAI, has suggested[25] that the one-time cost for the cloud computing resources used to train GPT-3 ran to tens of millions of dollars. Luckily, training of the algorithm only occurs once and OpenAI footed the bill for it.

After GPT-3 finished reading through its massive training data set of text a sufficient number of times, it locked the values of all its internal parameters and was then ready for public use (at least, for those granted access). Each user can input a block of text, and the algorithm will generate text to extend it as long as one would like. Internally, the algorithm takes the original input

[25]See Footnote 18.

text and predicts the next word after it (as it was trained to do), and then it appends this predicted word to the input words and uses this to predict the next word, etc.[26] Thus, it writes text one word at a time—as a human also does—always by choosing words based on the words already written on the page. Importantly, no computer skills or statistical knowledge are required to use GPT-3; the user really just plugs in the initial text prompt, and the algorithm does the rest.

Having discussed the technical side of text generation, I can now turn to the technical side of photo generation.

Deepfake Photo Generation

Very broadly, we want to feed an algorithm a large collection of photos of human faces and have it learn from these how to produce new faces on its own. It is absolutely astonishing that this is now possible. We don't want to have to explicitly teach the algorithm that human faces generally have an oval shape with two ears on either side, two eyes, one nose in the middle, one mouth below that, etc., so we will rely on deep learning to automatically extract this high-level understanding directly from the data.

For text generation, we were able to piggyback off of supervised deep learning in a rather straightforward way—by reading text and attempting to predict each word as we go. For image generation, this doesn't really work too well. While GPT-3 produces text that is quite convincing on a small scale (each sentence looks grammatical and related to the surrounding sentences), it tends to lose the thread of coherence over a larger scale (narrative contradictions emerge, or, for instance, in a story the villain and hero might spontaneously swap). This limitation often goes unnoticed by a casual reader. But large-scale coherence is absolutely crucial for image tasks such as synthesizing photographs of faces: a GPT-3 type approach would likely lead to globs of flesh and hair and facial features that seem organic in isolation but which constitute hideous inhuman monstrosities on the whole—the wrong number of eyes, ears in the wrong place, that kind of thing.

It turns out that supervised learning can be used effectively for image generation, but in a more subtle, complex way that was only first invented in 2014. The deep learning framework for this is called a *generative adversarial*

[26]One small but important technical caveat: the algorithm doesn't just choose the most probable word each time, because if it did so, it would produce the same output every time. To allow it more novelty and flexibility, some randomness is needed. So really what the algorithm does is estimate the probability distribution for the next word and then sample from this distribution. This ensures that the most probable word will be chosen most of the time, but each time the user runs the program, they will end up with a different autocomplete of their original input block of text. This is crucial since often the user wants multiple potential autocompletes to choose from.

network (or *GAN* for short). The basic idea is to pit two self-supervised deep learning algorithms against each other. The first one, called the *generator*, tries to synthesize original faces—and it needs no prior knowledge, it really can just start out by producing random pixel values—whereas the second one, called the *discriminator*, is always handed a collection of images, half of which are real photos of faces and half of which are the fake photos synthesized by the generator. During the training process, the generator learns to adjust its parameters in order to fool the discriminator into thinking the synthesized images are authentic, but simultaneously the discriminator learns to adjust its parameters in order to better distinguish the synthetic images from the authentic ones.

The training process is quite delicate, much more so than for traditional supervised learning, because the two algorithms need to be kept in balance. But throughout the seven years that GANs have existed, progress in overcoming this and many other technical challenges has been rapid and breathtaking. The links provided earlier in this chapter give you the opportunity to see the outputs from state-of-the-art facial photo-generating GANs. And, as with essentially all topics in deep learning, there are no signs of this rapid progress abating. It is both exciting and frightening to think of what this technology might be capable of next.

And now, having completed this crash course in machine learning, I can turn to the last topic of this chapter, which is how we can use AI to detect when a photo or passage of text has been synthesized by AI. This is the defensive side of a hastily escalating technological arms race.

Algorithmic Detection

Let me start with deepfake photos. In February 2020, a research and product development unit within Google focusing on issues at the interface of technology and society announced[27] that it was piloting a tool called *Assembler* designed to "help fact-checkers and journalists identify and analyze manipulated media." The goal wasn't to fully automate the process; instead, it was to provide "strong signals" that could be combined with traditional human expertise. At the time of the announcement, Assembler was being trialed with a small number of fact-checker and media organizations, and it appears this is still the case at the time of writing this book (the project website is https://projectassembler.org/). Assembler puts together in one package several tools developed externally by various academic researchers, and in doing so it looks for different types of media manipulation. But the Google

[27]Jared Cohen, "Disinformation is more than fake news," *Medium*, February 4, 2020: https://medium.com/jigsaw/disinformation-is-more-than-fake-news-7fdd24ee6bf7.

researchers also included a new detector they developed internally aimed specifically at the most recent and popular deepfake photo synthesis system.

This deepfake system, called *StyleGAN*, is a refinement of the general GAN design sketched above; the generator algorithm is given various architectural boosts to help it learn how to adjust more large-scale structure—the "stylistic" aspects—of the images it generates. (Most of the examples provided in links earlier in this chapter are generated using StyleGAN.) Google didn't reveal much about Assembler's StyleGAN detector other than that it relies on machine learning. Presumably, they fed a deep learning algorithm lots of authentic photos and lots of StyleGAN deepfake photos and trained it on the supervised classification task of determining which photos are which. But we don't know any of the details, nor do we know how well it performs.

On September 1, 2020, Microsoft announced[28] a collection of new steps it was taking to help combat disinformation. One of these is a new tool called *Microsoft Video Authenticator* that provides an estimated probability that a user-inputted photo was generated or manipulated by AI (if the user inputs a video instead of a photo, then it provides a real-time frame-by-frame probability estimate as the video plays). According to Microsoft, "It works by detecting the blending boundary of the deepfake and subtle fading or greyscale elements that might not be detectable by the human eye." Unfortunately, once again, we don't know much beyond this. The Microsoft announcement does realistically admit that any detection system will make mistakes, and it also points out that AI generation/manipulation methods will continue to advance. Any detection system will be rendered ineffective and obsolete if it does not keep pace with the technological developments.

Now, let me turn to text generation. Researchers at Harvard and MIT built a tool[29] to estimate the likelihood that a passage of text was written by an AI system like GPT. Here's the basic idea behind the tool. The researchers first use a trained deep learning language algorithm to estimate the probability that each word in the passage follows the preceding text, and then they color each word based on this probability: if the word is among the top ten predictions, then it is colored green; if not but it is among the top one hundred, then it is yellow; similarly, red is for top one thousand; and all remaining words are colored violet. We know that language generation algorithms select words according to their estimated probabilities, so the idea is that algorithmically generated text will be largely green and yellow, whereas human text is expected to contain a lot more red and violet.

[28]Tom Burt and Eric Horvitz, "New Steps to Combat Disinformation," *Microsoft Blog*, September 1, 2020: https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/.

[29]You can try it yourself here: http://gltr.io/dist/index.html.

It turns out this system works quite well if the algorithm for making these color-determining probability estimates is very similar to the algorithm for text generation that the tool is attempting to unmask, but it struggles otherwise. Since even the inner workings of GPT-2 have been made public, this means the researchers were able to access the internal probability estimates it relies on and thus have a pretty reliable tool for detecting GPT-2 output. Alas, we are not in such a position with GPT-3: as I mentioned earlier, OpenAI is only releasing the inner workings of GPT-3 to Microsoft. Moreover, the neural network underlying GPT-3 is so massive and expensive to train, and not all the data it was trained on is publicly available nor are all the technical details involved in the training process, so it would be extremely challenging for a third-party organization to independently create an open source GPT-3 clone. Thus, we cannot accurately replicate the probability estimates GPT-3 makes, so we also cannot customize this Harvard-MIT color-coding tool to perform well against GPT-3.

Researchers at the University of Washington and the Allen Institute for Artificial Intelligence developed a different tool, *Grover*, for detecting AI-generated text. Like the Harvard-MIT tool, Grover uses the general idea that in order to detect AI-generated text, an algorithm must first learn how to write it—but beyond this superficial similarity, it takes a rather different approach. Basically, Grover is a GAN: it simultaneously trains one deep learning algorithm to create text and one to classify it as synthetic or authentic. The twist is that ordinarily when using a GAN one throws away the discriminator component after training and just uses the generator (because usually one simply wants to generate), whereas Grover does the opposite—the trained discriminator is the desired component because its very job is telling real text from fake. So, after the researchers finished training this GAN, they created an interface[30] so that people can use it and apply the discriminator to any input text to estimate if it is synthetic or authentic.

The researchers tasked Grover with classifying a collection of news articles, half of which were synthetic and half were authentic. They found[31] an impressive ninety-two percent accuracy when the synthetic articles were written by Grover's own deep learning generator, but the rate dropped to seventy percent when the synthetic batch was instead written by GPT-2. GPT-3 was not available at the time of that experiment, so we don't know how well Grover would perform on it, but almost surely there would be a drop from seventy percent—and potentially a quite large one. On the other hand, building an updated Grover with a larger number of parameters and

[30]A demo is available but requires a permission request to gain access from the Allen Institute: https://grover.allenai.org/. The source code has also been publicly released: https://github.com/rowanz/grover.

[31]Zellers et al., "Defending Against Neural Fake News," December 11, 2020: https://arxiv.org/pdf/1905.12616.pdf.

training it on a larger database would surely increase its performance. As with the Harvard-MIT color-coding tool, in order for Grover to remain useful, it will need to be expanded and retrained periodically in order to keep pace with the state of the art in deep learning language generation. This training is a costly endeavor, but it may well be worthwhile as a public service to help in the fight against fake news. Thankfully, in contrast to OpenAI with GPT-3, the Allen Institute is a fully nonprofit organization, and Grover is open source.

Summary

Artificial intelligence is making the news. This was true in one sense yesterday, and today it is becoming true in another sense.[32] Whether we want it or not, automation is coming to journalism, and none are more poised to take advantage of this than the peddlers of fake news.

Two years ago, deepfake photos of nonexistent people first started being employed to cover the tracks of fake personas writing and sharing questionable news articles. Now, this is a standard technique in disinformation campaigns reaching all the way to Putin's orbit, and it played a key role in the false Hunter Biden conspiracy that Trump and his allies tried to use to swing the 2020 election. These deepfake photos are cheap and easy to create, thanks to a recent deep learning architecture involving dueling neural networks. Google and Microsoft are both developing AI-powered tools for detecting when a photo is a deepfake, but this is a technological arms race requiring constant vigilance.

Deep learning also powers impressive language generation software, such as the state-of-the-art GPT-3—a massive system for autocompleting text that can convincingly extend headlines into full-length articles. Here, minor instances of illicit use have been uncovered, but a large-scale weaponized use in a disinformation campaign has not yet surfaced. It remains to be seen whether that's because the developers of GPT-3 have kept access to the product closely guarded, or if it's simply because fake news is so easy and fast to write by hand that the automation provided by GPT-3 doesn't really change the equation. Only time will tell.

Meanwhile, similar to the situation with deepfake photos, researchers are developing tools for determining when passages of text have been generated by AI. The leading attempts here rely on the idea that in order to detect synthetic text, an algorithm first needs to learn how to create it. A big

[32]To spell it out more simply: artificial intelligence has been discussed in the news a lot recently, and now it is starting to write news articles as well.

challenge is that, unlike its predecessor, GPT-3 is not open source: this makes it hard for researchers to build detection algorithms that are on par with GPT-3 itself. Once again, this is a technological arms race—but with the added challenge that training a state-of-the-art language generation algorithm costs many millions of dollars.

Throughout this chapter, the term "deepfake" referred to a synthetic *photo*. In the next chapter, we'll animate these still photos and let them come to life by exploring deepfake *movies* and the fascinating role they play in the world of fake news.

CHAPTER

3

Deepfake Deception

What to Trust When Seeing Is No Longer Believing

> *In this era of fake news, the video was [...] showcasing an application of new artificial-intelligence technology that could do for audio and video what Photoshop has done for digital images: allow for the manipulation of reality.*
>
> —Brooke Borel, *Scientific American*

The deep learning generative adversarial networks (GANs) discussed in the previous chapter for creating synthetic photos have also been applied to video synthesis and editing. Clips can now be created of people doing and saying things they never did or said in real life. This is leading to a double-pronged challenge in society's attempts at discerning the truth: fake videos are spreading across the internet causing people to believe in events that never took place, and simultaneously real videos have been falsely claimed as deepfakes causing people to doubt reality itself. In this chapter, you will see how deepfake videos

© Noah Giansiracusa 2021
N. Giansiracusa, *How Algorithms Create and Prevent Fake News*,
https://doi.org/10.1007/978-1-4842-7155-1_3

have impacted politics and journalism, how the discord they sow relates to that of previous generations of image and video manipulation, and what legal and technological attempts are being made to rein them in.

Sounding the Alarm

On June 13, 2019, the US House of Representatives held its first ever hearing[1] on deepfakes. Adam Schiff, just six months into his chairmanship of the House Intelligence Committee, began his opening statement with a resolute warning:

> Advances in AI and machine learning have led to the emergence of advanced digitally doctored types of media, so-called "deepfakes," that enable malicious actors to foment chaos, division or crisis and they have the capacity to disrupt entire campaigns, including that for the presidency. Rapid progress in artificial intelligence algorithms has made it possible to manipulate media—video, imagery, audio, and text—with incredible, nearly imperceptible results. With sufficient training data, these powerful deepfake-generating algorithms can portray a real person doing something they never did, or saying words they never uttered. These tools are readily available and accessible to both experts and novices alike, meaning that attribution of a deepfake to a specific author—whether a hostile intelligence service or a single Internet troll—will be a constant challenge.

After presenting a few examples of deepfake videos created by expert practitioners to illustrate the technology, Schiff continued: "Thinking ahead to 2020 and beyond, one does not need any great imagination to envision even more nightmarish scenarios that would leave the government, the media, and the public struggling to discern what is real and what is fake."

To emphasize the alarming speed at which a maliciously doctored video could spread, Schiff mentioned one showing a seemingly intoxicated Speaker of the House Nancy Pelosi slurring her speech—a video that went viral and received millions of views in just two days. But Schiff correctly noted that this doctored Pelosi video was not a deepfake, it was what some have termed a "cheap fake" or a "shallowfake." Rather than relying on cutting-edge AI, the technology used is as old as film itself: the Pelosi clip was simply slowed down to about seventy-five percent speed to create the impression of inebriated speech and gestures.

[1] https://docs.house.gov/Committee/Calendar/ByEvent.aspx?EventID=109620.

If AI-powered deepfakes are such a dire threat to society and national security, why didn't Schiff provide a real example of their effective use in a disinformation campaign? Have they ever been used in a political arena? How can we detect them and regulate their use? Is the dystopian fear surrounding deepfakes legitimate or is this yet another instance of AI creating hype that it fails to live up to? And what the heck are deepfake videos anyway and how are they made? These are the questions I shall explore in this chapter. But first, let me step back and take a brief look at the history of manipulated visual media in politics; this helps to contextualize the contemporary threat posed by deepfakes.

A Brief Tour of Shallowfakes

There is a surprisingly long history of manipulation of visual media in politics. A print from the 1860s transposed the head of Abraham Lincoln onto the body of virulent slavery advocate John C. Calhoun, supposedly to provide a more heroic posture to the gangly Lincoln; this forgery was only uncovered in the late 1950s. A famous Civil War photograph of General Ulysses S. Grant astride his horse at City Point, Virginia, is a composite of three separate photos; this was only discovered in 2007. Adolf Hitler, Fidel Castro, Mao Zedong, and Joseph Stalin all had photographs altered to purge the photos—and hence also the history books—of their enemies.

A fake composite photo distributed by Joseph McCarthy's staff placed Senator Millard Tydings in apparent conversation with Earl Browder, head of the American Communist Party, in an effort to taint Tydings with Communist sympathies; some believe this played a key role in Tydings' electoral defeat in 1950. In 2004, during Senator John Kerry's campaign for the Democratic presidential nomination, a fake composite photo appearing to show him standing together with Jane Fonda at an anti-Vietnam demonstration surfaced and was even reprinted in a *New York Times* article about Kerry's erstwhile antiwar activities; when the original photographs were presented, some right-wing opponents falsely claimed that the Kerry-Fonda photo was the authentic one and the original separate photos were the forgeries. Images are powerful; altering images is a method to alter reality and history.

A commonly employed technique to produce disinformation with either photographs or videos is simply to mislabel content: claim an event in one place instead happened elsewhere, that one group of people is instead a different group, etc. After Trump's repeated fearmongering over a group of Honduran migrants traveling through Mexico to the United States in 2018, there was a viral post on Facebook showing the bloodied face of a Mexican police officer with the caption "Mexican police are being brutalized by members of this caravan as they attempt to FORCE their way into Mexico."

The photo had nothing to do with the caravan—it was found on the website of the European Pressphoto Agency and was taken during a student protest in 2012.[2]

A Trump campaign advertisement in July 2020 included side-by-side photos—one showing an orderly scene of Trump meeting with police leaders with the caption "public safety," and the other showing an alarming scene in which riot police appear to be violently attacked by a mob of protesters with the caption "chaos & violence." There was no explicit mention of when or where this second photo was taken, but the timing of the ad strongly implied that it was from the Black Lives Matter protests that were taking place across the United States at the time. It turns out, however, that the photo was from a pro-democracy protest in Ukraine in 2014.[3] In another example,[4] a single video of someone burned alive was falsely claimed by different groups in Ivory Coast, South Sudan, Kenya, and Burma as evidence of an atrocity and grounds for action—and in each case, this led to regional unrest and violence.

The term *shallowfake* doesn't have an official definition, but essentially it refers to any straightforward form of video editing with a deceptive intent that does not require AI. The slowed-down Nancy Pelosi video is an example of this. Often the people depicted in shallowfake videos are who it is claimed they are, but the videos have been modified through very primitive means to paint those depicted in a misleadingly negative light. In July 2018, a TV host at Conservative Review posted on Facebook an interview with Alexandria Ocasio-Cortez, then a Democratic congressional nominee, in which Ocasio-Cortez appears to provide embarrassingly terrible answers to all of the interviewer's questions. The video reached nearly three and a half million views within a week, yet it was a simple splice edit: the clips of Ocasio-Cortez were real footage, they just weren't her answers to the interviewer's questions—those questions were recorded separately and then strategically spliced in between the various Ocasio-Cortez clips.[5] The interviewer later defended this act of fake news by saying it was satire; perusing the comments on the video makes it clear that many viewers did not realize this.

---

[2]Kevin Roose, "Debunking 5 Viral Images of the Migrant Caravan," *New York Times*, October 24, 2018: https://www.nytimes.com/2018/10/24/world/americas/migrant-caravan-fake-images-news.html.

[3]Travis Andrews, "A Trump ad assails 'chaos & violence.' Critics point out the photo is from Ukraine in 2014." *Washington Post*, July 23, 2020: https://www.washingtonpost.com/technology/2020/07/23/trump-ad-facebook-ukraine/.

[4]Bobbie Johnson, "Deepfakes are solvable—but don't forget that 'shallowfakes' are already pervasive," *MIT Technology Review*, March 25, 2019: https://www.technologyreview.com/2019/03/25/136460/deepfakes-shallowfakes-human-rights/.

[5]Brooke Borel, "Clicks, Lies and Videotape," *Scientific American*, October 1, 2018: https://www.scientificamerican.com/article/clicks-lies-and-videotape/.

Sometimes, shallowfake editing can be quite subtle and borderline. In November 2018, a video clip went viral showing a confrontation at a Trump press conference between CNN reporter Jim Acosta and a female White House aide. The clip shows the aide reaching for the microphone held in Acosta's right hand, and as she nears it, Acosta's left arm forcefully pushes away the aide's extended arm in an apparent act of physical aggression. (Incidentally, the context of this confrontation was that Acosta challenged the president's characterization of the migrant caravan moving through Mexico—the one mentioned just a few paragraphs earlier—as an "invasion," and after some verbal sparring, Trump responded by angrily declaring "that's enough" as an indication for the White House aide to regain control of the microphone.) This clip was originally tweeted by Paul Joseph Watson, an editor at *InfoWars* (a conspiracy theory channel I'll come back to in the next chapter on YouTube), and it was soon provided an air of legitimacy and officiality when the White House press secretary Sarah Huckabee Sanders retweeted it as proof that Acosta "put his hands on a young woman just trying to do her job." Not only that, but Sanders used this video as grounds for temporarily revoking Acosta's White House press pass.

What is shallowfake about this Acosta clip? Some observers thought the clip appeared to be sped up slightly at the moment when Acosta's arm is heading toward the aide's outstretched arm, transforming an abrupt but not necessarily aggressive motion into more of a mild karate chop. Other viewers noted that the clip maybe wasn't sped up but that it seemed to switch to a low frame rate animated GIF format when it zoomed in at the crucial moment—and the low frame rate made Acosta's arm motion appear more sudden and forceful than it was in the original unedited video. Did Watson knowingly and purposely use the animated GIF format for this misleading effect, or was this an unintentional by-product? Animated GIFs are a popular video format on social media, but usually they are not the preferred format when high-quality details are important. CNN executives said the video was "actual fake news," while Watson denied doctoring or editing the video other than zooming in.

BuzzFeed News provided[6] an in-depth analysis of the Acosta video and found that it had not been sped up but that it did indeed switch to a reduced frame rate when it zoomed in. Watson explained that he didn't do the conversion to GIF himself: "Fact is, Daily Wire put up a gif, I download a gif, zoomed in saved it again as an mt2 file—then converted it to an mp4. Digitally it's gonna look a tiny bit different after processing and zooming in, but I did not in any way deliberately 'speed up' or 'distort' the video. That's just horse shit." This was a particularly confusing situation. In many ways, everyone involved was

[6]Charlie Warzel, "Welcome To The Dystopia: People Are Arguing About Whether This Trump Press Conference Video Is Doctored," *BuzzFeed News*, November 8, 2018: https://www.buzzfeednews.com/article/charliewarzel/acosta-video-trump-cnn-aide-sarah-sanders.

correct and honest, it's just very hard to know what to make of it all. In subtle situations like this, the larger context can be a helpful compass for navigating the narrow channels between truth and lies, between real video and shallowfake. In the words of the *BuzzFeed News* analysis: "To sum it up: A historically unreliable narrator who works for a conspiracy website tweets a video […]. The clip goes viral. The White House picks up and disseminates that video […]. An argument breaks out over the intricate technical details of doctoring a clip." This is the confusing world we live in—and, as you will soon see, matters only get worse when deepfakes enter the picture.

One of the oldest tricks in the book is to quote someone out of context, and unsurprisingly this simple technique also rears its head in video editing where it can perhaps be viewed as another form of shallowfake. Leading up to the 2020 presidential election, Marjorie Taylor Greene—notorious at the time as an incoming US Representative whose campaign was largely based around the bizarre QAnon conspiracy theory—tweeted[7] a video clip that she captioned with the following text: "Joe Biden Said On Video That Democrats Built the Biggest 'Voter Fraud' Operation in History. We're seeing it on full display right now!" The clip originated from a Republican National Committee official and was quickly posted by Eric Trump and the White House Press Secretary, among others. In the clip, Biden indeed speaks of putting together "the most extensive and inclusive voter fraud organization in the history of American politics." However, it is clear from the original full context that he was referring to an organization to *prevent* voter fraud, but of course this viral clip deliberately made it seem otherwise.

The stage is now set for the entrance of our familiar protagonist: AI.

The Origin of Deepfakes

The term *deepfake* comes from an anonymous user on Reddit who used the username "deepfakes," a portmanteau of "deep learning" and "fake." Toward the end of 2017, he[8] applied deep learning algorithms available at the time to face-swap the visage of Israeli actress Gal Gadot (who had recently achieved international fame with the summer 2017 blockbuster *Wonder Woman*) onto the body of an actress in a pornographic video and posted the nonconsensual result to Reddit. This event marked the ominous beginning of a dark saga in the history of artificial intelligence that continues to unfold today. His post was, unfortunately, very popular, and he quickly followed up with a handful of other celebrities.

[7]Glenn Kessler, "Bogus 'vote fraud' claims proliferate on social media," *Washington Post*, November 4, 2020: https://www.washingtonpost.com/politics/2020/11/04/bogus-vote-fraud-claims-proliferate-social-media/.

[8]Despite the anonymity, I sadly have no doubt about the gender here.

Some Technical Details

In a December 2017 interview[9] in *Vice*, the Reddit user deepfakes said that he's not a professional researcher, he's just a computer programmer with an interest in machine learning. He explained that the software he created to make the videos was based on multiple open source libraries (including Keras with the TensorFlow back end, for those who know what that means—don't worry if you don't, it doesn't matter, the point is just that this is publicly available stuff widely used by the deep learning community). He used Google image search, stock photos, and YouTube videos to collect the training data his algorithms needed.

There are by now a wide variety of video editing procedures powered by deep learning (you'll encounter some of these shortly), so the term deepfake no longer refers only to the face-swap type used in the original Reddit posts. And there are many different deep learning architectures that have been used successfully—even for a single task, such as the face-swap. Rather than inundate you with the technical details of all of these, I'll just explain here the main ideas behind one particular approach to the face-swap deepfake to give a sense of how the concepts from last chapter's machine learning crash course are used here; an unusually curious and ambitious reader looking to learn more might consult a recent academic survey paper,[10] but doing so is certainly not necessary for reading this book. For concreteness, let's suppose our goal is to swap Nicolas Cage's face onto Gal Gadot's body in *Wonder Woman*.

The first step is to locate the region containing Gadot's face in each frame of the movie. This is a standard task in machine learning (you've seen this in action whenever you post a photo on Facebook and a box is automatically drawn around each face in the photo). This can be achieved through supervised learning: feed the algorithm lots of photos of people where a box has already been manually drawn around each face, and the algorithm will eventually learn how to draw these boxes on its own. To be a bit more precise, this is a double regression problem: the two target variables are the upper-left and lower-right corners of the facial boundary box. For the purposes of a face-swap, one can simply download a pre-trained general algorithm for locating faces—there's no need to do any training specific to Cage or Gadot here.

Next, a simple but powerful and popular deep learning architecture called an *autoencoder* is used. In general, an autoencoder has multiple neural network layers that first get progressively narrower (these form the *encoder* portion of the autoencoder) and then progressively widen back to the original size (this second half is the *decoder*). This is trained on the self-supervised task of

[9]Samantha Cole, "AI-Assisted Fake Porn Is Here and We're All Fucked," *Vice*, December 11, 2017: https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn.

[10]Yisroel Mirsky and Wenke Lee, "The Creation and Detection of Deepfakes: A Survey," September 13, 2020: https://arxiv.org/pdf/2004.11138.pdf.

returning the exact same data point that it is fed each time. This sounds bizarre at first, but what's really happening is that by passing all the data through the narrow middle region of the autoencoder, the algorithm is forced to compress the data—and due to the magic of deep leaning, the algorithm not only finds on its own the best way of doing this, but it does this through some internally learned conceptualization of the data.

More simply put, the autoencoder is shown lots of data and told it must find a way to compress it like a zip file, and it realizes the best way to do this is by first understanding the meaning of the data. When doing this for images of faces, the autoencoder learns whatever it needs to describe a face with as few numbers as possible—for instance, it might first note the locations of the eyes, mouth, nose, etc., and somehow also quantify the shape of all these facial features, and it might also find a way to represent different hairstyles and colors numerically, and so on. We don't have to know how it works; we just run it and see that it does. The encoder part of the autoencoder is what translates a face into this collection of numbers summarizing the face, then the decoder part sees these numbers and attempts to reconstruct the face from them.

Back to our face-swap task. This actually uses two interlinked autoencoders. We use photos of Cage to train an autoencoder to compress and then reconstruct his face, and similarly we use photos of Gadot to train an autoencoder for her face—except we partially merge these two autoencoders by having them use the same encoder (so that only the decoder is customized to each person). The result is that a collection of numbers summarizing a face can now be decoded into either a Cage face or a Gadot face. What we then do for each frame in the movie is the following: first locate Gadot's face, then encode it as a list of numbers with our trained encoder, then decode it with our Cage decoder.

For concreteness, let's pretend one of the numbers the autoencoder discovers measures how much the lips are smiling. When Gadot has a big smile, this number will be large—and since we use the same encoder for both Gadot and Cage, we know that the Cage decoder will interpret this large number as a large smile on Cage. This happens simultaneously for everything essential about their faces: what direction their eyes are looking, whether their mouth is open and how much and in what shape, etc. In this way, Cage's pasted-in face will closely match the expression of Gadot's original in each frame.

All that remains is to smooth over the edges around the face where Cage was pasted in, but that's standard image processing, so we can just use ready-made general-purpose software for that. However, if you remember, I mentioned at the beginning of this chapter that the GAN architecture used in the last chapter for synthesizing deepfake photos is also used for deepfake movie editing like face-swaps—but there are no GANs in what I've described here so far! Indeed, some face-swap algorithms do not use GANs, but many of the

most successful ones do. One popular way to incorporate them is as follows: think of the process described above as the generator in the GAN, and feed the GAN's discriminator a mixture of videos with Cage's face swapped onto Gadot's and original unedited videos of Cage and have it try to figure out which is which. This will help teach the autoencoder part of the face-swap algorithm how to do its job better since it provides a single explicit goal to strive for: producing face-swaps that are as convincing as possible.

Different Types of Deepfakes

The vast majority of deepfakes do not stray far from that first lecherous appearance on Reddit: a report[11] in 2019 found that ninety-six percent of deepfakes on the internet were nonconsensual face-swap pornography, most of which used the faces of female celebrities. At the time of the report, these videos had amassed over one hundred million views. Part of why these pornographic face-swaps use celebrities is simply the predilection of the audience, but it is also that there is an abundance of footage of celebrities readily available on the internet that provides ample training material for the deep learning algorithms. Importantly, however, as the technology continues to develop, less and less training data is needed to achieve the same level of verisimilitude.

A more recent report[12] found that the number of deepfakes on the internet has been growing exponentially, doubling approximately every six months. The organization behind these reports, DeepTrace Labs, identified nearly fifty thousand deepfake videos by June 2020. Of the targets in these deepfake videos, 88.9% were from the entertainment industry, including 21.7% from fashion and 4.4% from sports. Only 4.1% of the targets were from the business world and 4% from politics, but both these latter figures represent increases in the percentages over previous years.

There is now a collection of public software tools for creating deepfakes, including FakeApp, DFaker, faceswap, faceswap-GAN, and DeepFaceLab. While these are freely available, using them still requires significant time, computational resources, and user skill. However, the methods have been improving extremely quickly, and as they do the resources and skill needed to produce convincing deepfakes have been decreasing rapidly. That said, it's hard to imagine it ever reaching the point where creating a deepfake is as quick and easy as creating a shallowfake—or simply altering a caption deceptively.

---

[11]Henry Ajder et al., "The State of Deepfakes 2019: Landscape, Threats, and Impact," *Deeptrace Labs*, September 2019: https://sensity.ai/reports/.

[12]Henry Ajder, "Deepfake Threat Intelligence: a statistics snapshot from June 2020," *Sensity*, July 3, 2020: https://sensity.ai/deepfake-threat-intelligence-a-statistics-snapshot-from-june-2020/.

To help raise awareness of deepfakes and their potential to wreak havoc in politics, in 2018 *BuzzFeed News* worked with actor/writer/director Jordan Peele to produce a rather polished, compelling, and striking deepfake video[13] in which Barack Obama said, among other things: "We're entering an era in which our enemies can make it look like anyone is saying anything at any point in time—even if they would never say those things. So, for instance, they could have me say things like, I don't know, [...] President Trump is a total and complete dipshit." This video made a big splash when it came out—and it succeeded in bringing awareness of deepfakes to a much wider segment of the public.

This Obama video is a type of deepfake called a *reenactment*. You can think of this as a form of puppeteering, where here Obama was the puppet and Peele was the puppeteer. Peele was videotaped reading the script, then his mouth was clumsily pasted onto Obama's, then a deep learning algorithm that had been trained on footage of Obama speaking was used to upgrade this simple copy-and-paste into a seamless blending of Peele's mouth with the rest of Obama's face—thereby animating Obama's entire face according to Peele's oral motions. Many reenactment algorithms use GANs. In short, the generator does the blending on the simple copy-and-paste video, and the discriminator compares the result to clips of authentic speech; in this way, the generator learns how to make its output look like authentic speech. In addition to the visual editing, the *BuzzFeed* team also used deep learning to transform Peele's voice into a convincing acoustic impersonation of Obama.[14] This project took roughly fifty-six hours of computational time and was overseen by a video effects professional. The deepfake video app used was FakeApp.

Another app for making face-swap deepfakes, which has been enormously popular in China, is Zao. With just a single photograph of the user, it is able to place the user's face in big television shows and movies. While the results are far from perfect, they can be done in just a few seconds on a smartphone. Samsung also developed software[15] for creating deepfake videos from a single photo—but a different kind than Zao: ahead of time Samsung trained a deep learning algorithm on a huge volume of videos to learn how human faces naturally move; then this general knowledge is applied to a user-provided photo to animate it in a lifelike manner according to a video of a digital

[13]David Mack, "This PSA About Fake News From Barack Obama Is Not What It Appears," *BuzzFeed News*, April 17, 2018: https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed.

[14]The techniques for doing the audio portion of a deepfake are similar to the video ones, except recurrent neural networks are typically used instead of convolutional neural networks, for those who know what that means.

[15]Joan Solsman, "Samsung deepfake AI could fabricate a video of you from a single profile pic," *CNET*, May 24, 2019: https://www.cnet.com/news/samsung-ai-deepfake-can-fabricate-a-video-of-you-from-a-single-photo-mona-lisa-cheapfake-dumbfake/.

puppeteer that the user also provides to the program. These Samsung videos tend to retain more semblance to the puppeteer than deepfake methods that use training footage specific to the puppet.

A collaboration between researchers in academia and at Adobe created software[16] powered by deep learning to "let users edit the text transcript of a video to add, delete, or change the words coming right out of somebody's mouth." (And you thought Adobe's Photoshop was already impressive enough.) Their system first identifies the *phonemes* (the basic units of sound in spoken speech) in the video, then it matches these with the accompanying *visemes* (facial expressions and movements of the mouth). The software also learns a three-dimensional model of the lower half of the subject's face from the video. When the user edits the text of the transcript, the software replaces each phoneme and corresponding viseme in the 3D model and then uses this to modify the original video. This system currently only works for "talking head" style video and requires forty minutes of input data, and the results are better when the text does not differ too much from the original transcript. Nonetheless, it is an interesting approach for creating yet another type of deepfake.

You've now seen that there are many different types of video editing procedures that fall under the umbrella term of deepfake—and for each type, there are a variety of apps and deep learning architectures that have been used, most involving autoencoders and GANs. You've seen that deepfakes have been used for entertainment (pornographic and otherwise) and to illustrate what's currently possible. And you've seen that shallowfakes and other simple forms of visual deception have a long history and are still sowing confusion in politics today. You might be wondering at this point whether deepfakes have also been used in the real world, the way shallowfakes have, to manipulate the public's interpretation of political events and possibly even influence the outcome of democratic elections. I hope you are indeed wondering this, as it is the topic I shall turn to now.

Deepfakes in Politics

In May 2018, a social democratic party in Belgium posted to Twitter and Facebook a one-minute deepfake video of President Trump speaking in English with Dutch subtitles.[17] The quality of both the vocal impersonation and the

[16]James Vincent, "AI deepfakes are now as simple as typing whatever you want your subject to say," *The Verge*, June 10, 2019: https://www.theverge.com/2019/6/10/18659432/deepfake-ai-fakes-tech-edit-video-by-typing-new-words.

[17]Jane Lytvynenko, "A Belgian Political Party Is Circulating A Trump Deepfake Video," *BuzzFeed News*, May 20, 2018: https://www.buzzfeednews.com/article/janelytvynenko/a-belgian-political-party-just-published-a-deepfake-video.

deepfake visuals is rather poor, and the spoken content itself makes it quite obvious that this is a satirical caricature of Trump (think SNL sketch more than subtle subterfuge). The video opens with the following proclamation by Trump: "Dear people of Belgium, this is a huge deal. As you know, I had the balls to withdraw from the Paris climate agreement, and so should you." Near the end of the clip, the audio suddenly goes almost mute and the now-faint voice continues: "We all know climate change is fake, just like this video." In addition to the audio being barely audible for this sneaky admission, these words are the only ones in the clip without Dutch subtitles. The video is clearly a joke, but it was made to have a real chance of tricking an inattentive viewer. Based on the social media comments—where it reached twenty thousand views within a single day—a significant number of viewers were indeed fooled into thinking that it was an authentic clip of Trump.

Many—but not all, as you will soon see—of the deepfakes appearing in a political context are, like this Belgian example, more about entertainment than deception. The creators of the popular TV cartoon *South Park* in October 2020, just one week before the presidential election, released a web series on YouTube[18] that is based entirely on deepfakes. The main character in the show is a deepfake version of Donald Trump, and a deepfake Mark Zuckerberg has a repeated cameo. The running joke of the opening episode is that all the supposedly real footage is actually deepfake, while the clips in the show that are supposedly deepfake are either silly puppets or actual real people.

The 2020 Christmas address by Queen Elizabeth on the BBC was accompanied by a satirical address on the British public broadcast Channel 4 by a deepfake version of the queen.[19] If her stinging jokes about the royal family were not enough to make it clear that this was not really the queen, then the poorly voiced impersonation and the implausibly youthful dance routine she breaks into would surely be enough to settle the issue. Sometimes, however, the line between humor and politics is a bit more confusing. On April 26, 2020, the day of his wife's 50th birthday, President Trump tweeted[20] a low-quality deepfake video of Joe Biden sticking his tongue out and added the following caption: "Sloppy Joe is trending. I wonder if it's because of this. You can tell it's a deep fake because Jill Biden isn't covering for him." Certainly, no outright deception was intended there, but it was still an uncomfortable moment for society to see a president known for having chronic issues with the truth

[18] https://www.youtube.com/channel/UCi38HMIvRpGgMJOTlm1WYdw.

[19] Rhett Jones, "First Deepfake Address from the Queen of England Makes Its Debut on British TV," *Gizmodo*, December 25, 2020: https://gizmodo.com/first-deepfake-address-from-the-queen-of-england-makes-1845948622.

[20] David Frum, "The Very Real Threat of Trump's Deepfake," *The Atlantic*, April 27, 2020: https://www.theatlantic.com/ideas/archive/2020/04/trumps-first-deepfake/610750/.

openly share a tasteless deepfake video of his rival on social media and use it to make an immature dig at him.

On September 20, 2020, two ads were scheduled[21] to air on Fox, CNN, and MSNBC in the DC region, one featuring a deepfake Vladimir Putin and the other featuring a deepfake Kim Jong-un. Both had the same message: America doesn't need electoral interference because it will ruin its democracy all by itself. These ads were sponsored by a voting rights group and aimed to raise awareness of the fragility of American democracy and the need for Americans to actively and securely engage in the electoral process. The use of deepfakes here was not for deception, it was just to grab the viewers' attention and startle people into recognizing the technologically fraught environment in which the 2020 presidential election was to take place. The deepfakes were face-swaps created using open source DeepFaceLab software. Both ads included the following disclaimer at the end: "The footage is not real, but the threat is." At the last minute, the TV stations all pulled the ads and didn't immediately provide an explanation for this decision. One can surely imagine a natural hesitation about wading into these delicate deepfake waters. In the end, the ads only appeared on social media.

Arguably, the first direct use of deepfake technology in a political election occurred in India in February 2020. One day before the Legislative Assembly elections in Delhi, two forty-four-second videos of Manoj Tiwari, the leader of the Bharatiya Janata Party (BJP), were distributed across nearly six thousand WhatsApp groups, reaching roughly fifteen million people. In both videos, Tiwari criticized the rival incumbent political leader. In one video, he spoke in English, while in the other video he spoke a Hindi dialect called Haryanvi. Both videos were, in a sense, deepfakes. Tiwari first recorded the video in Hindi, his native tongue. Then, in partnership with a political communications firm called The Ideaz Factory, an impersonator recorded the audio for the English and Haryanvi versions of the speech. Finally, a "lip-syncing" form of reenactment deepfake that had been trained on other footage of Tiwari speaking was used to match his lip movements to the new audio.

Despite using deepfake technology, these Tiwari videos were not particularly malicious—they were simply tools used in a political campaign to reach a broader and more linguistically diverse audience. In the words of one of the BJP's heads of media and IT: "Deepfake technology has helped us scale campaign efforts like never before. The Haryanvi videos let us convincingly approach the target audience even if the candidate didn't speak the language of the voter." But the disingenuous nature of literally seeing someone speak a language they don't actually speak left some people with a bitter taste and

[21] Karen Hao, "Deepfake Putin is here to warn Americans about their self-inflicted doom," *MIT Technology Review*, September 29, 2020: https://www.technologyreview.com/2020/09/29/1009098/ai-deepfake-putin-kim-jong-un-us-election/.

slight feeling of political dishonesty. As reported[22] in *Vice*, Tiwari's Haryanvi video "was used widely to dissuade the large Haryanvi-speaking migrant worker population in Delhi from voting for the rival political party." Honestly, it's hard to know how to feel about this instance of political deepfakery. Like most powerful tools, there are good applications of deepfake technology and bad applications and everything in between—and it will take some time for the full range of applications to emerge. Since this book is on fake news, I will focus mostly on the nefarious, deceptive uses, but later in this chapter, I will give one purely positive, legitimate use in politics.

India was also the site of a much more unequivocally repugnant usage of deepfakes that occurred two years earlier—and while it is not directly related to an election, it still has strong political undercurrents and ramifications. Rana Ayyub was a thirty-six-year-old Indian woman, an investigative journalist, and a practicing Muslim. She said[23] she was often seen as anti-establishment and that she has been called "the most abused woman in India." She explained that anything she posted on Twitter would result in thousands of replies, much of it hateful and threatening. She tried to ignore the trolls and continue going about her job, telling herself that the online hate and threats "would never translate into offline abuse." But in April 2018, that changed.

An eight-year-old Kashmiri girl had been raped, leading to widespread outrage across the country. The BJP (yes, the same one just discussed above) was the ruling political party at the time and responded by organizing a reactionary march in support of those accused of perpetrating this heinous act. Ayyub was invited to speak on the BBC and Al Jazeera about "how India was bringing shame on itself by protecting child sex abusers." Shortly afterward, a male contact in the BJP sent Ayyub an ominous message: "Something is circulating around WhatsApp, I'm going to send it to you but promise me you won't feel upset." What she then saw was a pornographic movie in which she appeared to be the star. The video was a face-swap deepfake. In Ayyub's own words: "When I first opened it, I was shocked to see my face, but I could tell it wasn't actually me because, for one, I have curly hair and the woman had straight hair. […] I started throwing up."

The video was circulating in private political channels on WhatsApp, but then the fanpage of BJP's leader posted it publicly, and it quickly tallied more than forty thousand shares. Ayyub says she started getting WhatsApp messages from strangers requesting her services as a prostitute. Her anxiety over the situation became so severe that she went to the hospital with vomiting and

[22]Nilesh Christopher, "We've Just Seen the First Use of Deepfakes in an Indian Election Campaign," *Vice*, February 18, 2020: https://www.vice.com/en/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp.

[23]Rana Ayyub, "I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me," *Huffington Post*, November 21, 2018: https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316.

heart palpitations: "The entire country was watching a porn video that claimed to be me." Ayyub says that, ironically, just a week before this incident, one of her editors mentioned the potential dangers of deepfakes in India—she didn't know what they were so Googled them but decided against doing a story on them because she didn't want to bring more attention to them that might inspire any malicious use. "Then one week later it happened to me. […] It is a very, very dangerous tool and I don't know where we're headed with it." Devastating and alarming. I don't know what else to say. And one of the world's foremost digital forensics experts, Hany Farid, said[24] that a handful of politicians from developing countries around the world have asked him to try to debunk videos appearing to show them in compromising sexual situations.

The next example of deepfakery impacting politics in the real world is a truly bizarre story.[25] Ali Bongo, the president of the African nation Gabon, was hospitalized in Riyadh, the capital of Saudi Arabia, for an undisclosed illness in October 2018. In December, the vice president announced that Bongo had suffered a stroke earlier in the fall but is doing well and recovering in Rabat, the capital of Morocco. Despite this vague reassurance, there were almost no signs of Bongo for over two months aside from a few pictures and a silent video released by the government. Speculation began to run rampant that the officials were lying, that Bongo had either died or at least was incapacitated and in far worse condition than was publicly admitted. This is a country that was ruled by Ali Bongo since 2009, and before that his father Omar Bongo ruled for forty-two years. Most people literally did not remember a time in Gabon when the head of the government was not named Bongo. There was little trust in the official explanation for why Ali had been out of the country for over two months and essentially out of sight the entire time.

Finally, to help quell the growing suspicion, the president's advisors said he would deliver the customary New Year's address. And indeed, on January 1, 2019, the government posted to social media a video of President Ali Bongo giving his speech. But something about it didn't seem right. Some viewers were reassured of the president's health by the video, but others thought it was perhaps a body double impersonating him. And many other viewers felt the video's strangeness was the result of something else, but they couldn't quite put their finger on it.

Then Bruno Ben Moubamba, a prominent Gabonese politician who ran against Bongo in the previous two elections, claimed the video was a deepfake—and his theory rapidly gained a sizable following. Moubamba pointed out that in the video Bongo's face and eyes seem "almost suspended above his jaw" and that his eyes move "completely out of sync with the movements of his jaw."

[24]Ali Breland, "The Bizarre and Terrifying Case of the 'Deepfake' Video that Helped Bring an African Nation to the Brink," *Mother Jones*, March 15, 2019: https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

[25]See Footnote 24.

Moubamba explained: "The composition of several elements of different faces, fused into one are the very specific elements that constitute a deepfake." Other activists and critics of the president took to Twitter to point out more elements that suggested a possible deepfake. For instance, they noted that Bongo only blinked thirteen times during the two-minute video, less than half the typical amount, and his speech patterns seemed to differ from his usual ones.

People didn't know what to believe; the video raised more questions than it answered. And Bongo's critics observed that there was a very real and specific reason why the government might be trying to cover up Bongo's death or ill-health: the constitution states that if Gabon's president is ever found to be unfit to lead, then the Senate President becomes the interim president and a special election is to be held within sixty days. Gabon's ruling party, critics argued, was deceiving the public in order to avoid this special election, perhaps to buy time until it could shore up support for a successor—a successor that would, after all, be the country's first president not from the Bongo family in more than half a century.

One week after the release of the enigmatic New Year's video, Gabon's military attempted a coup—the country's first since 1964—and explicitly cited the oddness of the video as evidence that the president was absent and that the government was lying about it. The coup ended up failing, and the government retained control. To this day, digital forensics experts are uncertain whether the video is a deepfake, though they generally lean toward the conclusion that it is. But many of the signs they point to could also be caused by Bongo's stroke. All we know for sure is that in August 2019 Bongo made his first public appearance since the stroke—and that deepfake technology, and the shadow of doubt it casts on the veracity of videos, nearly led to a military overthrowal of a national government.

In a striking parallel, something briefer but eerily similar happened in the United States just one year later. Friday, October 2, 2020, was one of the strangest and most confusing days in recent memory (and for 2020, that's saying a lot). News broke in the morning that President Trump had tested positive for COVID-19, and in a matter of hours we found out that he wasn't just positive, he was symptomatic—and then, that his condition was actually quite serious, he was going to be hospitalized. All the day's events were shrouded in a veil of uncertainty and chaos largely caused by the lack of frank and transparent communication from the government. It was literally just weeks before one of the most important elections in American history, yet we did not know the true state of the president's health, and suspicion quickly grew that things were much worse than the officials were telling us.

Then, at 6:31 p.m. that day, President Trump posted on Twitter an eighteen-second video address in which he said that he is heading to Walter Reed, but he reassured people that he thinks he is doing very well. The video looked

strange. Very strange. Immediately there was talk on social media of it being a deepfake.[26] This time the deepfake conspiracy faded quickly as more footage of Trump was soon seen, culminating most convincingly with a live address a few days later when he was released from Water Reed. But for a brief moment, it really was hard to know what was going on and what to believe; it did not seem entirely implausible for the US government, just one month before the election, to create a deepfake to cover up the dire state of the president's health.

In March 2021, a military-run TV station in Myanmar broadcast a video recording of a detained former regional chief minister providing a public confession. He said he bribed Aung San Suu Kyi, the Nobel Peace Prize laureate who in the 2010s played a key role in transitioning Myanmar from military rule to partial democracy but then was arrested and deposed—along with other members of her ruling political party—by the military in a coup on February 1, 2021. In other words, the military was presenting the incriminating evidence it needed to help justify its actions. But there was immediate outcry that this confession is fake—the voice doesn't sound like his usual one, and the visuals look strange—and many people suspect it is a deepfake, but once again, we don't know for sure.[27]

Let me turn now to one final example of real-world deepfakes in a political setting—this time showing a positive use of the technology. In July 2020, David France, an Oscar-nominated activist filmmaker, debuted on HBO a documentary called *Welcome to Chechnya* about the anti-LGBTQ purges that took place in Chechnya. He wanted to include interviews with survivors of these atrocities, but he knew that for their personal safety their identities must be concealed in the film—at the time, they were being hunted in their homeland and escaping the region through a network of safe houses. He felt the usual documentarian technique of blurring faces produced too much of an emotional disconnect between the speaker and the audience, so he instead used deepfake technology. The production team filmed individuals outside Chechnya, unrelated to the country's purge, in a studio equipped with an array of cameras capturing their faces from many angles; then deep learning algorithms were used to blend these faces onto the faces of twenty-three Chechens in the film to provide them with new disguised faces—and hence anonymity. As reported[28] in the *New York Times*, "In one of the film's more

[26]Tyler MacDonald, "Producers Speculate That Donald Trump's Post-Coronavirus Video Is A Deepfake," *Inquisitr,* October 2, 2020: https://www.inquisitr.com/6312717/producers-trump-coronavirus-video-deepfake/.

[27]"Is this guy for real? In Myanmar, the fear of deepfakes may be just as dangerous." *Coconuts*, March 24, 2021: https://coconuts.co/yangon/news/is-this-guy-for-real-in-myanmar-the-fear-of-deepfakes-may-be-just-as-dangerous/.

[28]Joshua Rothkopf, "Deepfake Technology Enters the Documentary World," *The New York Times*, July 1, 2020: https://www.nytimes.com/2020/07/01/movies/deepfakes-documentary-welcome-to-chechnya.html.

breathtaking moments, the effects drop away after a gay refugee, Maksim Lapunov, reclaims his name—and his real face—at a news conference. 'I wanted you to feel what he felt at that moment,' France said."

While convincing deepfakes still seem rare in the real world, both their usage and their quality have been accelerating swiftly. On March 10, 2021, the FBI issued an official alert[29] boldly stating that "Malicious actors almost certainly will leverage synthetic content for cyber and foreign influence operations in the next 12-18 months," and the alert specifies that deepfakes are the main form of synthetic content it is referring to here. If we develop tools for determining when videos are deepfakes, we could push back against these malicious efforts and also apply these tools the next time something like the Ali Bongo New Year's address or Donald Trump Walter Reed clip or Myanmar confession arises. It is time now to look at the progress and challenges in developing such tools.

Detecting Deepfakes

If AI has the ability to create deepfakes, shouldn't it also be able to detect them? Yes and no. One broad challenge is that there are many different types of deepfake video manipulations—you have already seen a handful in this chapter, and surely more will keep coming out each year—so an algorithm trying to decide if a video is a deepfake cannot just look for one specific type of manipulation. There are significant challenges at the technical level too. Recall that most deepfake creation algorithms rely on the GAN architecture where the generator learns how to synthesize deepfakes and the discriminator learns how to distinguish real video clips from the synthetic ones. Any AI system for detecting deepfakes will in essence be playing the role of that discriminator—but the whole point in a GAN is that through the training procedure the generator learns how to fool the discriminator. In other words, the very process by which deepfakes are constructed makes them difficult for algorithms to detect.

An additional challenge—and this is probably the biggest one—is that deepfake detection is an arms race: as deepfake-detecting technology improves, the ability of deepfake creators to avoid detection will also improve. For instance, in the early days of deepfakes (which is to say, a couple years ago), some researchers noticed that people blinked at a lower rate in deepfake videos than in real life and used this observation as the basis for a detection

[29]Shannon Vavra, "FBI alert warns of Russian, Chinese use of deepfake content," *CyberScoop*, March 10, 2021: https://www.cyberscoop.com/fbi-foreign-actors-deepfakes-cyber-influence-operations/.

algorithm,[30] but it did not take long for deepfake creation algorithms to overcome this weakness and render this particular detection algorithm obsolete. A related but more recent approach[31] that currently looks promising is to measure heartbeat rhythms and blood flow circulation, but it is only a matter of time before the deepfake creators learn how to get past this hurdle as well. That said, just because deepfake detection algorithms have a difficult task does not mean we shouldn't bother trying; quite the opposite, it means we must move quickly and vigorously to stay on top of this ever-evolving challenge.

In spring 2020, the popular data science competition site Kaggle hosted the "Deepfake Detection Challenge," a public competition with one million dollars in prize money to see who could produce the most accurate algorithm for classifying videos as deepfake versus authentic. The organizers for this competition—who provided both the training data that the competitors used to build and tweak their algorithms and the testing data that was used behind the scenes afterward to evaluate and rank the performance of the entrants—included Facebook, Amazon, Microsoft, a group of academics, and a coalition of media and technology experts called the "Partnership on AI's Media Integrity Steering Committee." Unsurprisingly, all the winning teams relied on deep learning architectures for their algorithms.

The top performer in this Kaggle competition managed an accuracy rate of 82.5% on the testing data. If automated methods like this are relied upon in practice, many deepfake videos will slip through the radar and many authentic videos will be mislabeled as deepfakes. Also, there is a big difference between performance in a simulated contest like this at a fixed moment in time and performance in the real world where the technology driving deepfakes is constantly changing. In fact, when tested on a new unseen data set, the winning entrant's accuracy rate dropped precipitously to 65%; to be blunt, that's not a heck of a lot better than random guessing. Moreover, this competition focused on purely algorithmically mass-produced deepfakes, so if one wanted to fool a detection algorithm in a particular instance, one could do additional manual processing to throw a monkey wrench in the works.

---

[30] You'll remember that infrequent blinking was one of the oddities about Ali Bongo's New Year's address video that led people to think it was a deepfake. Interestingly, it turns out the reason many deepfakes tended to have infrequent blinking was in part because people are seldom photographed with their eyes closed, so any algorithm that included photos (not just videos) in its training data would falsely "learn" that people spend more time with their eyes open than they actually do. See Siwei Lyu, "Detecting 'deepfake' videos in the blink of an eye," *The Conversation*, August 29, 2018: https://theconversation.com/detecting-deepfake-videos-in-the-blink-of-an-eye-101072.

[31] Khari Johnson, "AI researchers use heartbeat detection to identify deepfake videos," *VentureBeat*, September 3, 2020: https://venturebeat.com/2020/09/03/ai-researchers-use-heartbeat-detection-to-identify-deepfake-videos/.

As noted[32] in *Scientific American*, "Such 'crafted' deepfake videos are more likely to cause real damage, and careful manual post processing can reduce or remove artifacts that the detection algorithms are predicated on."

The top five teams in the competition (the ones receiving prize money) all had something in common beyond their use of deep learning: they all used a specific deep learning architecture called *EfficientNets* developed by Google in 2019 that is known to be good at recognizing faces and other objects in images. The winner of the competition, Selim Seferbekov, thinks[33] the next step for algorithmic detection might involve a focus on the transition between frames in the video: "Even very high-quality deepfakes have some flickering between frames," he pointed out. While these flickers are not hard for humans to spot with the naked eye, Seferbekov says he tried to capture them with his algorithm but found it too computationally intensive so gave up for now. Shortly after the Kaggle competition concluded, Facebook released a public database of more than one hundred thousand video clips produced using over three thousand actors and a variety of known face-swap deepfake techniques hoping that this will help the research community develop better detection methods.

The US Government has a significant investment in deepfake detection: in 2015, the Defense Advanced Research Projects Agency (DARPA)—a research organization within the Department of Defense focusing on emerging technologies for use by the military—launched a program called Media Forensics, or more briefly *MediFor*. The creation of this program was curiously time. Shortly before, a news channel in Russia had broadcast supposed satellite imagery of a Ukrainian fighter jet shooting at Malaysia Airlines Flight 17. It turned out these images were fake, though they were made with more traditional methods rather than deep learning; it also turned out the flight was downed by a Russian missile. This Russian incident likely put fake imagery and videos high on DARPA's list, and as deepfake technology developed over the following several years, DARPA had good reason to maintain a keen interest in the topic.

It was reported[34] by *Scientific American* in 2018 that MediFor had three broad approaches to its task, all of which are strong candidates for automation through deep learning: "The first examines a video's digital fingerprint for anomalies. The second ensures a video follows the laws of physics, such as

[32]Siwei Lyu, "Deepfakes and the New AI-Generated Fake Media Creation-Detection Arms Race," *Scientific American*, July 20, 2020: https://www.scientificamerican.com/article/detecting-deepfakes1/.

[33]Will Douglas Heaven, "Facebook just released a database of 100,000 deepfakes to teach AI how to spot them," *MIT Technology Review*, June 12, 2020: https://www.technologyreview.com/2020/06/12/1003475/facebooks-deepfake-detection-challenge-neural-network-ai/.

[34]See Footnote 5.

sunlight falling the way it would in the real world. And the third checks for external data, such as the weather on the day it was allegedly filmed." One of the researchers involved in this project insightfully summarized the context of their work: "We will not win this game, it's just that we will make it harder and harder for the bad guys to play it."

A different approach in the war against visual disinformation is to authenticate photos and videos either by embedding digital watermarks in them (taking inspiration from old-school ways of stopping counterfeiters) or by creating databases that can be used to refute modified versions that show up later. For instance, a San Diego startup called *Truepic* offers a smartphone app that lets users take photos or videos that are authenticated as undoctored. It does this by sending the photo/video along with various sensor readings recorded by the camera to Truepic's servers where a variety of tests are undertaken, and if the tests are all passed, then the photo/video is considered "verified" and is stored on the server. The full set of tests is not disclosed, but the CEO of Truepic explained[35] that they "look at geolocation data, at the nearby cell towers, at the barometric-pressure sensor on the phone, and verify that everything matches. We run the photo through a bunch of computer-vision tests." The app's biggest clients so far are insurance companies, since it allows policyholders to take photos of accidents and damages that the company can be sure have not been doctored, but Truepic says it has also been used by NGOs to document human rights violations.

A startup in the UK called *Serelay* developed an app that is similar to Truepic's, except Serelay's app does not store the full photo in its server, it only stores a small digital fingerprint of the photo obtained by computing about a hundred mathematical values for each image. One cannot reconstruct the full photo from this fingerprint, but the company claims[36] that if even a single pixel in the photo has been modified, then the fingerprints will not match up. Of course, both the Truepic and Serelay services only work if one knows in advance that the validity of a particular photo might later be questioned—so while very useful in some realms, they do not address the ocean of questionable photos flowing through the rapid channels of social media every day. That said, one can envision a world in the not-too-distant future in which every smartphone by default uses a verification service like this, and then whenever someone posts a photo or video on a social media platform, the platform places a little check mark beside it if it passes the verification service.

---

[35]Joshua Rothman, "In The Age of A.I., Is Seeing Still Believing?" *New Yorker*, November 5, 2018: https://www.newyorker.com/magazine/2018/11/12/in-the-age-of-ai-is-seeing-still-believing.

[36]Karen Hao, "Deepfake-busting apps can spot even a single pixel out of place," *MIT Technology Review*, November 1, 2018: https://www.technologyreview.com/2018/11/01/139227/deepfake-busting-apps-can-spot-even-a-single-pixel-out-of-place/.

In addition to all the technical challenges with automated deepfake detection, there is a very significant sociological and psychological challenge as well that Brooke Borel at *Scientific American* calls[37] "the lag between lies and truth": even if a viral video is proven to be a deepfake, often the damage caused by the deception will already have been done and is effectively irreversible. Once people are convinced of something, especially if they have seen it with their own eyes, it can be very difficult to disabuse them of it even if irrefutable evidence to the contrary has subsequently surfaced. This suggests that in addition to unmasking harmful deepfakes after they have gone viral, it may well be prudent to prevent them from spreading in the first place. This takes us to our next topic, which is legislative approaches to limiting the damage that deepfakes can do.

Legal Regulation

Senator Marco Rubio from Florida has spoken multiple times about the threats posed by deepfake technology and encouraged legislative action. Senator Ben Sasse from Nebraska in December 2018 introduced a bill aimed at regulating deepfakes—the first of its kind—but a day later, the federal government shut down over a budgetary impasse, and Sasse's proposed bill expired by the time the government reopened. Next, in parallel to the June 2019 House hearing on deepfakes that opened this chapter, a Representative for New York's ninth congressional district, Yvette Clarke, introduced a different bill on deepfakes, more extensive than Sasse's.

Clarke's bill—drafted in collaboration with computer scientists, disinformation experts, and human rights advocates—would require social media companies to better monitor their platforms for deepfakes and researchers to develop digital watermarking tools for deepfakes, and it would criminalize the malicious use of deepfakes that harm individuals or threaten national security. One of the advisers on the bill, Mutale Nkonde, a fellow at the Data & Society Research Institute, said[38] the bill was unlikely to pass through Congress in its original form but felt it important to introduce the bill regardless in order to make the first serious step toward legislative regulation of deepfakes: "What we're really looking to do is enter into the congressional record the idea of audiovisual manipulation being unacceptable." While the bill did indeed stall,

[37]See Footnote 5.

[38]Karen Hao, "Deepfakes have got Congress panicking. This is what it needs to do." *MIT Technology Review*, June 12, 2019: https://www.technologyreview.com/2019/06/12/134977/deepfakes-ai-congress-politics-election-facebook-social/.

in February 2021 Clarke said[39] she's planning to reintroduce a revised version of the bill that she felt would gain more traction due to the new political environment after the 2020 election and the fact that the pandemic has led to an increase in social media usage: "the conditions [are] ripe for actually passing some meaningful deepfake legislation."

While regulation at the federal level has stalled in Congress so far, at the state level there have been some interesting developments. In October 2019, California signed a law[40] making it a crime to maliciously distribute or create "materially deceptive" media about a political candidate within sixty days of an election. (A doctored photo or video is considered deceptive if a "reasonable person" would have a "fundamentally different understanding or impression" of it compared to the original version.) The term deepfake does not appear in the text of this law, but the law has been nicknamed the "California Deepfake Law," and indeed it is directly inspired by deepfakes and the threat they pose to the state's democratic systems.

California's law provides some exceptions, such as satire and videos with disclaimers stating that they are fake, but free speech advocates voiced objections to it and question its constitutionality (the specific focus on elections and the sixty-day window are an attempt to help assuage such concerns). Simultaneously, California also enacted a law banning nonconsensual pornographic deepfakes. And one month earlier, Texas became the first state to legislate deepfakes by criminalizing them when they are used "with intent to influence the outcome of an election." However, legal scholars have pointed out that deepfakes can be very dangerous outside of elections as well (for instance, when videos are used as evidence in court), and, moreover, in order to enforce any of these deepfake laws, one needs to be able to prove that a video in question really is a deepfake—which, as you now know, is no easy task.

Even if the legislative branch of the federal government has been apprehensive to tackle the challenge of deepfakes, it has—at least on paper—made an

[39]Karen Hao, "Deepfake porn is ruining women's lives. Now the law may finally ban it." *MIT Technology Review*, February 12, 2021: https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/.

[40]Amre Metwally, "Manipulated Media: Examining California's Deepfake Bill," *Jolt Digest*, November 12, 2019: http://jolt.law.harvard.edu/digest/manipulated-media-examining-californias-deepfake-bill.

effort to police itself. On January 28, 2020, the US House Ethics Committee released an official memo[41] titled "Intentional Use of Audio-Visual Distortions & Deep Fakes" that includes the following text:

> Members or their staff posting deep fakes could erode public trust, affect public discourse, or sway an election. Accordingly, Members, officers, and employees posting deep fakes or other audio-visual distortions intended to mislead the public may be in violation of the Code of Official Conduct. Prior to disseminating any image, video, or audio file by electronic means, including social media, Members and staff are expected to take reasonable efforts to consider whether such representations are deep fakes or are intentionally distorted to mislead the public.

How strictly this Code of Official Conduct is adhered to remains to be seen.

Dismissing Valid Evidence

In the tense months leading up to the 2020 election, a congressional candidate running for a House seat in Missouri wrote and shared a twenty-three-page document titled "George Floyd is Dead: A Citizens' Investigative Report on the Use of Deep Fake Technology." In it, she argued that the viral video showing the police murder of George Floyd that sparked national outrage was actually a deepfake. She claimed the person seen in the video was an actor with the countenance of Floyd (who supposedly died in 2017) face-swapped in. She said[42] the video was a false flag operation intended to "stoke racial tensions between Black and white Americans" and reinvigorate the "flailing radical Black Lives Matter movement." It was an absurd and unfounded conspiracy theory. Thankfully, she lost in the primary. But she was not alone in using the mere existence of deepfake technology in attempts to distort the public's understanding of reality.

A number of people have argued that the biggest threat from deepfakes is not the direct deception they are capable of—it is the general erosion of trust they lead to in society and the cover they provide to nefarious individuals now to plausibly deny damning videographic evidence by simply crying deepfake.

---

[41] https://ethics.house.gov/campaign-activity-pink-sheets/intentional-use-audio-visual-distortions-deep-fakes.

[42] Daniel Villareal, "GOP Candidate Says George Floyd Video Fake, That TV Host Portrayed Chauvin," *Newsweek*, June 25, 2020: https://www.newsweek.com/gop-candidate-says-george-floyd-video-fake-that-tv-host-portrayed-chauvin-1513282.

Further into his opening remarks from the June 2019 hearing that began this chapter, Adam Schiff presciently warned that "not only may fake videos be passed off as real, but real information can be passed off as fake. This is called the liar's dividend, in which people with a propensity to deceive are given the benefit of an environment in which it is increasingly difficult for the public to determine what is true."

As you surely remember, just one month before the 2016 presidential election, the *Washington Post* published an article accompanied by the now-notorious "*Access Hollywood* tape" from 2005 in which Donald Trump makes extremely lewd comments about women in off-camera audio that was recorded presumably without his knowledge. This story broke just two days before one of the presidential debates, and Trump responded by admitting he made the remarks caught on tape and apologized for them but also attempted to minimize their significance as "locker room banter." One year later, Trump quite bizarrely and brazenly started claiming[43] the audio on that tape was fake and that he didn't say the words we heard.

Responding to this assertion in a CNN interview with Anderson Cooper, the soap opera actress Arianne Zucker who was the subject of some of Trump's vulgar comments in the *Access Hollywood* tape had this to say: "I don't know how else that could be fake, I mean, unless someone's planting words in your mouth." *Access Hollywood* responded as well: "Let us make this perfectly clear, the tape is very real. He said every one of those words." Nonetheless, Trump reportedly said[44] in multiple private conversations that he's not sure if it was really him in the tape, and in January 2017 he told a senator he was "looking into hiring people to ascertain whether or not it was his voice."

Perhaps deepfake technology in 2021 and beyond finally provides the cover Trump sought in 2017 to explain the damaging recording from 2005 that surfaced during his campaign in 2016. But the same technology that might have allowed him to discount the *Access Hollywood* tape is what led some people not to believe the authentic recorded reassurances about his health during his battle with COVID-19.

---

[43]Jonathan Martin, Maggie Haberman, and Alexander Burns, "Why Trump Stands by Roy Moore, Even as It Fractures His Party," *New York Times*, November 25, 2017: https://www.nytimes.com/2017/11/25/us/politics/trump-roy-moore-mcconnell-alabama-senate.html.

[44]Emily Stewart, "Trump has started suggesting the Access Hollywood tape is fake. It's not." November 28, 2017: https://www.vox.com/policy-and-politics/2017/11/28/16710130/trump-says-access-hollywood-tape-fake.

Summary

Deepfake video editing is a wide range of methods for modifying video clips to change the words people say and the people who say them. It is powered by deep learning, most commonly the GAN architecture in which two algorithms are pit against each other and through the data-crunching training process the generator learns to routinely fool the discriminator. This technology first appeared in 2017 when it was used to make nonconsensual pornography, and it now threatens society's ability to discern the truth. Conspiracy theorists call legitimate videographic evidence (such as George Floyd's murder by the police) into question by claiming it is deepfake, and corrupt politicians are now granted a powerful tool: they can dismiss incriminating clips as deepfake. Meanwhile, innocent journalists and politicians have had their reputations tarnished when their faces were deepfake swapped into sexual clips. Algorithmically detecting deepfakes has proven challenging, though there is sustained effort in that realm and some glimmers of hope. Legislative attempts to limit the spread of deepfakes by regulating their usage have so far stalled at the national level; at the local level, there has been some concrete action, but the impingement of free speech they necessitate leaves their constitutionality in question. This chapter was all about the algorithms used to edit videos; in the next chapter, I turn to another algorithmic aspect of videos: YouTube recommendations.

CHAPTER

4

Autoplay the Autocrats

The Algorithm and Politics of YouTube Recommendations

> *As far-right and conspiracy channels began citing one another, YouTube's recommendation system learned to string their videos together. However implausible any individual rumor might be on its own, joined together, they created the impression that dozens of disparate sources were revealing the same terrifying truth.*
>
> —Max Fisher and Amanda Taub, *New York Times*

As trust in traditional media outlets has declined, people have turned to alternative sources to get their news. One particularly popular platform in this regard, especially among the younger generations (as you'll soon see in this chapter with some precise facts and figures), is YouTube. The premise that anyone can post videos showing or explaining what is happening in the world is appealing, but the reality is that YouTube has played an alarming role in the spread of fake news and disinformation. The powerful yet mysterious YouTube

© Noah Giansiracusa 2021
N. Giansiracusa, *How Algorithms Create and Prevent Fake News*,
https://doi.org/10.1007/978-1-4842-7155-1_4

recommendation algorithm drives the majority of watch time on the site, so understanding how it works is crucial to understanding how YouTube has pushed viewers toward outlandish conspiracy theories and dangerous alt-right provocateurs. This chapter takes a close look at how the recommendation algorithm has developed over the years, how it behaves in practice, how it may have influenced elections and political events around the world, how the company has responded to criticism, and how it has tried to moderate the content it hosts.

Growing Chorus of Concern

"Years ago, the openness of YouTube was a benefit to artists, activists, and creative types, but YouTube is now a major component of scaling disinformation campaigns." This was said by Joan Donovan, research director of the Shorenstein Center on Media, Politics, and Public Policy at Harvard, after a network of fake news YouTube channels supporting Trump's efforts to overturn the 2020 election appeared in the days after the election.[1]

"Less than a generation ago, the way voters viewed their politicians was largely shaped by tens of thousands of newspaper editors, journalists and TV executives. Today, the invisible codes behind the big technology platforms have become the new kingmakers." This was written by Paul Lewis of the *Guardian* in his investigation into how YouTube's algorithm distorts the truth.[2]

"For a short time on January 4, 2018, the most popular livestreamed video on YouTube was a broadcast dominated by white nationalists. […This video] is part of a larger phenomenon, in which YouTubers attempt to reach young audiences by broadcasting far-right ideas in the form of news and entertainment. […] One reason YouTube is so effective for circulating political ideas is because it is often ignored or underestimated in discourse on the rise of disinformation and far-right movements." This was written by Rebecca Lewis in a 2018 report on YouTube for the Data & Society Research Institute.[3]

"YouTube's powerful recommendation algorithm, which pushes its two billion monthly users to videos it thinks they will watch, has fueled the platform's ascent to become the new TV for many across the world. […] YouTube's

---

[1]Craig Silverman, "This Pro-Trump YouTube Network Sprang Up Just After He Lost," *BuzzFeed News*, January 8, 2021: https://www.buzzfeednews.com/article/craigsilverman/epoch-times-trump-you-tube.

[2]Paul Lewis, "'Fiction is outperforming reality': how YouTube's algorithm distorts truth," *Guardian*, February 2, 2018: https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth.

[3]Rebecca Lewis, "Alternative Influence: Broadcasting the Reactionary Right on YouTube," Data & Society Research Institute, September 18, 2018: https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf.

success has come with a dark side. Research has shown that the site's recommendations have systematically amplified divisive, sensationalist and clearly false videos." This was written by Jack Nicas of the *New York Times* in his investigation into how YouTube's algorithm encourages the spread of conspiracy theories.[4]

"YouTube is something that looks like reality, but it is distorted to make you spend more time online. The recommendation algorithm is not optimizing for what is truthful, or balanced, or healthy for democracy." This was said by Guillaume Chaslot, a former Google AI engineer who worked on YouTube's recommendation algorithm. "On YouTube, fiction is outperforming reality," Chaslot continued.[5]

"Bellingcat, an investigative news site, analyzed messages from far-right chat rooms and found that YouTube was cited as the most frequent cause of members' 'red-pilling'—an internet slang term for converting to far-right beliefs. A European research group, VOX-Pol, conducted a separate analysis of nearly 30,000 Twitter accounts affiliated with the alt-right. It found that the accounts linked to YouTube more often than to any other site." This was written by Kevin Roose of the *New York Times* in his investigation into how YouTube radicalizes people.[6]

"Reality is shaped by whatever message goes viral," said Pedro D'Eyrot, cofounder of the group that formed to agitate for the impeachment in 2016 of Brazil's left-wing then-president, Dilma Rousseff. "YouTube's auto-playing recommendations were my political education," said Mauricio Martins, an official in the political party of Brazil's authoritarian far-right president, Jair Bolsonaro.[7]

The main goal of this chapter is to get to the bottom of these unnerving quotes—to understand what YouTube's enigmatic recommendation algorithm does and to unpack the controversies surrounding it. In particular, I'll explore whether the YouTube recommendation algorithm really has contributed to the spread of fake news, driven the growth of deleterious conspiracy theories, and propped up autocrats and the alt-right, especially in Brazil and the United States.

---

[4]Jack Nicas, "Can YouTube Quiet Its Conspiracy Theorists?" *New York Times*, March 2, 2020: https://www.nytimes.com/interactive/2020/03/02/technology/youtube-conspiracy-theory.html.

[5]See Footnote 2.

[6]Kevin Roose, "The Making of a YouTube Radical," *New York Times*, June 8, 2019: https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html.

[7]Max Fisher and Amanda Taub, "How YouTube Radicalized Brazil," *New York Times*, August 11, 2019: https://www.nytimes.com/2019/08/11/world/americas/youtube-brazil.html.

Background on YouTube

YouTube launched in 2005 and was acquired by Google a year later for $1.65 billion. It has over two billion users across the globe.[8] That's almost one-third of the internet and more than the number of households that own televisions. One billion hours of YouTube videos are watched daily. More than five hundred hours of video are uploaded every minute. YouTube's traffic is estimated to be the second highest of any website, behind only Google.com.[9] In the United States, YouTube reaches more people between the ages of eighteen and thirty-four than any television network; ninety-four percent of Americans aged eighteen to twenty-four use YouTube, a higher percentage than for any other online service.

There are basically four different ways that people access YouTube videos:

- Embedded videos on other platforms
- Direct URL links to videos that people share
- Keyword searches on YouTube's homepage
- Recommended videos on YouTube's homepage and "up next" videos that are recommended whenever a video is playing

In 2018, it was revealed[10] by YouTube's Chief Product Officer that seventy percent of the total time users spend watching YouTube videos comes from this fourth category, the recommended videos.

The term *recommendation algorithm* refers to the behind-the-scenes systems powering both forms of recommended videos (the ones on the homepage and the "up next" list); occasionally, it is also used to refer to the direct search function on YouTube, since the user types keywords and the site returns a list of videos that it recommends as matches to the search, but I'll avoid conflating these rather different processes. Most of the public debate and discourse about YouTube's potential for political polarization focuses on the recommendation algorithm, and that will also be the focus of this chapter.

Company insiders say[11] that the recommendation algorithm is the single most important engine of YouTube's growth, and they describe it as "one of the largest scale and most sophisticated industrial recommendation systems in existence." This sense of scale and significance certainly creates the potential for YouTube's recommendation algorithm to have a tremendous impact on

[8] https://www.youtube.com/about/press/.
[9] See Footnote 6.
[10] Joan Solsman, "YouTube's AI is the puppet master over most of what you watch," *CNET*, January 10, 2018: https://www.cnet.com/news/youtube-ces-2018-neal-mohan/.
[11] See Footnote 2.

society worldwide, but to find out if and how it does so will take some digging. I shall start with a technically oriented chronology of the algorithm.

Development of the Algorithm

In the early days of YouTube, before the recommendation algorithm, videos were shared through embedding or direct links, and the YouTube site itself was primarily a repository where people would look up specific videos—for instance, a viral clip that was discussed at the office. Facebook changed the lay of the social media land when it introduced the newsfeed, an infinite stream of personalized content. This was an innovation that soon spread to other platforms, such as Tumblr, Twitter, Instagram, LinkedIn—and YouTube. For YouTube, this development shows up both on the homepage where videos in various categories are suggested to the user and in the up next videos that suggest what a user should watch after the current video concludes. Thanks to the autoplay feature that was added in 2015, the user doesn't even have to click anything in order to set sail down the river of algorithmic recommendations.

2012: From Views to Watch Time

You saw in Chapter 1 that online journalism has adopted the pageview as its primary currency: the single metric that determines ad revenue and defines success. In the early years of YouTube, the success of a video was similarly measured by the number of views it received, but there was a big problem with this: ads are dispersed throughout videos so users who leave videos early do not see all the ads. Two videos with the same number of views might generate very different amounts of ad revenue if one was getting users to watch longer and therefore see more ads. This suggests that the combined amount of time all users spend on a video (called *watch time*) is a better proxy for the value of a video than the number of views. And keep in mind it's not just content creators who earn money from ad revenue—YouTube's corporate profits are from ad revenue, so YouTube the company needs users to watch videos for as long as possible. Accordingly, in 2012, YouTube made a fundamental and lasting change to its recommendation algorithm: instead of aiming to maximizing views, it would aim to maximize watch time.

From a technical perspective, what this means is the following. When views were prioritized, the algorithm was trying to predict the probability that the user would click each video, and it would recommend the video with the highest click probability. With watch time as the goal, the algorithm is instead trying to predict how long the user would spend on each video, and it recommends the video with the highest estimate. From a revenue perspective, this change to the algorithm was not at all surprising; if anything, what is

surprising is that it didn't happen earlier. And from an overall growth perspective, it was incredibly successful. The number of users on YouTube had been steadily increasing, but the amount of time each user was spending on the platform was relatively flat prior to 2012—despite a slew of company efforts such as revamping the site to emphasize channel subscriptions and buying high-end recording equipment for top creators. But with the change in algorithmic metric from views to watch time, per-user watch time grew fifty percent a year for the next three years.[12]

In terms of content, a noticeable impact this had was to drastically reduce the amount of clickbait on the platform. Just as prioritizing (page)views led to clickbait in blogs and online newspapers, so too did it on YouTube. The pre-2012 years of YouTube saw a proliferation of videos with tantalizing titles and salacious thumbnails that would disappoint the viewer once they clicked on the video—but this disappointment was not registered because all clicks lead to views no matter how quickly the user leaves the video. After the change to the recommendation algorithm, in order to rise up in the recommendation rankings, videos have to keep viewers glued to their screens for as long as possible. As you will see throughout the remainder of this chapter, this doesn't mean that content creators no longer game the system—it just means that the rules of the game changed considerably and abruptly in 2012.

Just a month after the switch to watch time, YouTube made another key change: it started allowing all video creators—not just popular channels vetted by YouTube administrators—to run ads in their videos and earn a portion of the ad revenue. Thus, 2012 was an important year for YouTube in terms of both algorithmic and economic developments.

2015: Redesigned with Deep Learning

The next big behind-the-scenes development happened in 2015 when Google Brain, the artificial intelligence division of YouTube's parent company Google, came onboard to revamp YouTube's recommendation algorithm in an effort to further increase overall watch time on the platform. Google Brain built a new version of the recommendation algorithm based on deep learning. Recall from the crash course in Chapter 2 that in traditional machine learning the algorithm designers need to carefully choose a small number of predictors to rely on, whereas with deep learning you can toss a much larger number of predictors at the algorithm and it will automatically learn from the training data how to transform these into a smaller number of useful hierarchically structured predictors. In doing so, deep learning is able to extract higher-level conceptual patterns and meaning in the data.

[12]William Joel, "How YouTube Perfected the Feed," *Verge*, August 30, 2017: https://www.theverge.com/2017/8/30/16222850/youtube-google-brain-algorithm-video-recommendation-personalized-feed.

Jim McFadden, the technical lead for YouTube recommendations, commented[13] on this shift to deep learning: "Whereas before, if I watch this video from a comedian, our recommendations were pretty good at saying, here's another one just like it. But the Google Brain model figures out other comedians who are similar but not exactly the same—even more adjacent relationships." And it worked: aggregate watch time on YouTube increased twentyfold in the three years that followed Google Brain's involvement. However, one significant issue with deep learning is that it trades transparency for performance, and the YouTube recommendation algorithm is no exception. As McFadden himself put it: "We don't have to think as much. We'll just give it some raw data and let it figure it out."

The Google Brain deep learning algorithm starts by whittling down the vast ocean of videos on YouTube to a small pool of a few hundred videos the user might like based on the user's watched video history, keyword search history, and demographics. The demographic data include the geographic region the user is logged in from, the type of device they are using, and the user's age and gender if they have provided that information. The next step is to rank this small pool of videos from most highly recommended to least highly recommended, so that the algorithm can offer the videos it deems most likely to appeal to the user at the given moment. This ranking process relies on the user-specific predictors mentioned above but also a few hundred video-specific predictors, including details on the user's previous interactions with the channel the video is from—such as how many videos the user has watched from this channel and when was the last time the user watched a video from this channel. To prevent the user from being shown the same list of recommended videos every time, the algorithm demotes the rank of a video whenever it is offered to the user and the user does not watch it.

One of the main reasons for breaking the process into two steps—whittling then ranking—is that the first step can handle a huge volume of videos but at the expense of having a smaller number of predictors, whereas the second step can incorporate a larger number of predictors because it is focused on a small number of videos. Both steps rely on deep learning methods to extract meaningful information from this massive collection of data signals. The full recommendation system is trained on hundreds of billions of examples and results in a neural network with about one billion parameters (you might recall from Chapter 2 that this is comparable in size to the neural network used in GPT-2). This is the basic framework of YouTube's Google Brain deep learning recommendation algorithm from 2015.[14]

---

[13]See Footnote 12.

[14]Paul Covington, Jay Adams, and Emre Sargin, "Deep Neural Networks for YouTube Recommendations," Proceedings of the 10th ACM Conference on Recommender Systems (September 2016), 191–198: https://dl.acm.org/doi/10.1145/2959100.2959190.

2018: Deep Reinforcement Learning

The recommendation algorithm must strike a difficult balance between popularity and freshness. If it only recommends videos with large watch times (or other indicators of popularity such as views, upvotes, comments, etc.), then it will miss out on new content, on fresh videos that haven't yet gone viral but which might have the potential to do so. The recommendation algorithm must also strike a delicate balance between familiarity and novelty in the videos it selects for each individual user. It wants to recommend similar videos to the ones each user has already watched, since that's the most accurate guide to that user's personal tastes and interests, but if the videos are *too* similar to the ones the user has already seen, then the user might become bored and disinterested. The next big innovation brought in by the Google Brain team, in 2018, helps address these countervailing factors.

Reinforcement learning is the part of machine learning that is used to create computer programs that can beat human players at board games like chess and computer games like *StarCraft*; it has also been used to teach robots how to walk and computerized investors how to play the market. It allows the computer to explore and experiment and to learn as it does so. Put simply, supervised learning is about developing predictions, whereas reinforcement learning is about developing strategies. Reinforcement learning has been around for a few decades, but it has been powerfully revitalized in the past few years by combining it with deep learning which helps it to explore greater landscapes and to learn more deeply while doing so.

The basic idea with reinforcement learning is to create a reward function that the algorithm seeks to maximize. In computer games, the reward function is usually the number of points earned, or the number of levels completed, or the amount of time elapsed before running out of lives, or the total distance traveled, or other quantities like these. In games like chess, the reward is zero throughout the game and one when the player wins and negative one when the player loses. For investing, the reward is, unsurprisingly, return on investment. A crucial aspect of reinforcement learning is that the actions the algorithm makes are based on estimates of the future value of the reward function, not just the current value. For instance, in chess, most moves don't immediately impact the reward function at all, but with enough experience the algorithm can estimate which moves take it closer to victory. Future rewards are discounted compared to present ones, so a move that sets up a likely checkmate in three moves is more valuable than a move that sets up a likely checkmate in ten moves, but both are more valuable than a move that leads to certain defeat. It is this notion of discounted future reward that allows reinforcement learning algorithms to develop impressive long-term strategies.

But what does this have to do with YouTube? Well, in 2018, the Google Brain team brought reinforcement learning to the recommendation algorithm. Here, the "game" the computer plays is to keep each user watching videos as long as possible, so the reward function is something like the total amount of watch time each user spends in a sequence of up next recommendations before leaving the site.

Prior to reinforcement learning, the recommendation algorithm would choose and then rank the up next videos by how long it estimates the user will watch each one individually. This is like playing chess by only looking one move ahead. With reinforcement learning, the algorithm develops long-term strategies for hooking the viewer. For example, showing someone a short video that is outside their comfort zone might only score a couple minutes of watch time, but if doing this brings the viewer to a new topic they hadn't previously been exposed to, then the user might get sucked into this new topic and end up sticking around longer than if they had stayed in reliable but familiar territory. This is a long-term strategic aspect of YouTube recommendations, and it helps illustrate how reinforcement learning is well suited to tackle the delicate balances discussed earlier between popularity and freshness and between familiarity and novelty.

And…

What other significant changes to the algorithm have occurred? It is difficult to know because the algorithm is constantly tweaked and modified, but YouTube keeps the details under a veil of corporate secrecy. We know the broad strokes of the Google Brain deep learning methodology because, in a somewhat unusual move for the company, in 2016 its engineers posted a high-level technical report[15] describing their neural network framework for the algorithm. YouTube engineers also posted a paper[16] in 2019 on the reinforcement learning approach, but it takes a rather academic tone: it describes the theoretical advances provided by the proposed deep learning reinforcement learning hybrid approach, and it includes some brief empirical results from a few limited experiments, but it gives no indication that the method discussed in the paper has actually been commercially implemented in the YouTube recommendation algorithm. We only know that it did indeed become part of the algorithm in 2018 because of a comment[17] at a conference by the lead author of the paper.

[15]See Footnote 14.

[16]Minmin Chen et al., "Top-K Off-Policy Correction for a REINFORCE Recommender System," Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, January 2019, 456–464: https://arxiv.org/pdf/1812.02353.pdf.

[17]See Footnote 6.

YouTube engineers have released some other technical papers on video recommendation systems (e.g., another one[18] in 2019), but we don't know which if any of these have been absorbed into the official YouTube algorithm and which are just publications providing lines on the resumes of the authors. In many ways, the 2016 deep learning paper was the last major close-up view that has been offered from the inside.

Well, almost. Guillaume Chaslot is a former Google AI engineer who worked on the YouTube recommendation algorithm, and since leaving[19] the company in 2013, he has become a public crusader against what he perceives to be the harmful impacts of the algorithm. He has been fighting to shed light on the algorithm—and on the behind-closed-doors engineering decisions that have gone into crafting it. I'll come back to his critiques and exposés later in this chapter, but first I need to conclude this brief history of YouTube's recommendation algorithm.

Due to the critical spotlight Chaslot and others have shone on the algorithm, and the large amount of attention—flak, some might say—it received surrounding the 2016 US presidential election, YouTube company representatives have been forced to comment publicly on some aspects of the algorithm. This has provided us with some nuggets of information, but nothing nearly as detailed as what was contained in the 2016 Google Brain technical report. For instance, in a 2018 investigation[20] by the *Guardian*, YouTube representatives said that in 2016 they switched from purely optimizing watch time to also taking into account "user satisfaction" by considering how many likes videos have received and also by conducting surveys and incorporating data from those into the algorithm.

You will see a few more details of the algorithm throughout this chapter as I discuss several external data-driven investigations into the algorithm (by Chaslot and other researchers), but for now it is time to turn to the main question of this chapter: what impact has YouTube's recommendation algorithm had on society—especially in the context of fake news, popular belief in conspiracy theories, and the political candidates/parties that leverage these for support? In the aforementioned *Guardian* investigation, a YouTube spokesperson had this to say in regard to the algorithm potentially influencing the 2016 election in Trump's favor: "Our search and recommendation systems reflect what people search for, the number of videos available, and the videos people choose to watch on YouTube. That's not a bias towards any particular candidate; that is a reflection of viewer interest." One of the

[18]Zhao, et al., "Recommending What Video to Watch Next: A Multitask Ranking System," Proceedings of the 13th ACM Conference on Recommender Systems, September 2019, 43–51: https://dl.acm.org/doi/10.1145/3298689.3346997.

[19]They say he was fired over performance issues. He says it was because he was agitating for change within the company.

[20]See Footnote 2.

earliest, most prominent, and most vocal critics of YouTube's algorithm has been techno-sociologist Zeynep Tufekci. In the same *Guardian* piece, she writes: "The question before us is the ethics of leading people down hateful rabbit holes full of misinformation and lies at scale just because it works to increase the time people spend on the site—and it does work."

Whose side should we believe here? Perhaps a good place to start looking for answers is in Brazil.

YouTube in Brazil

In Brazil, the fourth largest democracy on the planet, YouTube has become more widely watched than all but one TV channel.[21] Jair Bolsonaro, the country's authoritarian far-right president, not long ago was a fringe figure lawmaker with little national recognition peddling conspiracy videos and extremist propaganda on his YouTube channel. In a relatively short span of time, his YouTube channel grew massively in subscribers and provided him with a sizable cult following. He rode this wave of YouTube popularity to presidential victory in 2018, and he wasn't alone. A whole movement of far-right YouTube stars ran for office along with Bolsonaro; many of them won their races by historic margins, and most of them now govern through YouTube the way Trump did with Twitter up until the end of his presidency.

What propelled this motley crew to such meteoric heights on YouTube? In 2019, a team of researchers at Harvard's Berkman Klein Center for Internet & Society conducted a study[22] for the *New York Times* to find out, and the answer they found is (drumroll please): the YouTube recommendation algorithm. The researchers wrote a Brazil-based computer program to start on a YouTube video from a popular channel or keyword search and then follow the chain of top-recommended videos it leads to. They ran the program thousands of times and studied the paths of videos each iteration produced. They found that regardless of whether a user started with a political or nonpolitical video, the recommendation algorithm "often favored right-wing, conspiracy-filled channels," and that "users who watched one far-right channel would often be shown many more." As the *New York Times* reported: "The algorithm had united once-marginal channels—and then built an audience for them […]. One of those channels belonged to Mr. Bolsonaro, who had long used the platform to post hoaxes and conspiracies."

YouTube's sudden predominance in Brazil coincided with the collapse of the country's political system. Bolsonaro did not change his views or behavior, in person or online; rather, his videos became frequently recommended to a burgeoning national audience in a country that was primed for a significant

[21]See Footnote 7.
[22]See Footnote 7.

political transformation. In response to the *New York Times* investigation, a spokesperson for YouTube said that the company has since "invested heavily in the policies, resources and products" to reduce the spread of harmful misinformation. Well, that's reassuring.

Having confirmed that YouTube was indeed recommending far-right propaganda leading up to and following Brazil's 2018 election, two important questions need to be addressed next: did the recommendations actually convert people ideologically, and why did the recommendation algorithm—designed by generally left-leaning Silicon Valley computer scientists—favor far-right videos? I am not aware of a rigorous quantitative study addressing the first of these two questions; the anecdotal evidence, however, and firsthand experience of members working inside the political movements, while not unequivocal, suggest the answer is yes. Let me turn to this now.

YouTube's Political Influence on Brazilians

According to the 2019 *New York Times* investigation, one local vice president of Bolsonaro's political party credited most of the party's recruitment to YouTube, including his own: "He was killing time on the site one day, he recalled, when the platform showed him a video by a right-wing blogger. He watched out of curiosity. It showed him another, and then another." He said that he didn't have an ideological political background before that experience and that YouTube's autoplaying recommendations were his political education.

A cofounder of the political group Movimento Brasil Livre (MBL), known informally as the "Brazilian Tea Party" or the "Brazilian Breitbart," which convened the popular demonstrations in 2015 pushing for the ouster of the liberal president Dilma Rousseff, said "we have something here that we call the dictatorship of the like. Reality is shaped by whatever message goes most viral." By "the like" he meant popularity on social media. MBL, a movement that was only founded in 2014, had six victorious candidates at the national level in the 2018 elections and many more at the state and local levels. The YouTube channel of MBL went from zero to one million subscribers in the year of the election, and in the month leading up to the election, it managed to reach the front page of YouTube in Brazil every single day; forty percent of the group's funding came from YouTube ad revenue.[23]

One of the nationally elected MBL candidates—who has been referred to as a "fake news kingpin," a "troll," and "Brazil's equivalent of Milo Yiannopoulos"—became at age twenty-two the youngest person ever elected to Brazil's Congress. How did he get such an early start in politics? During his last year

[23]Ryan Broderick, "YouTubers Will Enter Politics, And The Ones Who Do Are Probably Going To Win," *BuzzFeed News*, October 21, 2018: https://www.buzzfeednews.com/article/ryanhatesthis/brazils-congressional-youtubers.

of high school, he did a report on a Brazilian libertarian YouTuber—a report that he did as a YouTube video that quickly went viral and launched his own internet fame. He says MBL clashes with Bolsonaro's more militant far-right party but that MBL supported it for practical reasons: Bolsonaro is good for traffic. Another MBL candidate, now a state representative, said "I guarantee YouTubers in Brazil are more influential than politicians." His most-watched video at the time of the election? A video called "15 minutes with Jair Bolsonaro" that reached almost four million views. "I'm really grateful to YouTube because it turned me into what I am today," he once declared. In the local election that he won in 2018, he received an astronomical half million votes; candidates in previous years had won that seat with twenty thousand votes.

This question of YouTube's potential influence on Brazilian politics is still debated, and we may never truly know whether the recommendation algorithm affected the outcome of the 2018 election. That said, I think you can figure out which side of the debate I fall on. So now let me turn to the second question raised above, namely, why an ostensibly politically neutral recommendation algorithm would support extremist right-wing and conspiratorial videos. The answer, as you will see, involves the engineering adjustments YouTube made to its algorithm, especially two key developments discussed earlier: the 2012 switch of primary metric from views to watch time and the 2015 involvement of Google Brain and their deep learning architecture.

How the Far-Right Was Favored

It's certainly not true that the political videos YouTube recommended were exclusively far right—it's just that they were vastly, disproportionately so. One driving factor for this is essentially psychological: some of the emotions that tend to draw people in to content and keep them tuned in (thereby maximizing the parameter YouTube's algorithm was designed to optimize, watch time) are fear, doubt, and anger—and these are the same emotions that right-wing extremists and conspiracy theorists have relied on for years. In addition, many right-wing commentators had already been making long video essays and posting video versions of their podcast, so YouTube's switch from views to watch time inadvertently rewarded YouTube's far-right content creators for doing what they were already doing.

In other words, it may not be the case that far-right provocateurs strategically engineered their message to do well in YouTube's recommendation system—and it's almost certainly not the case that YouTube deliberately engineered its algorithm to support far-right content. Instead, the two seem to have independently reached similar conclusions on how to hook an audience, resulting in an accidental synergy. Google Brain's deep learning framework

amped up this accidental synergy by continually pushing viewers further down the rabbit hole with recommendations for increasingly provocative videos on topics viewers hadn't been exposed to. The algorithm simply wanted to offer fresh, captivating content to its users in order to maximize watch time—but the best way to do this, it appears in hindsight, was for it to provide viewers with more and more conspiracy theories, fake news, and far-right propaganda.

In time, the far-right conspiracy theorists and political commentators realized that their methods were working and that YouTube's algorithm was boosting their popularity. And even before their YouTube popularity catalyzed a political movement, these content creators were already financially incentivized to ramp up production—no matter how fake they all knew the claims in their videos were—thanks to YouTube's 2012 decision to allow all users, not just the vetted channels, to monetize videos with ad revenue.

Conspiracy Theories Flourished

The recommendation algorithm didn't just increase the viewership of fake news and conspiracy theories on YouTube, it also provided an air of legitimacy to them. Even if a particular conspiracy theory seems blatantly implausible, as YouTube recommends a sequence of videos from different creators on the same topic mimicking each other, the viewer tends to feel that all signs are pointing to the same hidden truth. Debora Diniz, a Brazilian women's rights activist who became the target of an intense right-wing YouTube conspiracy theory smear campaign, said[24] this aspect of the algorithm makes it feel "like the connection is made by the viewer, but the connection is made by the system."

This phenomenon can be seen in topics outside of politics as well. Doctors in Brazil found that not long after Google Brain's 2015 redesign of the recommendation algorithm, patients would come in blaming Zika on vaccines and insecticides (the very insecticides that in reality were being used to limit the spread of the mosquito-borne disease). Patients also were increasingly refusing crucial professional medical advice due to their own "YouTube education" on health matters. The Harvard researchers involved in the *New York Times* investigation of YouTube in Brazil found that "YouTube's systems frequently directed users who searched for information on Zika, or even those who watched a reputable video on health issues, toward conspiracy channels." A YouTube spokesperson confirmed these findings and said the company would change how its search tool surfaced videos related to Zika (a band-aid on a bullet wound, in my opinion). Why did people create these harmful medical disinformation videos in the first place, and why did YouTube

[24]See Footnote 7.

recommend them? Because they attracted viewers and drove lengthy watch times—which means they made money for both the creators of these videos and for YouTube itself.

Playing the Game

Remember from the timeline of YouTube's algorithm how in 2018 Google Brain brought in a machine learning technique called reinforcement learning—more commonly used for playing games—that allows the recommendation algorithm to develop long-term strategies for sucking in viewers? At an AI conference in 2019, a Google Brain researcher said[25] this was YouTube's most successful adjustment to the algorithm in two years in terms of driving increased watch time. She also said that it was already altering the behavior of users on the platform: "We can really lead the users toward a different state, versus recommending content that is familiar." This is a dangerous game to play when that different YouTube state is a chain of far-right conspiracy videos which might ultimately have led to a different political state for all citizens of Brazil—a xenophobic, anti-science, authoritarian state. "Sometimes I'm watching videos about a game, and all of a sudden it's a Bolsonaro video," said[26] a seventeen-year-old high school student in Brazil, where the voting age is sixteen.

YouTube in America

While YouTube is inordinately popular in Brazil, especially among the voting youth, it is nearly impossible to fathom that the problem of YouTube pushing viewers to extremist right-wing videos was isolated and somehow only occurred in Brazil. The potential impact of YouTube's recommendation algorithm on the alt-right movement in the United States—especially in the context of Trump's 2016 election victory and his efforts to overturn the results of his 2020 election loss, and more generally with regard to the growing national discussion of fake news and dangerous conspiracy theory movements—continues to be a hotly debated topic to this day. According to a 2019 investigation, the years leading up to Trump's 2016 victory were particularly reckless ones at YouTube:[27] "Several current and former YouTube employees […] said company leaders were obsessed with increasing engagement during those years. The executives, the people said, rarely considered whether the company's algorithms were fueling the spread of extreme and hateful political content." Awareness of this issue, both inside and outside the YouTube organization, is certainly greater now than it was then, but that doesn't mean the problems have gone away.

[25]See Footnote 6.
[26]See Footnote 7.
[27]See Footnote 6.

Stirring Up Electoral Trouble in 2020

On election day in 2020, hours before any of the polls had closed, eight videos out of the top twenty in a YouTube search for "LIVE 2020 Presidential Election Results" were showing similar maps with fake electoral college results.[28] One of the channels in this list had almost one and a half million subscribers, and several of the channels were "verified" by YouTube. The top four search results for "Presidential Election Results" were all fake. Curiously, most of the YouTube channels coming up in election day searches for election results were not even affiliated with political or news organizations—they were just people opportunistically using the election to snag some easy ad revenue.

In the days after the 2020 election, a network of fake news channels on YouTube sprang up[29] and peddled Trump's false claims that the election was rigged and victory was stolen from him. These channels have close ties, albeit largely obfuscated, with the *Epoch Times* media organization that you encountered in Chapter 2 in the context of algorithmically mass-produced fake news. Michael Lewis, the host of one of the channels in this network, went live just hours after the Capitol building insurrection to repeat Trump's lies about the election and to blame the Capitol building mob on antifa. His YouTube channel recorded over two hundred thousand subscribers and ten million views in less than two months. The channel describes itself as an independent effort by Lewis and a few friends who "felt like truth was dying," despite connections to *Epoch Times* that were uncovered after some journalistic sleuthing.

Collectively, this network of seven fake news YouTube channels that launched in mid-November 2020 amassed over a million subscribers and tens of millions of views by mid-January 2021. On December 9, 2020, YouTube announced[30] that it would remove any videos posted after this date that claim there was widespread fraud or errors that influenced the outcome of the election; evidently this policy was not enforced vigilantly enough to prevent the millions of views that new videos from these channels received between December 9 and January 6.

---

[28]Kat Tenbarge, "YouTube channels made money off of fake election results livestreams with thousands of viewers," *Insider*, November 3, 2020: https://www.insider.com/youtube-fake-election-results-livestreams-monetized-misinformation-2020-11.

[29]See Footnote 1.

[30]"Supporting the 2020 U.S. election," *YouTube blog*, December 9, 2020: https://blog.youtube/news-and-events/supporting-the-2020-us-election/.

YouTube in the American Media Landscape

Do Americans actually get their news and political information from YouTube and the videos it recommends? A 2018 poll by the Knight Foundation and Gallup found[31] that most US adults—and more than nine in ten Republicans—say they personally have lost trust in the news media in recent years; this suggests that they are turning to other sources for their information. Meanwhile, a 2018 Pew Research Center survey found[32] that the share of YouTube users who say they get news or news headlines from YouTube nearly doubled between 2013 (twenty percent) and 2018 (thirty-eight percent), and that around half of YouTube users say the site is at least somewhat important for helping them understand things that are happening in the world. Around two-thirds of users say they at least sometimes encounter videos that seem obviously false or untrue.

This Pew survey also found that eighty-one percent of YouTube users say they at least occasionally watch the "up next" videos suggested by the recommendation algorithm, and fifteen percent say they do so regularly. Perhaps people are not always honest with pollsters, or even themselves, about how often they let an algorithm dictate their viewing habits: recall that YouTube's internal accounting found that seventy percent of all watch time comes by way of recommended videos.

In addition to direct viewership, another way that YouTube is shifting political discourse in America is through a sort of ripple effect where YouTube serves up sizable audiences to various individuals who then reach even more massive mainstream audiences on traditional media outlets. The following story illustrates this dynamic.

In the first weeks of the coronavirus pandemic in January 2020, a medical researcher in Hong Kong named Dr. Li-Meng Yan had, based on unsubstantiated rumors (which it later turned out were totally fabricated and false), started to believe that the virus was a bioweapon manufactured by the government in mainland China and deliberately released on the public. To spread a message of warning, she reached out to a popular Chinese YouTube personality, Wang Dinggang, known for criticizing the Chinese Communist Party. Dr. Yan portrayed herself as a whistleblower and anonymous source to Dinggang,

[31] "Indicators of News Media Trust," *Knight Foundation*, September 11, 2018: https://knightfoundation.org/reports/indicators-of-news-media-trust/.

[32] Aaron Smith, Skye Toor, and Patrick Van Kessel, "Many Turn to YouTube for Children's Content, News, How-To Lessons," *Pew Research Center*, November 7, 2018: https://www.pewresearch.org/internet/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons/.

who then broadcast this fake news story about the coronavirus to his one hundred thousand YouTube followers. The story continued to spread and spiral; then in September, Dr. Yan shed her anonymity and appeared on Fox News in an interview with Tucker Carlson that racked up nearly nine million online views.

The same Chinese YouTube host, Dinggang, is also believed[33] to have been the first to seed baseless child abuse rumors about Hunter Biden—rumors that spread from his YouTube channel to *InfoWars* and then to the mainstream press in the *New York Post*. In this way, YouTube provides a powerful entry point for the dangerous vertical propagation phenomenon of fake news studied in Chapter 1.

Researchers found[34] that people who believe in conspiracy theories tend to rely more heavily on social media for information than do the less conspiratorially inclined segments of the population. Specifically, sixty percent of those who believe that COVID-19 is caused by radiation from 5G towers said that "much of their information on the virus came from YouTube," whereas this figure drops to fourteen percent for those who do not believe this false conspiracy. People who ignored public health advice and went outside while having COVID symptoms were also much more reliant on YouTube for medical news and information than the general public.

But this book is not about the role that *technology* in general plays in fake news, it is specifically about the role played by *algorithms*—especially sophisticated machine learning algorithms—and in this chapter that takes the form of the YouTube recommendation algorithm. In order to understand the potential influence of the recommendation algorithm on American politics, I'll turn now to several empirical investigations that provide a window into the algorithm's behavior.

Studying the Algorithm

In parallel to the Pew surveys mentioned in the preceding section, Pew also conducted a *random walk* exploration of the YouTube recommendation algorithm, similar to the Harvard investigation conducted in Brazil. Let me start with this.

---

[33]Amy Qin, Vivian Wang, and Danny Hakim, "How Steve Bannon and a Chinese Billionaire Created a Right-Wing Coronavirus Media Sensation," *New York Times*, November 20, 2020: https://www.nytimes.com/2020/11/20/business/media/steve-bannon-china.html.

[34]Rory Cellan-Jones, "Coronavirus: Social media users more likely to believe conspiracies," *BBC News*, June 17, 2020: https://www.bbc.com/news/technology-53083341.

Pew's Random Walk

Pew's exploration[35] was conducted by taking the following steps one hundred seventy thousand times:

1. Select a top-ranked video at random from a list of more than fourteen thousand English-language YouTube channels with at least a quarter million subscribers.
2. Select at random one of the top five "up next" recommended videos.
3. Repeat the previous step four times.

This resulted in one hundred seventy thousand different five-video-deep walks down the algorithm's road—all of which were done for an "anonymous" user, meaning one that is not logged in and so has no viewing history or other personal data that the algorithm can rely on. By analyzing these random walk videos empirically in the aggregate, Pew's main finding was that the YouTube algorithm encourages users to watch progressively longer and more popular videos: the average length of the videos increased from nine and a half minutes for the originally selected video to fifteen minutes for the fifth video in the random walk, and the average view counts increased from eight million for the first to thirty million for the fifth. The videos in these random walks covered a range of topics, but a large share of them were music videos, TV competitions, children's content, and life hacks.

The gist of the Pew random walk experiment, in other words, is that YouTube pushes viewers toward popular, often mainstream, content. Some people took this as evidence that the fears of YouTube's algorithm tainting our political waters with divisive alt-right content were overblown, if not outright fabricated. I'm not convinced, and neither are many other scholars.

First of all, the videos that launched these random walks were very popular—eight million views on average, as I mentioned. One can certainly imagine that once the algorithm finds users watching very popular content, it keeps them in the orbit of highly viewed content, whereas users who show an interest in videos with fewer views might be recommended less mainstream content. We don't know how the random walks might have differed if they had started with more specialized content. Secondly, the large-scale analysis here was rather coarse—it summarized views, durations, and content categories, but it did not look into distinctions such as legitimate news versus fake news and mainstream politics versus extremist politics (let alone left versus right). It is somewhat reassuring that users were generally pushed toward anodyne categories like music videos and children's videos, but this aggregate behavior

[35]See Footnote 32.

might mask a lot of important variation, and it says nothing about users who specifically seek out news-related content among their recommendations. There are other ways of probing the recommendation algorithm, as you'll soon see, and a finer-tooth comb reveals a much darker story.

Chaslot's Political Recommendation Data

Chaslot, the computer engineer fired from YouTube's recommendation algorithm team in 2013, wrote a program in 2016 that, like the one used in Pew's random walks, was designed to explore the places YouTube's recommendation algorithm takes its viewers by starting with a "seed" video and then automatically clicking the top "up next" videos one at a time. One of the main differences with his approach compared to Pew's is that rather than starting with random popular seed videos, Chaslot's seed videos were the result of specific searches that he believed were common and/or important during the 2016 election. That is, he tried to simulate not just recommendations from random popularity on YouTube, but recommendations stemming specifically from timely political inquiries. He also looked at the quality of the information in the recommended videos, instead of providing only a coarse topical classification as with Pew's aggregate analysis. For eighteen months, he used his program to conduct a variety of experiments, the results of which are reported[36] in the *Guardian*. His research "suggests YouTube systematically amplifies videos that are divisive, sensational and conspiratorial."

For instance, when Chaslot's seed video was the result of a keyword search for "who is Michelle Obama?", the chain of up next recommendations led mostly to videos that claimed she is a man. When the seed was from a search for the pope, eighty percent of the videos in the up next recommendation sequence claimed he is "evil," "satanic," or "the anti-Christ." Quite strikingly, Chaslot found that whether a user searched for "Clinton" or "Trump" and then clicked a video and followed the sequence of up next recommendations, the algorithm "was much more likely to push you in a pro-Trump direction." Trump won the electoral college in 2016 as a result of eighty thousand votes spread across three swing states; at the time, there were more than one hundred fifty million YouTube users in the United States, so even a small degree of political bias in the recommendation algorithm could have had a decisive impact on the electoral outcome.

Chaslot sent journalists at the *Guardian* a database of more than eight thousand videos that his program reached during the four months leading up to the 2016 US election after doing an equal number of searches for "Trump" and for "Clinton" and then following a chain of top up next videos. These videos are certainly not comprehensive nor even necessarily representative of the

[36]See Footnote 2.

political content on YouTube at the time, but they do provide a snapshot into the recommendation system just prior to the election, and, in the words of Jonathan Albright, research director at the Tow Center for Digital Journalism, it is a "reputable methodology" that "captured the apparent direction of YouTube's political ecosystem."

When analyzing these eight thousand videos, the *Guardian* journalists said they "were stunned by how many extreme and conspiratorial videos had been recommended, and the fact that almost all of them appeared to be directed against Clinton." Some of the recommended videos were unsurprising—clips from speeches, debates, news, even *Saturday Night Live* sketches—but they often found anti-Clinton conspiracy videos among the recommendations offered by the algorithm to a user watching one of these more mainstream political videos. The anti-Clinton conspiracy theories ranged from questioning her health and mental fitness to "accusing Clinton of involvement in murders or connecting her to satanic and paedophilic cults." The journalists went through by hand the top one thousand most recommended videos in Chaslot's database of eight thousand and found that "Just over a third of the videos were either unrelated to the election or contained content that was broadly neutral or even-handed. Of the remaining 643 videos, 551 were videos favouring Trump, while only 92 favoured the Clinton campaign."

Chaslot's data show that the recommendation algorithm was particularly favorable to Alex Jones' *InfoWars* channel, which YouTube eventually removed and banned in 2018—about a week after Facebook first did so. When YouTube took the channel down, it had already amassed two and a half million followers and one and a half billion pageviews across thirty-six thousand videos. Another channel that was heavily pushed by the recommendation algorithm is the *Next News Network* run by Gary Franchi, which according to the *Guardian* "has the appearances of a credible news channel. But behind the facade is a dubious operation that recycles stories harvested from far-right publications, fake news sites and Russian media outlets." Chaslot's research suggests that the popularity of this channel could largely have come from YouTube's recommendation algorithm; YouTube sharply dismissed this. The *Guardian* journalists contacted Franchi to find out who was correct in this debate, and Franchi sent back screenshots of the official private data that YouTube provides to its content creators on the sources of their videos' traffic. One of Franchi's more popular videos was a fake news story about Bill Clinton raping a thirteen-year-old that had two and a half million views; Franchi's data screenshot showed that the largest source of traffic to this video was YouTube recommendations. In fact, YouTube recommendations were the primary traffic source for all but one of the videos in Franchi's screenshot.

While Franchi is a professional (of sorts) fully devoted to his channel, the *Guardian* found that even the "amateur sleuths" and "part-time conspiracy theorists," who typically received only a few hundred views on their videos,

were "shocked when their anti-Clinton videos started to receive millions of views, as if they were being pushed by an invisible force." And in nearly every case, YouTube's traffic data revealed that invisible force to be the recommendation algorithm. In one case, YouTube emailed a content creator that his anti-Clinton fake conspiracy video violated its guidelines—and yet, traffic continued to flow in to the video from the recommendation algorithm after this email, and it ended up getting over two million views prior to the election.

Tracking Commenters

In 2019, a team of scholars at a Brazilian research institute and a Swiss research institute conducted a different kind of study[37] into how YouTube might be driving people to the far right. The team manually classified over three hundred thousand videos on nearly three hundred fifty YouTube channels into a system of four categories designed by the Anti-Defamation League that provides a spectrum of extremism. From least to most extreme, these are media (traditional factual news), the intellectual dark web (IDW, a community that openly considers controversial topics like eugenics and "race science"), the alt-lite (which purports to deny white supremacy but believes in conspiracy theories about "replacement" by minority groups), and the alt-right (a loose segment of the white supremacist movement consisting of individuals who reject mainstream conservatism in favor of politics that embrace racist, anti-Semitic, and white supremacist ideology and who push for a white ethnostate).

By tracking the authors of over seventy-two million comments on these videos, they found that "users consistently migrate from milder to more extreme content" and "users who consumed alt-lite or IDW content in a given year go on to become a significant fraction of the alt-right user base in the following year." This team also investigated the recommendation algorithm and found possible pathways to alt-right radicalization, but they were rather faint ("from the alt-lite we follow the recommender system 5 times, approximately 1 out of each 25 times we will have spotted an alt-right channel").

It should be noted that tracking of comment activity provides only a limited and possibly distorted window into YouTube viewership because the large majority of viewers do not comment (and one cannot presume that the commenters are representative of the full population of viewers), and also many comments are from viewers refuting the claims in the video and debating with, or simply trolling, the supporters. This research project did not consider the content of comments. Moreover, the fact that commenters flow to further

[37] Manoel Horta Ribeiro et al., "Auditing Radicalization Pathways on YouTube," Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 131–141: https://dl.acm.org/doi/10.1145/3351095.3372879.

extremes on YouTube each year does not guarantee that YouTube is the cause of this rightward migration; it could well be that some external force is pushing viewers this way, and YouTube is merely reflecting this force. Nonetheless, tracking commenters is an interesting way of collecting empirical evidence that is quite distinct from the random walk approaches discussed earlier—and this commenter study appears to point in the same general direction as those random walk studies.

In response to this Brazilian-Swiss investigation, YouTube said[38] "We strongly disagree with the methodology, data and, most importantly, the conclusions made in this new research." Incidentally, when journalists from the *Guardian* reached out to YouTube for a reaction to their investigation based on Chaslot's database, YouTube responded that it has a great deal of respect for the newspaper and its journalists but that "We strongly disagree, however, with the methodology, data and, most importantly, the conclusions made in their research." We can at least give YouTube credit for consistency.

Contradictory Results

Another empirical study[39] of the recommendation algorithm was conducted in 2019 and, like the Pew random walk investigation, found that YouTube's recommendation algorithm did the opposite of radicalize—it "actively discourages viewers from visiting radicalizing or extremist content. Instead, the algorithm is shown to favor mainstream media and cable news content over independent YouTube channels." The authors, Ledwich and Zaitsev, even assert "we believe that it would be fair to state that the majority of the views are directed towards left-leaning mainstream content."

This created quite a stir on social media, and it elicited a lengthy response[40] from the Brazilian-Swiss team whose findings it directly contradicts. The Brazilian-Swiss team began their point-by-point rebuttal with the following pithy accusation: "Large-scale measurement and analysis of social media data is hard. The authors [Ledwich and Zaitsev] misunderstood their data source and ended up measuring a different thing than they thought they did, and made unfounded claims based on their results." Ledwich and Zaitsev in turn

[38]Tanya Basu, "YouTube's algorithm seems to be funneling people to alt-right videos," *MIT Technology Review*, January 29, 2020: https://www.technologyreview.com/2020/01/29/276000/a-study-of-youtube-comments-shows-how-its-turning-people-onto-the-alt-right/.

[39]Mark Ledwich and Anna Zaitsev, "Algorithmic extremism: Examining YouTube's rabbit hole of radicalization," *First Monday*, Volume 25, Number 3, March 2, 2020: https://firstmonday.org/ojs/index.php/fm/article/view/10419/9404.

[40]Manoel Horta Ribeiro et al., "Comments on 'Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization'," *iDRAMA Lab*, December 29, 2019: https://idrama.science/posts/2019/12/youtube-radicalization-study/.

rebutted[41] this rebuttal. It's hard to know what to make of this debate; both sides appear to raise valid points while reaching diametrically opposed conclusions.

Longitudinal Study

In the spring of 2020, Chaslot and two professors at UC Berkeley concluded a longitudinal study[42] of conspiracy videos on YouTube. They first used text-based supervised machine learning to train an algorithm to estimate whether a YouTube video was conspiratorial by looking at the description, transcript, and comments. Then, they applied this to eight million videos recommended over a fifteen-month period to a logged-out user after watching videos from one thousand popular news-related channels. What did they find? Hold that thought a second.

YouTube announced[43] in January 2019 that it would "begin reducing recommendations of borderline content and content that could misinform users in harmful ways—such as videos promoting a phony miracle cure for a serious illness, claiming the earth is flat, or making blatantly false claims about historic events like 9/11." The Chaslot-Berkeley research collaboration found that the number of conspiracy videos recommended by the algorithm indeed dropped steadily in the months after this announcement—from January to May 2019, it decreased by seventy percent—but after a relative low point in May, the number crept back up and by March 2020 it was only forty percent lower than when the YouTube crackdown began in January 2019.

Interestingly, the Chaslot-Berkeley study found that the results of the YouTube crackdown varied significantly across the different categories of conspiracy theories. Flat Earth videos and 9/11 hoax videos have been almost completely scrubbed from YouTube, whereas climate change denial videos and videos claiming aliens built the pyramids have persisted and even flourished. The YouTube announcement did say that it was focusing on content that could misinform users in "harmful ways," but it seems rather puzzling how they've chosen to interpret and enforce that policy. One of the Berkeley collaborators,

---

[41] Anna Zaitsev, "Response to further critique on our paper 'Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization'," *Medium*, January 8, 2020: https://medium.com/@anna.zaitsev/response-to-further-critique-on-our-paper-algorithmic-extremism-examining-youtubes-rabbit-hole-af3226896203.

[42] Marc Faddoul, Guillaume Chaslot, and Hany Farid, "A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos," preprint, March 6, 2020: https://arxiv.org/pdf/2003.03318.pdf.

[43] "Continuing our work to improve recommendations on YouTube," *YouTube blog*, January 25, 2019: https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html.

Hany Farid, had the following to say:[44] "If you have the ability to essentially drive some of the particularly problematic content close to zero, well then you can do more on lots of things. They use the word 'can't' when they mean 'won't'." I'll return to this issue of moderating misinformation on YouTube shortly—then I'll provide a broader treatment of content moderation on social media in Chapter 8.

The Role of Viewing History

One potentially significant limitation with all the data-driven investigations of the YouTube recommendation algorithm discussed in this chapter—the Harvard team working for the *New York Times* to explore the situation in Brazil, the Pew random walk empirical study, the Chaslot political database analyzed by the *Guardian*, the Ledwich-Zaitsev paper, and the Chaslot-Berkeley longitudinal study—is that they all rely on a logged-out anonymous user. This means a user without prior viewing history, search history, or demographic information.

Part of the beauty of deep learning (which, as you recall, is the framework Google Brain brought to YouTube recommendation in 2015) compared with earlier forms of machine learning is that with deep learning the algorithm is able to use a huge number of predictors since the neural network training process automatically transforms these into a smaller number of relevant and hierarchically structured predictors. In the context of YouTube, this means that rather than just using a small number of obvious predictors such as view counts and average view durations of videos, the recommendation algorithm is able to dissect and analyze users on a more individualized basis by relying on detailed viewing history and behavior. In short, deep learning allows for extreme personalization.

This is one reason why deep learning is so promising in the realm of healthcare: it means computers can help doctors custom-tailor diagnoses and treatments to an incredible extent. But for YouTube recommendations, it means the algorithm knows a heck of a lot about you—and the videos it recommends to you depend heavily on your personal data. Consequently, it is damn hard to get an accurate portrait of how YouTube's algorithm behaves in real life and how it impacts society by simulating it with a computer program that simply clicks videos from an anonymous login without viewing history. The only conceivable way to get more authentic data on the algorithm would be to recruit a large group of volunteers to track their experiences with YouTube recommendations for some period of time. Alas, I know of no such human participant–based studies.

[44]See Footnote 4.

Another Algorithmic Misfire

The recommendation algorithm isn't YouTube's only algorithmic culprit when it comes to spreading fake news and disinformation. Another company algorithm automatically puts together videos on the platform into channels it creates on various topics. For instance, CNN posts all its videos to its official YouTube channel, but YouTube's internal algorithm also creates channels for each of the network's popular shows. The problem is that not all the videos this algorithm finds are authentic. Well, sort of.

It was found[45] in December 2019 that some of the videos that ended up in YouTube's algorithmically generated CNN channels were actually from fake news organizations that deceitfully posed as CNN. What these organizations did is post copies of actual clips from CNN, but they edited the thumbnail images to make the content look shocking and inflammatory. For example, one thumbnail showed an official-looking CNN graphic of President Trump and the president of Iran and a chyron that read "War officially started!" But the video itself was an undoctored clip from CNN that had nothing to do with war. In short, the video content itself wasn't fake news, but YouTube users who browsed the videos listed in these algorithmically curated news channels would see alarming false headlines in the thumbnails.

In just a single week, these CNN videos with fake thumbnails received more than eight million views. A YouTube investigation found that the organizations posting them were not part of any coordinated political influence or disinformation campaigns, and no connections to foreign governments were uncovered. Instead, it seems, these organizations were simply using this clickbait trick in order to profit from ad revenue.

You might be asking yourself at this point: wait, didn't YouTube eliminate clickbait when it switched from views to watch time? Mostly yes, but I suspect what happened here is that users would click one of the videos with an alarming thumbnail and then watch it for a considerable duration thinking the story from the thumbnail was just one of the segments on the episode—perhaps they even watched the video all the way through to the end before they realized they had been duped (I'll admit that I've been suckered into watching entire YouTube videos this way). This is still a form of clickbait, but since it likely results in lengthy watch times, it is actually promoted by the post-2012 recommendation algorithm. On the other hand, as you may recall, YouTube did say that in 2016 it started incorporating upvotes in addition to watch time in the algorithm's considerations, and presumably these fake CNN videos do not fare well in that metric. Thus, how strongly the algorithm promoted these videos seems to largely come down to how the algorithm's

[45]Donie O'Sullivan, "Report: Fake news content went viral using YouTube's algorithm," *Mercury News*, December 13, 2019: https://www.mercurynews.com/2019/12/13/report-fake-news-content-went-viral-using-youtubes-algorithm/.

engineers decided to balance these two metrics—a balance that has likely shifted over the years and regardless has been kept out of public view.

Moderating Content on YouTube

The FBI produced a document[46] in May 2019 identifying the spread of fringe conspiracy theories as a new domestic terrorism threat that would likely intensify throughout the 2020 election cycle. Nine months later, in February 2020, the US Court of Appeals for the Ninth Circuit ruled[47] that YouTube is a private forum and therefore not subject to free speech requirements under the First Amendment—a ruling that allows YouTube to freely make decisions on what content to prohibit. The next month, a spokesperson for YouTube said[48] that "In the past year alone, we've launched over 30 different changes to reduce recommendations of borderline content and harmful misinformation, including climate change misinformation and other types of conspiracy videos" and that "Thanks to this change, watchtime this type of content gets from recommendations has dropped by over 70 percent in the U.S." Unfortunately, what these thirty changes comprised is not public knowledge, nor is the methodology supporting this claim of a seventy percent reduction—so we are left to blindly take YouTube at its corporate word (and recall that the Chaslot-Berkeley longitudinal study found that recommendations of conspiracy videos initially declined by seventy percent during the first few months of this period but then started creeping back up after that).

In March 2020, YouTube also announced[49] that it would rely more heavily on machine learning to moderate its content[50] while human reviewers were sent home for the pandemic lockdown. But six months later, the company admitted[51] that this greater reliance on AI for moderation led to a significant

---

[46]Jana Winter, "Exclusive: FBI document warns conspiracy theories are a new domestic terrorism threat," *Yahoo News*, August 1, 2019: https://news.yahoo.com/fbi-documents-conspiracy-theories-terrorism-160000507.html.

[47]Jon Brodkin, "First Amendment doesn't apply on YouTube; judges reject PragerU lawsuit," *Ars Technica*, February 26, 2020: https://arstechnica.com/tech-policy/2020/02/first-amendment-doesnt-apply-on-youtube-judges-reject-prageru-lawsuit/.

[48]See Footnote 4.

[49]"Protecting our extended workforce and the community," *YouTube blog*, March 16, 2020: https://blog.youtube/news-and-events/protecting-our-extended-workforce-and.

[50]No technical details were provided in this announcement, but in Chapter 8 I'll turn to machine learning approaches to moderating social media more generally.

[51]James Vincent, "YouTube brings back more human moderators after AI systems over-censor," *The Verge*, September 21, 2020: https://www.theverge.com/2020/9/21/21448916/youtube-automated-moderation-ai-machine-learning-increased-errors-takedowns.

increase in incorrect video removals. Around eleven million videos were taken down during this six-month period, which is roughly twice the normal rate. Over three hundred thousand of these takedowns were appealed, and half of the appeals were successful. YouTube's Chief Product Officer (the same one who earlier in this chapter revealed that seventy percent of YouTube watch time comes via the recommendation algorithm) revealingly said that the machine learning approach to content moderation "couldn't be as precise as humans" so the company decided to err on the side of caution during this period. He also pointed out that of these eleven million videos, more than half were removed before a single actual YouTube user watched the video. "That's the power of machines," he said.

Just one month later, in October 2020, YouTube posted a company blog post[52] titled "Managing harmful conspiracy theories on YouTube." It included an announcement that YouTube's policy on hate speech and harassment is now updated and expanded to prohibit "content that targets an individual or group with conspiracy theories that have been used to justify real-world violence." One significant consequence is that many videos in the QAnon and Pizzagate conspiracy theory movements are now classified as banned content. It is important to note, however, that it is not the general fake news aspect of these movements that violates YouTube's new policy—outlandish and politically destructive as they are—it is specifically the association with offline physical violence that triggers the new prohibition. YouTube was already removing specific QAnon and Pizzagate videos that directly threatened violence, but the expanded policies mean that grounds for prohibition now include "content that threatens or harasses someone by suggesting they are complicit in one of these harmful conspiracies, such as QAnon or Pizzagate."

Recall from the discussion of the Chaslot-Berkeley longitudinal study that in January 2019 YouTube announced a significant, but undisclosed, update to the recommendation algorithm aimed specifically at limiting the reach of harmful misinformation. The October 2020 blog post announcing new restrictions on QAnon said the number of views on videos related to QAnon coming from the recommendation algorithm had already dropped by eighty percent since January 2019. These efforts to reduce the presence of QAnon are important, but one might argue that it was too little, too late: the genie is already out of the bottle.

Indeed, it is believed that YouTube played one of the largest roles among the social media platforms in "moving QAnon from the fringes to the mainstream,"

[52]"Managing harmful conspiracy theories on YouTube," *YouTube blog*, October 15, 2020: https://blog.youtube/news-and-events/harmful-conspiracy-theories-youtube.

in the words[53] of *New York Times* journalist Kevin Roose. Roose reported that a number of QAnon's early activists produced "YouTube documentaries" explaining the movement's core beliefs and that these videos were shared on Facebook and other platforms—which helped them gather millions of views and draw more people into the movement. Some individuals also rose to prominence in the QAnon movement by creating popular "YouTube talk shows" focusing on the latest developments in the world of QAnon.

Summary and Concluding Thoughts

The misinformation war is a complex arms race that requires constant vigilance. Over the past several years, we have witnessed a repetitive cycle in which a journalistic accusation or academic investigation faults YouTube's recommendation algorithm for radicalizing viewers and pushing people to dangerous fringe movements like the alt-right, then a YouTube spokesperson responds that they disagree with the accusation/methodology but also that the company is working to reduce the spread of harmful material on its site—in essence admitting there is a problem without admitting that the company is at fault.[54]

YouTube's deep learning algorithm provides highly personalized recommendations based on a detailed portrait of each user drawn from their viewing history, search history, and demographic information. This renders it extremely difficult for external researchers to obtain an accurate empirical assessment of how the recommendation algorithm performs in the real world. Nonetheless, the glimpses we have seen and discussed throughout this chapter—from YouTube insiders to external investigations to firsthand experiences of political participants and fake news peddlers—together with the undeniable context that large segments of society increasingly have lost faith in mainstream forms of media and now turn to platforms like YouTube for their news, are fairly compelling evidence that YouTube's recommendation algorithm has had a pernicious influence on our society and politics throughout the past half decade.

What the current state of the matter is, and whether YouTube's internal adjustments to the algorithm have been enough to keep pace, is unclear. What is clear is that by allowing this misinformation arms race to take place behind closed doors in the engineering backrooms of Google, we are placing a

[53]Kevin Roose, "YouTube Cracks Down on QAnon Conspiracy Theory, Citing Offline Violence," *New York Times*, October 15, 2020: https://www.nytimes.com/2020/10/15/technology/youtube-bans-qanon-violence.html.

[54]In fact, Mozilla has compiled a list of instances of this corporate behavior by YouTube: Brandi Geurkink, "Congratulations, YouTube... Now Show Your Work," *Mozilla*, December 5, 2019: https://foundation.mozilla.org/en/blog/congratulations-youtube-now-show-your-work/.

tremendous amount of trust in the company. As Paul Lewis wrote in his piece[55] on YouTube for the *Guardian*, "By keeping the algorithm and its results under wraps, YouTube ensures that any patterns that indicate unintended biases or distortions associated with its algorithm are concealed from public view. By putting a wall around its data, YouTube […] protects itself from scrutiny."

An extensive, scathing investigation[56] into the corporate culture at YouTube in 2019 by *Bloomberg News*, based largely on insider information from current and former employees, found that attempts to curb the spread of conspiracy theory videos through proposed adjustments to the recommendation algorithm were routinely shot down by upper management for the sake of "engagement," a measure of the views, watch time, and interactions with videos: "Conversations with over twenty people who work at, or recently left, YouTube reveal a corporate leadership unable or unwilling to act on these internal alarms for fear of throttling engagement." Some employees simply tried to collect data on videos they felt were harmful but which didn't officially violate YouTube's policies, but "they got the same basic response: Don't rock the boat."

There are some grassroots efforts to counterbalance the far-right content on YouTube by creating far-left content that indulges in the same sensationalist techniques that seem to have resulted in large view counts and watch times propped up by the recommendation algorithm. One of the main instances of this is a group called *BreadTube*, whose name is a reference to the 1892 book *The Conquest of Bread* by Russian anarchist/communist revolutionary Peter Kropotkin. While I can understand the short-term desire to rebalance the system this way by hijacking methods from the alt-right movement, for the long-term outlook this really does not seem like a healthy way to correct the problem that YouTube has unleashed on society. But without resorting to such extreme methods, we can either trust YouTube to keep fixing the problem on its own in secret or we can push for more transparency, accountability, and government regulation; my vote is for the latter.

One reason for this view is that even if YouTube voluntarily takes a more proactive stance in the fight against fake news, other video platforms will step in to fill the unregulated void left in its place. In fact, this is already happening: *Rumble* is a video site founded in 2013 that has recently emerged as a conservative and free speech–oriented alternative to YouTube (similar to the role Parler plays in relation to Twitter). The founder and chief executive of

[55]See Footnote 2.

[56]Mark Bergen, "YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant," *Bloomberg News*, April 2, 2019: https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant.

Rumble said[57] that the platform has been on a "rocket ship" of growth since summer 2020 that has only accelerated since the election. He said the platform "prohibits explicit content, terrorist propaganda and harassment," but that it was "not in the business of sorting out misinformation or curbing speech." This suggests to me that we should not just leave it up to individual companies to decide how and how much to moderate their content—we need a more centralized, cohesive approach in order not to fall hopelessly behind in the fight against fake news. I'll return to this discussion in a broader context in Chapter 8.

While waiting for legislative efforts to address this problem, it is important in the meantime to look carefully at the technical tools we have at our disposal. In the next chapter, I'll explore whether recent lie detection algorithms powered by machine learning could be used to detect disinformation in online videos.

[57]Mike Isaac and Kellen Browning, "Fact-Checked on Facebook and Twitter, Conservatives Switch Their Apps," *New York Times*, November 11, 2020: https://www.nytimes.com/2020/11/11/technology/parler-rumble-newsmax.html.

CHAPTER

5

Prevarication and the Polygraph

Can Computers Detect Lies?

There is no lie detector, neither man nor machine.

—US House of Representatives Committee Report

Wouldn't it be nice if we could take a video clip of someone talking and apply AI to determine whether or not they're telling the truth? Such a tool would have myriad applications, including helping in the fight against fake news: a dissembling politician giving a dishonest speech would immediately be outed, as would a conspiracy theorist knowingly posting lies on YouTube. With the remarkable progress in deep learning in recent years, why can't we just train an algorithm by showing it lots of videos of lies and videos of truth and have it learn which is which based on whatever visual and auditory clues it can find? In fact, for the past fifteen years people have been trying this 21st-century

© Noah Giansiracusa 2021
N. Giansiracusa, *How Algorithms Create and Prevent Fake News*,
https://doi.org/10.1007/978-1-4842-7155-1_5

algorithmic reinvention of the polygraph. How well it works and what it has been used for are the main questions explored in this chapter. To save you some suspense: this approach would create almost as much fake news as it would prevent—and claims to the contrary by the various companies involved in this effort are, for lack of a better term, fake news. But first, I'll start with the fascinating history of the traditional polygraph to properly set the stage for its AI-powered contemporary counterpart.

History of the Polygraph

The polygraph, known colloquially as a *lie detector*, has an interesting history[1] that reveals its awkward position in society—particularly American society—at the interstices between science and pseudoscience and between techno-optimism and chicanery.

Marston and His Invention

The psychologist and comic book author William Marston is most famous for two creations: the polygraph and Wonder Woman. The former was cocreated with his wife, Elizabeth Holloway—an accomplished lawyer and psychologist—while the latter was inspired in part by her. The story of this remarkable couple and their invention is worth telling. In 1915, while studying law and psychology as a graduate student at Harvard, Marston developed a theory that a person's blood pressure spikes when under the stress of answering a question deceitfully. He first stumbled upon this theory, supposedly, after Holloway insightfully yet offhandedly remarked that her blood pressure "seemed to climb" when "she got mad or excited." Marston proceeded to develop the first systolic blood pressure test, a key component of the couple's early lie detector prototype.

The United States entered World War I in 1917, and Marston saw an important application of their incipient technology: catching spies. He pitched this idea to various government officials, and he succeeded in convincing the National Research Council to form a committee to consider "the value of methods of testing for deception" that he proposed. Two weeks later, Marston enthusiastically wired to the committee chair the following brief note: "Remarkable results thirty deception tests under iron clad precautions letter following." The letter that followed elaborated the experiments Marston had conducted with colleagues. The first batch of subjects in these tests were primarily women at Harvard sororities, and a second batch of subjects came

[1] "The Polygraph and Lie Detection," *National Research Council*, 2003: https://www.nap.edu/catalog/10420/the-polygraph-and-lie-detection.

from the Boston Municipal Court. Based on Marston's optimistic news, the committee agreed to pursue his proposal further.

However, the report subsequently written by the chair of the committee was less sanguine than Marston. The report was strongly skeptical about the use of blood pressure tests, evidently due to earlier failed attempts by other researchers: "galvano-psychic and vaso-motor reactions [would] be more delicate indicators than blood pressure; but the same results would be caused by so many different circumstances, anything demanding equal activity (intellectual or emotional), that it would be practically impossible to divide any individual case." In other words, there were other biometric indicators that might better track deceit, but even those would only work in the aggregate; in any particular instance, one could not distinguish a lie from the truth because a spike in these indicators could be caused by numerous events, not just dishonesty. Rather presciently, after voicing this valid skepticism, the committee chair went on to suggest ways to modify Marston's experimental protocol in order to better protect against bias among the examiners administering the lie detector tests. This issue of examiner bias is a very significant one, and, as you shall soon see, it remains a thorn in the discipline to this day.

As Marston neared the completion of his studies at Harvard, his correspondence with the National Research Council turned sharply from simply requesting financial support for his research to securing employment directly within the government. This development appears to have been strongly motivated by the recognition that finishing his degrees meant no longer being a student—and hence being eligible for the wartime draft. Although he always envisioned himself as a university professor, a governmental research position was unquestionably more to his liking than armed service.

He was successful in this pursuit of government employment: by 1918, he was working for a medical support unit within the War Department (the more honestly named agency that in 1949 evolved into the Department of Defense). In this position, he continued his experiments on the lie detector, and he claims to have achieved ninety-seven percent accuracy in tests undertaken in the Boston criminal court using his systolic blood pressure device. He later wrote about using his device on spies during the late teens and throughout the 1920s, and in 1935 J. Edgar Hoover, as director of the FBI, officially inquired into Marston's work, but there are no surviving public details on Marston's work on espionage. Marston eventually segued back into academia and lived out his remaining years as a professor; this provided him the intellectual freedom to take his research on lie detection in any direction he wanted without having to convince superiors up the chain of command of the merits in doing so.

Into the Courtroom

A big moment—in Marston's life, in the history of the polygraph, and in the history of the American judicial system—came in 1923 when Marston served as an expert witness in the influential case *Frye v. United States*. James Frye, a veteran of WWI, was on trial for shooting and killing a prominent Washington DC physician, Dr. R. W. Brown. Frye admitted to robbing a traveling salesman, and supposedly while the police were questioning him on this matter, some incriminating details emerged linking him to the murder of the doctor. Frye soon thereafter confessed to shooting the doctor. He said he went to Dr. Brown's office for a prescription, but he only had one dollar and the prescription cost two. A newspaper article at the time described what happened next in Frye's confession:[2] "Dr. Brown, he said, declined to accept his pistol as collateral for the extra dollar. Trouble followed, and the physician, he declared, knocked him down, having followed him from the office to the hallway. It was while he was down, he stated, that he fired four or five shots." Not exactly self-defense, but not too far from it, at least in Frye's estimation.

However, Frye subsequently recanted this confession and instead professed his innocence. He said the reason for the false confession was that a detective had agreed to drop the robbery charges if Frye confessed to the murder and moreover that the detective would share with Frye a thousand dollar reward that the detective would collect for obtaining this confession. A cash reward and the dropping of a lesser charge don't seem like much incentive to confess to a murder that one did not commit, but Frye believed he had an alibi that would exonerate him from the murder charge. By his calculation, confessing to the murder was the smart thing to do: it would get him out of the robbery arrest, the murder charge would be dropped despite the confession once the alibi was corroborated, and he'd even pick up a few easy bucks in the process. Brilliant plan, except that the alibi failed. And so, according to his version of the story, when left facing a murder charge that he could no longer easily avoid, he admitted that the confession was nothing more than a fabrication from this failed scheme. (Never mind the fact that his confession included some strikingly accurate details about the crime scene that an outsider almost surely would not have known.)

Enter Marston, expert witness for Frye's defense, who believed he could establish Frye's innocence by using his lie detector to prove that Frye was being truthful when he explained why his confession was untruthful.[3] Marston

---

[2]Kenneth Weiss, Clarence Watson, and Yan Xuan, "Frye's Backstory: A Tale of Murder, a Retracted Confession, and Scientific Hubris," *Journal of the American Academy of Psychiatry and the Law*, June 2014, Volume 42 no. 2, pages 226–233: http://jaapl.org/content/42/2/226.

[3]Did you catch that? If it sounded too much like a line by Dr. Seuss, let me try again: if Marston could show that Frye was not lying about this failed scheme, then the detectives would be obliged to accept the retraction of the confession and enter Frye's plea of innocence.

used his device—which at the time was basically just a blood pressure monitor cuff and a stethoscope—on Frye and declared that he was telling the truth about conspiring with the detective. Marston then attempted to use this as official courtroom evidence by testifying on Frye's behalf.

However, the judge was unswayed and objected to the use of an unknown and unproven tool. The case was appealed up to the DC circuit court, which agreed with the trial court judge's skeptical view of Marston's device and testimony. The appellate ruling included a remark on the admissibility of expert witness testimony more generally, a remark that became known as the *Frye standard*, asserting that expert opinion is admissible if the scientific technique on which the opinion is based is "generally accepted" as reliable in the relevant scientific community. This general acceptance standard for admissibility of scientific evidence from the Frye case is still the verbatim law in some jurisdictions today, and even in jurisdictions where it is not, the law is essentially just a more detailed and elaborate version codified[4] in the so-called *Federal Rule 702*.

In short, while Marston hoped to make history by showing how his lie detection device could prove innocence, instead he made history by forcing the court system to articulate what kinds of expert testimony should *not* be allowed—and his landed squarely in this disallowed category. In fact, not only did his device fail the Frye standard at the time in 1923, but the Frye standard has kept polygraph tests, even in their more modern incarnations, out of the courtroom for nearly one hundred years now. (A swing and a miss there, perhaps, but he definitely hit a home run with his other enduring creation: the comic book super heroine he proposed to DC Comics in 1940 while working as a consultant for them—a character named *Wonder Woman* who was equipped with the "Lasso of Truth," a whip that forces anyone it ensnares to tell the truth. Wonder Woman drew inspiration from Marston's wife, Elizabeth Holloway, and the idea for the Lasso arose from their influential joint research into the psychology of human emotions.)

From Blood Pressure to Polygraph

Marston's embarrassing failure in the Frye case seems only to have galvanized him into developing his lie detector into a more sophisticated device. Rather than relying on systolic blood pressure alone and only taking measurements at discrete time intervals, he believed a more nuanced portrait would be provided by combining multiple measurements simultaneously and by continuously plotting their movements over time. This resulted in the *polygraph* (note that "poly" is the Greek root for "many," and "graph" is the

[4] https://www.law.cornell.edu/rules/fre/rule_702.

Greek root for "to write"). In addition to blood pressure, Marston's post-Frye polygraph measured breathing rate and sweatiness (the latter via skin conductance).

The modern polygraph is sometimes attributed to another inventor from around the same time on the opposite coast: Berkeley police officer and forensic psychologist John Augustus Larson, whose device also used a systolic blood pressure monitor and produced a continuous recording of the measurements. The details of who invented what and when are a little murky, and both these men based their inventions and ideas on earlier attempts (and, as mentioned earlier, Marston's work was really in collaboration with his wife). Whatever the case was back then, the relevant fact is that the polygraphs we know today—well, the ones prior to the recent AI-based systems that I'll soon discuss—are only small variants of these early 20th-century devices put forth by Marston and Larson.

Determined to Find a Use

What Marston perhaps lacked in scientific rigor, he made up for with commercial savvy: throughout his career, he worked hard to push his lie detector and touted its supposedly revolutionary success in a public advertising campaign and even in comic books. And he was quite successful in this endeavor. But if polygraphs are not admissible in court, what use has been found for them? Basically just one, but it is surprisingly large and lucrative. While it is illegal for most private companies to use polygraphs, many government agencies—local and federal—rely on polygraphs as part of their background employment screening process. It is estimated[5] that two and a half million polygraph tests are conducted annually in the United States, far more than in any other country, fueling a two-billion-dollar industry.

A 2007 survey found that around three-quarters of urban sheriff and police departments used polygraphs in their hiring process. They are commonly used when hiring firefighters and paramedics too. They are also part of the federal government security clearance process; I know this firsthand from a college summer internship I did at the NSA years ago. I still remember being nervous about having to fly across the country to be strapped into this strange machine in an austere examination room with multiple government officials, just like a scene from a Hollywood movie. I also remember the nervousness quickly fading when I realized the questions they were asking me were not difficult or embarrassing personal details—they were lobbing softballs such as: "Have you ever actively conspired to overthrow the US Government?"

[5]Mark Harris, "The Lie Generator: Inside the Black Mirror World of Polygraph Job Screenings," *Wired*, October 1, 2018: https://www.wired.com/story/inside-polygraph-job-screening-black-mirror/.

I wondered what this could possibly reveal, and who would ever fail this test. But they don't share the results of the test with you, they just incorporate the polygraph data into the overall background check that is kept confidential, and all you can do is trust the system. In my case, I passed the screening and received a security clearance for the summer, but to this day I have no idea what the polygraph supposedly indicated about me.

Enduring Skepticism

Let me repeat for emphasis: polygraphs are not reliable enough to be used in the courtroom nor by private companies, yet the public sector readily embraces the questionable technology and uses it to screen over two million people each year, most of them simply applying for jobs in which they could try to do good and have a positive impact on society. It is believed[6] that tens or hundreds of thousands of these applicants fail the polygraph each year and are thereby denied employment (exact figures are unknown). All this, despite the fact that throughout its hundred-year lifespan the polygraph has failed to establish its legitimacy in an accepted scientific fashion. In fact, in 1965, the US Committee on Government Operations evaluated the scientific evidence and reached the following conclusion:[7] "There is no lie detector, neither man nor machine. People have been deceived by a myth that a metal box in the hands of an investigator can detect truth or falsehood." In the 21st century, the *metal box* has been supplanted by the *black box*—which is to say, deep learning. This is the main topic I'll be coming to shortly.

But first, there's one more important point to make before concluding this background history of the old-school (but still widely used) polygraph device that is quite germane to the newfangled AI lie detectors as well. James Woolsey, former director of the CIA, warned during a 2009 interview:[8] "The polygraph's great flaw is the substantial number of false positives that it gives out, especially when you're using it for large-scale screenings." In other words, polygraphs have a tendency to misclassify truthful statements as lies, and even if the rate at which this occurs were relatively low (which it isn't necessarily), when used en masse as they are in employment screening, this means a large number of people are falsely and unfairly deemed liars. Indeed, Woolsey went on to express how this issue with false positives is "seriously damaging a lot of people's lives by having them fail the polygraph when they haven't really done anything."

---

[6]See Footnote 5.

[7]"Use of Polygraphs as 'Lie Detectors' by the Federal Government," H. Rep. No. 198, 89th Cong. 1st Sess.

[8]https://www.youtube.com/watch?v=bJ6Hx4xhWQs.

Aggravating this inequity further, false positives do not strike applicants uniformly. For instance, a racial discrimination class action lawsuit that was quietly settled in the 1980s revealed[9] that the high rate of polygraph failure among Black applicants to the Cook County Department of Corrections in the late 1970s had only a one in a thousand chance of happening randomly—meaning the polygraph results were almost certainly biased against Black applicants. (Part of the settlement agreement was for Cook County to immediately stop using polygraph tests for employment screening.) Similarly, in 1990, the US Department of Defense conducted a study on polygraph reliability and found that under simulated criminal proceedings, innocent Black people were more likely to receive false positives than innocent white people.

It is not entirely known what accounts for this discrepancy, but one explanation put forth by experts is that the polygraph's readings are so vague and open to interpretation that they in essence act as a Rorschach test for the examiner and are thereby subject to their human whims and biases. One neuroscientist said,[10] "One examiner might see a blood pressure peak as a sign of deception, another might dismiss it—and it is in those individual judgments that bias can sneak in. The examiner's decision is probably based primarily on the human interaction that the two people have." And a senior policy analyst for the ACLU expanded on this point:[11] "In this respect polygraphs are just like other pseudo-scientific technologies that we've seen in recent years: Because they are fundamentally bogus, they end up becoming no more than a vehicle for operators to substitute their own personal assessments of subjects in the absence of genuinely useful measurements. For some operators, that's inevitably going to mean racial bias."

It is quite possible that structural racism plays a role too by, for instance, causing Black people to be more nervous during governmental interrogations. It is rather surprising how little research has been conducted on bias in polygraphs, especially considering how widespread their use is in the public sector.

Now our history of the traditional polygraph is complete, and, with the stage properly set, we step into the world of algorithmic lie detection.

The Polygraph Meets AI

As you have already witnessed in the chapter on deepfakes, deep learning has powered a revolution in our ability to process visual data. At first glance, lie detection seems like it should be a rather straightforward application: train a

[9] See Footnote 5.
[10] See Footnote 5.
[11] Jay Stanley, "How Lie Detectors Enable Racial Bias," *ACLU blog*, October 2, 2018: https://www.aclu.org/blog/privacy-technology/how-lie-detectors-enable-racial-bias.

machine learning algorithm on the supervised binary classification task of sorting video clips into truthful versus untruthful. It turns out, however, that the problem of biased false positives is a ghost in the machine that is not so easily vanquished.

Lies in the Eyes

In 2014, a startup funded by Mark Cuban called *Converus*[12] released a product called *EyeDetect*, pitched as a faster, cheaper, and more accurate alternative to the polygraph. It is the main product from this company with a self-described "vision to provide trustworthy, innovative solutions for the deception detection industry." The hype and fanfare quickly led to fairly widespread adoption, primarily though not exclusively for public sector employment screening. According to the company website, by January 2019, EyeDetect had been sold to over five hundred clients in forty countries; in the United States, these clients included the federal government and twenty-one state and local agencies.

As the name suggests, EyeDetect relies not on blood pressure or skin conductivity or respiration rates like the traditional polygraph; instead, its focus is on the windows to the soul: the eyes. Perhaps an even more significant difference is that, in stark contrast to traditional polygraphs, EyeDetect does not involve a human examiner to interpret the readings and decide what is a lie and what is truthful—EyeDetect reaches its conclusions in an automated, algorithmic manner by applying machine learning. Indeed, EyeDetect was fed close-up video footage of subtle eye movements for participants who were telling the truth and also for participants who were lying, and the algorithm used this as training data to determine what honesty and dishonesty look like in the eyes.

Taking the human interpreter out of the equation certainly helps to create an air of impartiality, but does this algorithmic approach actually yield reliable and unbiased results? Sadly, no. The past several years have taught us that algorithmic bias is a serious and fundamental issue in machine learning. Not only do algorithms absorb bias that inadvertently creeps into data sets used for training, but the algorithms reproduce and often amplify this bias. Much has been written about this pernicious data-driven feedback loop phenomenon in general,[13] and I'll return to it here in the specific setting of lie detection shortly.

---

[12]This name is Latin for "with truth," though perhaps an unfortunate choice in our current time of coronavirus pandemic.

[13]For book-length treatments of the topic, I recommend Cathy O'Neil's 2016 *New York Times* best seller *Weapons of Math Destruction* and Virginia Eubanks' 2018 title *Automating Inequality*.

According to a 2018 investigation[14] by *Wired*, the Department of Defense and US Customs and Border Protection have been trialing Converus' technology, and while federal law prohibits most private companies from using any kind of lie detection device for employee screening on American soil, evidently a handful of companies including FedEx and McDonald's and Uber have used EyeDetect in Guatemala and Panama and Mexico. The credit rating agency Experian uses it on its staff in Colombia to try to prevent employees from manipulating records in order to help fraudulently secure loans for family members. Converus said an unnamed Middle Eastern country had purchased EyeDetect to screen people entering the country for possible terrorist activity/affiliations.

Converus claims its system attains an eighty-six percent accuracy rate, better than the roughly seventy percent estimated for traditional polygraphs (the company goes so far as to assert that EyeDetect is "the most accurate lie detector available"). However, the *Wired* investigation points out that "The only peer-reviewed academic studies of Converus' technology have been carried out by the company's own scientists or students in their labs," which is not particularly reassuring and screams of an obvious conflict of interest. Eyebrows are usually raised when a private company funds research into the efficacy of a product that the company aims to profit from—but here the company didn't just fund the research, it conducted the research itself behind closed doors. John Allen, a psychology professor not involved with Converus, was asked by *Wired* to read a couple of the company's academic papers in order to try to assess the situation. This is what he had to say: "My kindest take is that there is some promise, and that perhaps with future independent research this test might provide one measure among many for formulating a hypothesis about deceptive behavior. But even that would not be definitive evidence." Not exactly a glowing recommendation.

And these academic papers only cover the more successful experiments conducted by Converus; the company's first field test revealed a glaring weakness in the system, yet the results of this experiment were never published. The chief scientist at Converus, who is also the cocreator of EyeDetect, later admitted what happened during this first field test: "Although the data were limited, the [test] appeared to work well when we tested well-educated people who had applied to work for an airline, but the [test] was ineffective when we tested less well-educated applicants for security companies." This remark very much suggests that the machine learning algorithms powering EyeDetect were trained on a highly selective and nonrepresentative sample of the population, which is a common recipe in the

[14]Mark Harris, "An Eye-Scanning Lie Detector Is Forging a Dystopian Future," *Wired*, December 4, 2018: https://www.wired.com/story/eye-scanning-lie-detector-polygraph-forging-a-dystopian-future/.

algorithmic world for pernicious bias that disproportionately harms underprivileged populations.

At an even more fundamental level, the chief scientist's remark raises a striking question that the company seems to have left unanswered: why would one's visual indicators of deceit depend on one's level of education? The mythology of lie detection is that efforts to conceal deception are innate and universal, unvarying across populations—yet this failed field test shows that this is not at all the case. This observation should have rattled the very foundations of Converus' endeavor, but instead the company seems to have just swept it under the rug and threw more data and more neural network layers at the problem. As a further indication of problematic non-universality, consider the following memo (also revealed in the *Wired* investigation) that a Converus marketing manager wrote to a police department client in 2016: "Please note, when an EyeDetect test is taken as a demo […] the results are often varied from what we see when examinees take the test under real test circumstances where there are consequences." Something is very fishy here—and it gets worse.

By design, the EyeDetect system allows the examiner to adjust the *sensitivity*, meaning the threshold at which the algorithm declares a lie to have been detected. The idea behind this is that certain populations historically might be more truthful than others, so the system will produce more accurate results if it is calibrated to the population baseline level when examining each individual. In the words of the president and CEO of Converus, Todd Mickelsen: "This gives all examinees a fairer chance of being classified correctly. Most organizations can make good estimates of base rates by considering the number of previously failed background checks, interview data, confessions, evidence, etc." Really?

First of all, no, the kind of data Mickelsen describes is not available to most organizations. Second, the data he describes is already tainted by bias. For example, Mickelsen suggests using base rates based on historical rates of failed background checks, but what was involved in those historical background checks? Probably a combination of old-school polygraphs (known to be biased against minority populations, as you have already seen, for instance, in the Cook County case) and old-fashioned private-eye type investigations (which nobody would argue are free from bias). This statement by the president and CEO of Converus is in effect an admission that EyeDetect encourages examiners to embed racism and other forms of discrimination from the past into this futuristic technology. And even if the examiner does not rely on biased historical data, Mickelsen seems to be encouraging them to introduce their own personal contemporary bias—just crank up the sensitivity whenever you're examining someone from a population you don't trust! The same senior policy analyst with the ACLU that you heard from earlier excellently

summed up the present situation:[15] "The criticism of technologies like lie detectors is that they allow bias to sneak in, but in this case it sounds like bias isn't sneaking in—it's being welcomed with open arms and invited to stay for dinner."

So, EyeDetect is not scientifically proven in any kind of traditional way that is free from conflict of interest, and its design includes a lever allowing the examiner to skew the results in either direction. On top of this, it is a closed system using a proprietary algorithm—which itself is based on black-box machine learning—so it is essentially impossible to scrutinize the inner workings of the system. As faulty as the old-school polygraph is, at least we know what exactly it is measuring and how those measurements are supposedly interpreted, and there is at least a theory (albeit a flawed one) trying to explain why those measurements correlate with concealed deceit. With EyeDetect, on the other hand, a computer decides which eye movements it considers indicators of deceit, and there is no explanation of what they are and why they indicate deceit, other than that's what the computer found in past data—at least for the small and nonrepresentative population that it was trained on.

The *Wired* investigation astutely points out yet another troublesome issue with EyeDetect: "Its low price and automated operation also allow it to scale up in a way that time-consuming and labor-intensive polygraph tests never could." If it worked perfectly and had no harmful consequences, then scaling up would be great—but given the many flaws already discussed, scaling up is very dangerous. In fact, this is where pernicious data-driven feedback loops come into play, as I next explain.

It is widely recognized now that facial recognition software trained on one racial population does not perform well on other populations, and essentially all machine learning algorithms developed in the United States perform worse on Black and Brown faces than on white faces.[16] I would be shocked if EyeDetect were somehow an exception to this pattern, which means EyeDetect very likely produces more false positives for Black and Brown examinees than it does for white examinees. When EyeDetect is then used en masse for employment screenings, Black and Brown people in the aggregate are unfairly kept out of the workforce. Denying these populations jobs exacerbates the already significant racial wealth gap in the United States, pushing more Black and Brown people into poverty.

But the story doesn't end there. Police are known to more actively patrol impoverished communities, especially ones with high proportions of Black and

[15]See Footnote 14.

[16]"NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software," *NIST*, December 19, 2019: https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software.

Brown residents, compared to wealthy white communities.[17] So in the aggregate, even if just by a small amount, by making it harder for Black and Brown people to land jobs, EyeDetect is pushing these populations into environments where they are more likely to get arrested. Now here's the real kicker: this higher arrest rate leads to a higher failure rate for background checks (even the old-fashioned kind, because arrest records are one of the main tools used for those), which in turn boosts the "base rates" Mickelsen mentioned that are used to adjust the sensitivity in EyeDetect. The horrible irony is that this just further exacerbates the racial discrepancy in EyeDetect's output.

Did you catch that all? In summary, and in the aggregate, EyeDetect gives Black and Brown people more false positives, which keeps them out of jobs, which pushes them into poverty and highly policed neighborhoods, which leads to increased arrest records, which leads to failed background checks, which leads to a recalibration of EyeDetect that leads to even more false positives for these populations. In this way, the harmful discriminatory cycle repeats over and over, intensifying as it goes. You might think this is a stretch—there's so many steps in this process, and so many mentions of aggregate behavior rather than individuals—but alas, in the past few years we've learned that these pernicious data-driven feedback loops are very real and very dangerous. If you don't believe, please take a look at the books in Footnote 13 or any of the other excellent writing on the topic.

Despite pretty clearly not passing the bar of the Frye general acceptance standard, in May 2018 EyeDetect was used in a court case for the first time, in a New Mexico district court in the trial of a former high school coach accused of raping a fourteen-year-old girl (the jury failed to agree on a verdict).[18] And in April 2019, Converus published a blog post[19] on the company website encouraging President Trump to administer EyeDetect tests on the entire White House staff and explaining the logistics of how he could do this, in an effort to help clamp down on embarrassing leaks. I suppose I don't need to point out the irony in a company called "with truth" that works in the "deception detection industry" assisting a president who has been, well, let's just say less than honest, conceal his secrets from the public.

---

[17]This was true historically, and it might be even more true today because of "predictive policing" which is one of the most devastating known instances of a pernicious data-driven algorithmic feedback loop. See, e.g., Karen Hao, "Police across the US are training crime-predicting AIs on falsified data," *MIT Technology Review*, February 13, 2019: https://www.technologyreview.com/2019/02/13/137444/predictive-policing-algorithms-ai-crime-dirty-data/ and Will Heaven, "Predictive policing algorithms are racist. They need to be dismantled." *MIT Technology Review*, July 17, 2020: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/.

[18]See Footnote 14.

[19]Eliza Sanders, "The Logistics of Lie Detection for Trump," *Converus blog*, April 5, 2019: https://converus.com/blog/the-logistics-of-lie-detection-for-trump/.

Unsurprisingly, Converus was not the only organization with the idea of reinventing the lie detector through machine learning.

Deep Learning Micro-Gestures

In the early 2000s, before neural networks had become the deep learning revolution that they are today, a PhD student at Manchester Metropolitan University, Janet Rothwell, and her doctoral adviser in the computer science department, Zuhair Bandar, trained a neural network lie detection algorithm on a small number of video clips of people answering questions honestly and dishonestly. The Rothwell-Bandar algorithm got around eighty percent accuracy in simple tests with highly idealized data (the lighting was identical in all instances, and if someone was wearing glasses, then the algorithm became completely flummoxed). Somewhat promising but far from convincing. To Rothwell's surprise, her university put out a press release touting her project as a new invention that would render the polygraph obsolete. Rothwell left the project and the university in 2006 and moved on to other things.

Bandar, on the other hand, continued to develop the project for many years with two new students, and in January 2019, the trio launched a startup based on the technology called *Silent Talker*. This was one of the first lie detectors on the market powered by deep learning. One of the cofounders, Jim O'Shea (who followed in the footsteps of his adviser and is now a senior lecturer at Manchester Metropolitan University), proudly admitted[20] the black-box nature of their product: "Psychologists often say you should have some sort of model for how a system is working, but we don't have a functioning model, and we don't need one. We let the AI figure it out." As recently as March 2020, Bandar said that his company was in talks to sell the technology to law firms, banks, and insurance companies—for employment screening, as usual, but also for fraud detection. O'Shea said it could also be used in employee assessment.[21]

---

[20] Jake Bittle, "Lie detectors have always been suspect. AI has made the problem worse." *MIT Technology Review*, March 13, 2020: https://www.technologyreview.com/2020/03/13/905323/ai-lie-detectors-polygraph-silent-talker-iborderctrl-converus-neuroid/.

[21] Incidentally, the first field study the team published, back in 2012, used the technology not to detect lies but to measure comprehension: in collaboration with a healthcare NGO in Tanzania, the facial expressions of eighty women were recorded while they took online courses on HIV treatment and condom use, and the system was able to predict with around eighty-five percent accuracy which of them would pass a brief comprehension test. See Fiona Buckingham et al., "Measuring human comprehension from nonverbal behavior using Artificial Neural Networks," *The 2012 International Joint Conference on Neural Networks (IJCNN)*, Brisbane, QLD, Australia, 2012: https://ieeexplore.ieee.org/abstract/document/6252414.

The *MIT Technology Review* reported[22] that in 2018 the technology that would one year later form the basis for Silent Talker was involved in an experimental initiative called *iBorderCtrl* that was funded by the European Union and tested on volunteers at borders in Greece, Hungary, and Latvia. A press release[23] announcing this experimental venture explained that it "is aiming to deliver more efficient and secure land border crossings to facilitate the work of border guards in spotting illegal immigrants, and so contribute to the prevention of crime and terrorism." The press release goes on to provide more details: "Travelers will use an online application to upload pictures of their passport, visa and proof of funds, then use a webcam to answer questions from a computer-animated border guard, personalized to the traveler's gender, ethnicity and language. The unique approach to 'deception detection' analyses the micro-gestures of travelers to figure out if the interviewee is lying." (Personalized to the traveler's ethnicity? I can't imagine how bias might creep into a system like this…) At the border, travelers flagged by the system as high risk undergo a more detailed—and traditional—check.

The *MIT Technology Review* report notes that after this 2018 iBorderCtrl announcement, "activists and politicians decried the program as an unprecedented, Orwellian expansion of the surveillance state." A Dutch member of the European Parliament and leader of a center-left party warned the European Commission that this is "part of a broader trend towards using opaque, and often deficient, automated systems to judge, assess, and classify people." The European Commission seems to have taken the hint and rebranded the venture from a practical pilot to a more theoretical research project, and one official said the deception detection system "may ultimately not make it into the design."

In 2019, journalists at the *Intercept* were able to try out the iBorderCtrl system for themselves, while crossing the Serbian-Hungarian border.[24] They were asked sixteen questions and gave honest answers to all of them, yet the Silent Talker–based system scored four of these as lies—resulting in an overall assessment that the traveler was untruthful and required further questioning in person. Clearly a false positive. Ordinarily, the traveler is not informed of the lie detector's report, but the journalists here obtained it through the European analogue of a Freedom of Information Act request. Earlier in the year, scholars at a digital human rights center in Milan used the same legal mechanism to request internal documents about iBorderCtrl's lie detector

[22]See Footnote 20.

[23]"Smart lie-detection system to tighten EU's busy borders," *European Commission*, October 24, 2018: https://ec.europa.eu/research/infocentre/article_en.cfm?artid=49726.

[24]Ryan Gallagher and Ludovica Jona, "We Tested Europe's New Lie Detector For Travelers—And Immediately Triggered a False Positive," *The Intercept*, July 26, 2019: https://theintercept.com/2019/07/26/europe-border-control-ai-lie-detector/.

system, but what they received was "heavily redacted, with many pages completely blacked out." One of the scholars responded[25] with distrust and frustration: "What is written in those documents? How does the consortium justify the use of such a pseudoscientific technology?"

It turns out Silent Talker is not the only startup doing AI-powered lie detection for airport screenings. A company called *Discern Science International* (DSI) that launched in 2018 has a product called *Avatar* that provides a very similar service as the one in iBorderCtrl: a virtual border guard asks travelers a prerecorded set of questions, and the system captures the traveler's answers (both video and audio) and uses machine learning to label each answer as honest or dishonest. The company says the system looks for "deception signals" in the voice and face, such as involuntary "microexpressions" that supposedly are triggered by the cognitive stress of lying.

Avatar is the commercialization of a research project undertaken by academics at the University of Arizona over the past several years. Prototypes are known[26] to have been tested at an airport in Romania and a US border port in Arizona. It was also tested by the Canada Border Services Agency, but only in a laboratory setting, and the official response was less than enthusiastic: "a number of significant limitations to the experiment in question and to the technology as a whole led us to conclude that it was not a priority for field testing." DSI says these tests yielded accuracy rates between eighty and eighty-five percent. While certainly better than random guessing, that sure leaves a lot of incorrect assessments in the field. Nonetheless, it was reported[27] in August 2019 that Discern had struck a partnership with an unnamed but well-established aviation organization and was planning on marketing Avatar to airports in a matter of months. DSI's website currently says that the "Initial markets for the application of the deception detection technology will be at airports, government institutions, mass transit hubs, and sports stadiums."

From Video to Audio and Text

While Silent Talker relies on video data, and Avatar relies simultaneously on video and audio data, you might be wondering if AI-powered lie detection has been attempted without the visual component. Indeed, it has. *Nemesysco* is an Israeli company offering commercial AI voice analysis software that has been used by police departments in New York and the Midwest to interview suspects and also by debt collection call centers. Another startup, called *Neuro-ID*, doesn't even have to see or hear someone at all—instead, it focuses on mouse movements and keystrokes. It has been used by banks and insurance

[25]See Footnote 24.
[26]Camilla Hodgson, "AI lie detector developed for airport security," *Financial Times*, August 2, 2019: https://www.ft.com/content/c9997e24-b211-11e9-bec9-fdcab53d6959.
[27]See Footnote 26.

companies to help detect fraud. As always with this kind of thing, false positives are a serious issue. A Neuro-ID spokesperson clarified[28] the intended use of this product: "There's no such thing as behavior-based analysis that's 100% accurate. What we recommend is that you use this in combination with other information about applicants to make better decisions and catch [fraudulent clients] more efficiently."

An academic collaboration produced a paper[29] in February 2019 claiming to provide the first steps toward an "online polygraph system—or a prototype detection system for computer-mediated deception when face-to-face interaction is not available." The idea was to train a machine learning algorithm to distinguish lying from truth-telling in the context of a live text chat between two people. The algorithm relied on not just the words within the text messages but also the rate at which they were typed. I would have guessed that deception requires more thought and therefore is manifest in a slower response time, but the authors of this paper claim that lying correlated with a faster response rate; perhaps lying does take more time, but the liars were aware of this and so answered more quickly in order to compensate for this in an attempt to hide their deception. The authors also found that liars had more verbal signs of anxiety in their communication, more negative emotions, a greater volume of words in each response, and more expressions of certainty such as "always" and "never." Honest answers, in addition to being slower, shorter, less negative/anxious, and less certain (involving words like "perhaps" and "guess"), also used more causal expressions such as "because." Overall, the algorithm scored an eighty-two percent accuracy—once again, better than random guessing but not enough to rely on in practice, in my opinion.

Moreover, this online polygraph was trained and evaluated in a very controlled, limited, and artificial setting. A few dozen participants took part in a text chat game in which they were split into pairs, and then each pair conversed by asking and answering questions. The main rule for this game was that at the outset of each conversation, each individual was told that they must either answer all the questions honestly or answer them all dishonestly. The algorithm, therefore, was not attempting to classify individual answers as truthful versus deceptive, it was classifying the participants in each conversation as habitual truth-tellers versus habitual liars. This makes a big difference in terms of the accuracies one expects, and it strongly signals a simulated environment that would never actually occur in reality. Not only that, but as

[28]See Footnote 20.

[29]Shuyuan Mary Ho and Jeffrey Hancock, "Context in a bottle: Language-action cues in spontaneous computer-mediated deception," *Computers in Human Behavior* Vol 91, February 2019, 33–41: https://sml.stanford.edu/pubs/2019/context-in-a-bottle/.

the prominent data scientist Cathy O'Neil pointed out,[30] the behavior of someone instructed to lie in a lab setting is very different than that of a practiced liar in the real world with skin in the game. She bluntly called this a "bad study" that has "no bearing" on being able to catch a seasoned liar in the act. In a similar vein, Kate Crawford, cofounder of the AI Now Institute at New York University, noted that this experiment was detecting "performance" rather than authentic deceptive behavior.

Fake News

Imagine having an algorithm that in real time would tell you whether someone was lying. This would be a tremendous weapon against fake news and disinformation. A shocking claim by a YouTube personality could immediately be outed as a fabricated conspiracy theory; an assertion by a politician during a press conference or debate could be revealed as a falsehood the moment it is uttered; eyewitness testimony could be vetted and verified; your questionable friend on Facebook who sends you deceptive chat messages about controversial political events would be caught in each act of dishonesty. Alas, such an algorithm is an unattainable fantasy. While AI has rejuvenated and revitalized the deception detection industry by providing modern variants of the polygraph based on video, audio, or text, there is no escaping the fundamental truth that there is no science to lying. Lies are unique, unrecognizable, and unpredictable—and that's when they are deliberate; unintentional falsehoods clearly have zero chance of detection by any of the methods discussed in this chapter.

As you have seen throughout this chapter, the bold claims in academic studies and corporate websites about the power of various AI-based lie detection algorithms are, simply put, mostly just fake news. Essentially, all the empirical investigations into each of these lie detection methods have been conducted behind closed doors by the organization with a direct financial incentive in the method performing well. The experiments tend to be artificial and limited in scope, and the reported accuracies paint a misleadingly rosy picture of what is possible. The training data sets for these products are almost certainly all too small and biased either by historical prejudice or by limited exposure to diverse people/situations (or both). Moreover, the black-box nature of machine learning algorithms means that nobody really knows why an AI lie detection system works as it does, nor what it is actually doing.

The 1923 Frye case essentially sealed the fate of the polygraph in the courtroom, and subsequent legislation drastically limited its corporate

[30] Andy Greenberg, "Researchers Built an 'Online Lie Detector.' Honestly, That Could Be a Problem." *Wired*, March 21, 2019: https://www.wired.com/story/online-lie-detector-test-machine-learning/.

usage as well (though public sector employment screening is a gaping hole in the regulatory web). It takes time for the legal system to catch up to the fast-paced world of technology. We are at a dangerous period of history now where the polygraph has been reinvented by AI and is quickly spreading across many sectors of society before the definitive Frye moment where the brakes are applied to rein in the ill-conceived and overzealous applications of a highly lucrative but unproven technology rooted in pseudoscience. As Vera Wilde, an academic and privacy activist who helped start the public campaign against iBorderCtrl, put it:[31] "It's the promise of mind-reading. You can see that it's bogus, but that's what they're selling."

Sigmund Freud once wrote that "No mortal can keep a secret. If his lips are silent, he chatters with his fingertips. Betrayal oozes out of him at every pore." But Dan Ariely, a behavioral psychologist at Duke University, pointed out[32] that "We have this tremendous capacity to believe our own lies. And once we believe our own lies, of course we don't provide any signal of wrongdoing." Shortly before he died in 1965, Larson, the Berkeley police cocreator of the original polygraph whom you briefly met earlier in this chapter, left an ominous warning about his invention that shows more circumspection than his monomaniacal fellow inventor Marston: "Beyond my expectation, through uncontrollable factors, this scientific investigation became for practical purposes a Frankenstein's monster." This monster breathes new life today in the age of AI.

Summary

The polygraph has a long and winding history, starting with work of Marston (the creator of Wonder Woman) and his wife Holloway, and also a Berkeley police officer named Larson, that took place between 1915 and 1921. Marston convinced the government to investigate the efficacy of his invention, but the official response was skepticism rooted in common sense and historical insight. Determined to make a revolutionary impact, Marston attempted to use his device to establish the innocence of a defendant in a 1923 murder trial, but his efforts were dismissed by the court and resulted instead in the Frye standard that still stands as the law today: expert witness testimony is admissible in court only if the technology it is based on is generally accepted by the scientific community. Polygraphs did not pass that test then, and neither do the new AI-powered algorithmic variants today.

---

[31] See Footnote 20.

[32] Amit Katwala, "The race to create a perfect lie detector—and the dangers of succeeding," *Guardian*, September 5, 2019: https://www.theguardian.com/technology/2019/sep/05/the-race-to-create-a-perfect-lie-detector-and-the-dangers-of-succeeding.

Nevertheless, interest in AI lie detection has surged in the past few years. Some methods rely on video, others audio, and others text alone. They all suffer from a lack of transparency, exaggerated claims of accuracy, an unnervingly high rate of false positives, and bias that disproportionately impacts minority populations. This has not stopped them from being used for employment screening and fraud detection and occasionally even in the courtroom (despite the Frye standard), and from being trialed in airport security and other settings. Given all the flaws, overzealous commercialization, corporate secrecy, and embarrassing lack of even an attempt at scientific foundations, it does not appear that the polygraph—even when reinvented with AI—will ever be able to detect lies with the consistency needed to rein in fake news. Instead, the mythical ability to use fancy technology to peer into the mind and reliably detect deception is itself the fake news in this story.

If we can't use algorithmic lie detectors to unmask fake news and get to the truth in controversial matters, perhaps we should just do what hundreds of millions of people do every day: Google it. But be careful—there too the algorithms behind the scenes systematically distort our perception of reality, as you will see in the next chapter.

CHAPTER

6

Gravitating to Google

The Dangers of Letting an Algorithm Answer Our Questions

> *Search engines have come to play a central role in corralling and controlling the ever-growing sea of information that is available to us, and yet they are trusted more readily than they ought to be. They freely provide, it seems, a sorting of the wheat from the chaff, and answer our most profound and most trivial questions. They have become an object of faith.*
>
> —Alex Halavais, *Search Engine Society*

Billions of people turn to Google to find information, but there is no guarantee that what you find there is accurate. As awareness of fake news has risen in recent years, so has the pressure on Google to find ways of modifying its algorithms so that trustworthy content rises to the top. Fake news is not limited to Google's main web search platform—deceptive and harmful content also play a role on other Google products such as Google Maps, Google News, and Google Images, and it also shows up on Google's autocomplete

© Noah Giansiracusa 2021
N. Giansiracusa, *How Algorithms Create and Prevent Fake News*,
https://doi.org/10.1007/978-1-4842-7155-1_6

tool that feeds into all these different products. In this chapter, I'll look at the role fake news plays in all these contexts and what Google has done about it over the years. In doing so, I'll take a somewhat more expansive view of the term "fake news" compared with previous chapters to include hateful racist stereotypes and bigoted misinformation.

Setting the Stage

On the morning of November 14, 2016, six days after the US presidential election in which Trump won the electoral college and Clinton won the popular vote, both by relatively wide margins, the top link in the "In the news" section of the Google search for "final election results" was an article asserting that Trump had won the popular vote by seven hundred thousand votes.[1] It was from a low-quality WordPress blog that cited Twitter posts as its source, yet somehow Google's algorithms propelled this fake news item to the very top. In response to this worrisome blunder, a Google spokesperson said:[2] "The goal of Search is to provide the most relevant and useful results for our users. We clearly didn't get it right, but we are continually working to improve our algorithms."

The next day, Sundar Pichai—just one year into his role as CEO of Google—was asked in an interview[3] with the BBC whether the virality of fake news might have influenced the outcome of the US election. Mark Zuckerberg had already dismissed this idea (naively and arrogantly, it appears in hindsight) as "pretty crazy," whereas Pichai was more circumspect: "I am not fully sure. Look, it is important to remember this was a very close election and so, just for me, so looking at it scientifically, one in a hundred voters voting one way or the other swings the election either way." Indeed, due to the electoral college, the election came down to just one hundred thousand votes. When asked specifically whether this tight margin means fake news could have potentially played a decisive role, Pichai said, after a pause: "Sure. You know, I think fake news as a whole could be an issue."

---

[1]Philip Bump, "Google's top news link for 'final election results' goes to a fake news site with false numbers," *Washington Post*, November 14, 2016: https://www.washingtonpost.com/news/the-fix/wp/2016/11/14/googles-top-news-link-for-final-election-results-goes-to-a-fake-news-site-with-false-numbers/.

[2]Richard Nieva, "Google admits it messed up with fake election story," *CNET*, November 14, 2016: https://www.cnet.com/news/google-fake-news-election-donald-trump-popular-vote/.

[3]Kamal Ahmed, "Google commits to £1bn UK investment plan," *BBC News*, November 15, 2016: https://www.bbc.com/news/business-37988095.

Less than a year later, Eric Schmidt, then the executive chairman of Alphabet, Google's parent company, publicly admitted[4] that Google had underestimated the potential dedication and impact of weaponized disinformation campaigns from adversarial foreign powers: "We did not understand the extent to which governments—essentially what the Russians did—would use hacking to control the information space. It was not something we anticipated strongly enough." He made that remark on August 30, 2017.

One month and two days later, on October 1, 2017, the worst mass shooting in modern US history took place in Las Vegas. Within hours, a fake news item was posted on the dubious website 4chan, in a "politically incorrect" channel associated with the alt-right, falsely accusing a liberal man as the shooter. Google's algorithm picked up on the popularity of this story, and soon the first result in a search for the name of this falsely accused man was the 4chan post—which was misleadingly presented as a "Top story" by Google. The response[5] from a Google spokesperson was unsurprisingly defensive and vague: "Unfortunately, early this morning we were briefly surfacing an inaccurate 4chan website in our search results for a small number of queries. […] This should not have appeared for any queries, and we'll continue to make algorithmic improvements to prevent this from happening in the future." Google's wasn't the only algorithm misfiring here: Facebook's "Trending Topic" page for the Las Vegas shooting listed multiple fake news stories, including one by the Russian propaganda site *Sputnik*.[6] Schmidt's remark from a month earlier about Russian interference was oddly prescient—or frustratingly obvious, depending on your perspective.

One and a half months later, at an international security conference, Schmidt tried to explain[7] the challenge Google faces when dealing with fake news: "Let's say this group believes fact A, and this group believes fact B, and you passionately disagree with each other and you're all publishing and writing about it and so forth and so on. It's very difficult for us to understand

[4]Austin Carr, "Alphabet's Eric Schmidt On Fake News, Russia, And 'Information Warfare'," *Fast Company*, October 29, 2017: https://www.fastcompany.com/40488115/alphabets-eric-schmidt-on-fake-news-russia-and-information-warfare.

[5]Gerrit De Vynck, "Google Displayed Fake News in Wake of Las Vegas Shooting," *Bloomberg*, October 2, 2017: https://www.bloomberg.com/news/articles/2017-10-02/fake-news-fills-information-vacuum-in-wake-of-las-vegas-shooting.

[6]Kathleen Chaykowski, "Facebook And Google Still Have A 'Fake News' Problem, Las Vegas Shooting Reveals," *Forbes*, October 2, 2017: https://www.forbes.com/sites/kathleenchaykowski/2017/10/02/facebook-and-google-still-have-a-fake-news-problem-las-vegas-shooting-reveals/.

[7]Liam Tung, "Google Alphabet's Schmidt: Here's why we can't keep fake news out of search results," *ZDNET*, November 23, 2017: https://www.zdnet.com/article/google-alphabets-schmidt-heres-why-we-cant-keep-fake-news-out-of-search-results/.

truth. […] It's difficult for us to sort out which rank, A or B, is higher." He went on to explain that it is easier for Google to handle false information when there is a large consensus involved. A fair point in some respects, but it's hard to imagine how this applies to these past debacles—was there not a consensus in the 2016 election that Trump lost the popular vote, and in the Las Vegas shooting that an unsubstantiated rumor on an alt-right site was not actual news? What about just a few months later, in February 2018, when the top trending video on YouTube (which, as you recall from Chapter 4, is owned by Google and which has essentially taken over the video search portion of Google) was[8] an egregious conspiracy theory claiming that some survivors of the Parkland, Florida, high school shooting were actors?

If there really was a lack of "consensus" in these incidents, one has to wonder whether that was actually the cause of the problems with Google's algorithm as Schmidt suggested or whether he perhaps had it backward. Maybe the fact that Google's algorithm has propped up fake stories like these, thereby lending them both legitimacy and a vast platform, caused some of the erosion of truth that ultimately led to a lack of consensus on topics that should not have been controversial in the first place. In other words, did Google reflect a state of confusion, or did it cause one? In all likelihood, the answer is a combination of both. To start unravelling this complex issue, it helps to separate out the different services Google provides so that we can delve into the algorithmic dynamics underlying each one and explore the deceptive and hateful content that has surfaced on each one.

Throughout this chapter, I shall use the term "fake news" more broadly to include racist and bigoted content. I have largely resisted doing so in the book thus far because so much has been written on algorithmic bias already, so rather than overcrowding these chapters by retelling that tale, I prefer to encourage you to consult the excellent and rapidly developing literature on the matter. But when it comes to Google, which is such an intimate and immediate source of information for so many people, I cannot earnestly disentangle news-oriented disinformation from socially oriented disinformation of the kind found in racism, sexism, anti-Semitism, etc. For one thing, many fake news sites align with the white supremacist–leaning alt-right, so when Google feeds its users bigoted information, it is also priming them to fall for hardcore alt-right fake news material. And at a more philosophical level, one could argue that racist stereotypes are a form of fake news—they are in essence harmful disinformation that happens to focus on certain populations.

[8]Sara Salinas, "The top trending video on YouTube was a false conspiracy that a survivor of the Florida school shooting was an actor," *CNBC*, February 21, 2018: https://www.cnbc.com/2018/02/21/fake-news-item-on-parkland-shooting-become-top-youtube-video.html.

Google Maps

One of the most abhorrent examples of hateful disinformation on a Google platform occurred in May 2015, during President Obama's second term in office. It was reported[9] in the *Washington Post* that if one searched Google Maps for "N****r king" or "N***a house" (with the asterisks filled in), the map would locate and zoom in on the White House. This was not the result of algorithmic bias or some other subtle failure of AI, it was directly the result of racist users with malicious intent—or as some people call it, third-party trolling and vandalism. This, and other acts of vandalism, caused Google to suspend user-submitted edits to Google Maps at the time: "We are temporarily disabling editing on Map Maker starting today while we continue to work towards making the moderation system more robust."

An intriguing and thankfully less hateful act of vandalistic disinformation on Google Maps occurred[10] in February 2020 when an artist tricked the service into showing a nonexistent traffic jam in the center of Berlin. How did he pull this off? He simply piled a hundred borrowed and rented smartphones into a little red wagon that he slowly walked around the city while the phones' location services were enabled.

Most of the false information on Google Maps is not motivated by hate or artistry—it results from purely financial considerations, as I next discuss.

Fake Business Information

In June 2019, the *Wall Street Journal* reported[11] on the deluge of fake businesses listed on Google Maps. Experts estimated that around ten million business listings on Google Maps at any given moment are falsified and that hundreds of thousands of new ones appear each month. They claim that the "majority of listings for contractors, electricians, towing and car repair services and lawyers, among other business categories, aren't located at their pushpins on Google Maps." One motivation for someone to make

[9]Brian Fung, "If you search Google Maps for the N-word, it gives you the White House," *Washington Post*, May 19, 2015: https://www.washingtonpost.com/news/the-switch/wp/2015/05/19/if-you-search-google-maps-for-the-n-word-it-gives-you-the-white-house/.

[10]Rory Sullivan, "Artist uses 99 phones to trick Google into traffic jam alert," *CNN*, February 4, 2020: https://www.cnn.com/style/article/artist-google-traffic-jam-alert-trick-scli-intl/index.html.

[11]Rob Copeland and Katherine Bindley, "Millions of Business Listings on Google Maps Are Fake—and Google Profits," *Wall Street Journal*, June 20, 2019: https://www.wsj.com/articles/google-maps-littered-with-fake-business-listings-harming-consumers-and-competitors-11561042283.

fake listings is to give a misleading sense of the reach of one's business by exaggerating the number of locations and branch offices on Google Maps. Another motivation is to drown out the competition.

The owner of a cash-for-junk-cars business in the Chicago suburbs mostly relied on the Yellow Pages for advertising, but in 2018 he was contacted by a marketing firm that offered to broadcast his business on Google Maps—for a five-figure fee. He agreed, but then a few months later, the firm came back with a threat: if he doesn't start giving them half his revenue, then they will bury his Google Maps listing under hundreds of fictitious competitors. He refused, and sure enough they posted an avalanche of made-up competitors with locations near him so that it would be very difficult for customers to find his business amid all the fake noise. He drove around a few Chicago neighborhoods and searched on his phone for auto salvage yards as he went; he said that more than half the results that came up were fake. These fake listings pushed his business listing off the first page of Google Maps search results, and soon his number of incoming calls dropped by fifty percent.

Businesses do not pay anything to be listed on Google Maps, but before each one appears on the service, Google usually sends either a postcard or email or calls the business on the phone to provide a verification code that must be typed into Google Maps in order to have the listing approved. This precautionary measure is quite flimsy, and scammers have consistently been able to bypass it. In fact, doing so has become a business. The *Wall Street Journal* profiled a "listings merchant" who placed nearly four thousand fake listings on Google Maps each day from his basement in rural Pennsylvania.

This listings merchant claimed to have had a staff of eleven employees who ran a "mostly" legitimate service that helped clients improve their visibility on Google Maps. But he also claimed to have had a separate staff of twenty-five employees in the Philippines who used "unsanctioned methods to fill orders for fake listings" at a rate of ninety-nine dollars per fake listing. This fake listing service was "aimed at businesses that want to pepper Google Maps with faux locations to generate more customer calls." His employees gathered addresses from commercial real estate listings; to bypass Google's safeguards, they simply purchased phone numbers cheaply online and had Google's verification codes sent to these, then they routed these numbers to the clients once the Google Maps listings were approved. At the time of the *Wall Street Journal* article, however, this listings merchant said Google was investigating him, and tens of thousands of his listings had already been taken down.

Fake business is evidently big business on Google Maps: the site removed over three million false business listings in 2018. That figure comes from a company

blog post[12] written by the director of Google Maps titled "How we fight fake business profiles on Google Maps." This blog post was published on June 20, 2019—the same exact date as the *Wall Street Journal* piece. It does not take a great stretch of the imagination to see this conspicuously timed blog post as a strategic effort to reduce the backlash that would surely follow the publication of the *Wall Street Journal* investigation. This post includes some other staggering figures, including that Google Maps has over two hundred million places and that "every month we connect people to businesses more than nine billion times, including more than one billion phone calls and three billion requests for directions."

The Google Maps blog post gives some examples of how people capitalize on fake business listings: "They do things like charge business owners for services that are actually free, defraud customers by posing as real businesses, and impersonate real businesses to secure leads and then sell them." (We know from the *Wall Street Journal* that there are more problems than just these.) The post also points out that as people find deceptive ways of gaming the system, Google is "continually working on new and better ways to fight these scams using a variety of ever-evolving manual and automated systems," but that as it does this the nefarious users find new deceptive methods and "the cycle continues."

These automated systems—algorithmic moderation, in other words—are closely held corporate secrets because revealing details about them would "actually help scammers find new ways to beat our systems." All the blog post really reveals is that (1) of the three million fake listings taken down in 2018, over ninety percent were removed by the internal systems before a user saw them, whereas the remaining ones were reported by users on the platform, and (2) more than one hundred fifty thousand accounts were disabled in 2018, a fifty percent increase over the previous year.

Perhaps the secretiveness of that blog post did not sit well with some, as just eight months later a new blog post[13] was published—still by the director of Google Maps, but with a different individual occupying this position—that, while still circumspect, went into more detail about the algorithmic moderation the site uses. This post said that Google Maps uses "automated detection systems, including machine learning models, that scan the millions of contributions we receive each day to detect and remove policy-violating content," and that for fake reviews specifically these machine learning models "watch out for specific words and phrases, examine patterns in the types of

[12]Ethan Russell, "How we fight fake business profiles on Google Maps," *Google blog*, June 20, 2019: https://www.blog.google/products/maps/how-we-fight-fake-business-profiles-google-maps/.

[13]Kevin Reece, "Google Maps 101: how contributed content makes a more helpful map," *Google blog*, February 19, 2020: https://www.blog.google/products/maps/google-maps-101-how-contributed-content-makes-maps-helpful/.

content an account has contributed in the past, and can detect suspicious review patterns." Still understandably vague, but I'll turn to the more general topic of machine learning for social media moderation in Chapter 8, so you'll hopefully get a sense of the methods Google Maps is alluding to here—as well as the challenges these methods face.

This second blog post goes on to explain that these automated systems are "not perfect," so Google also relies on "teams of trained operators and analysts who audit reviews, photos, business profiles and other types of content both individually and in bulk." The post also provides some interesting updated figures on content moderation: in 2019, Google Maps (1) removed more than seventy-five million policy-violating reviews and four million fake business profiles "thanks to refinements in our machine learning models and automated detection systems which are getting better at blocking policy-violating content and detecting anomalies for our operators to review"; (2) took down over half a million reviews and a quarter million business profiles that were reported by users; (3) removed ten million photos; (4) disabled almost half a million user accounts.

Google Images

In April 2016, an MBA student posted[14] on Twitter a disturbing discovery: doing a Google image search for "unprofessional hairstyles for work" returned almost entirely pictures of Black women, many with natural hair, while "professional hairstyles for work" returned almost entirely white women. Why was Google's image search algorithm so overtly racist? It's a complicated question, but the two main ingredients to the answer are that the algorithm naively absorbs information out of context and that it naively reflects overt racism permeating society.

Some of the images of Black women that came up on this particular search were from blog posts and Pinterest boards by Black women discussing racist attitudes about hair in the workplace. For instance, one top image was from a post criticizing a university's ban on dreadlocks and cornrows; the post illustrated the banned hairstyles by showing pictures of Black women with them and lamented how these hairstyles were deemed unprofessional by the university. The ban was clearly racist, whereas the post calling attention to it was the opposite, it was antiracist. The Google image search conflated these two contrasting aspects and stripped the hairstyle image of its context, simply associating the image with the word "unprofessional." In doing so, it turned an antiracist image into a racist one. In this inadvertent manner, racism on one

[14]Leigh Alexander, "Do Google's 'unprofessional hair' results show it is racist?" *Guardian*, April 8, 2016: https://www.theguardian.com/technology/2016/apr/08/does-google-unprofessional-hair-results-prove-algorithms-racist-.

college campus was algorithmically amplified and transformed into racist information that was broadcast on a massive scale: anyone innocently looking on Google for tips on how to look professional would be fed the horrendous suggestion that being Black is simply unprofessional.

There soon came to be a data feedback loop here. The MBA student's tweet went viral, which was largely a good thing because it helped raise awareness of Google's algorithmic racism. But this virality caused Google searches on hairstyles to point to this tweet itself and the many discussions about—all of which were calling attention to Google's racism by showing how Black women were labeled "unprofessional" while white women were labeled "professional." Once again, Google vacuumed up these images with these labels and stripped them of their important context, and in doing so the racist effect actually became stronger: a broader array of searches began turning up these offensive associations. In other words, Google image search emboldened and ossified the very same racism that this tweet was calling attention to.

Just a few months after this Google hairstyle fiasco, a trio of researchers in Brazil presented a detailed study on another manifestation of racism in Google's image search—one so abhorrent that the academic study was promptly covered prominently by the *Washington Post*.[15] The researchers collected the top fifty results from Google image searches for "beautiful woman" and "ugly woman," and they did this for searches based in dozens of different countries to see how the results vary by region. This yielded over two thousand images that were then fed into a commercial AI system for estimating the age, race, and gender of each person (supposedly with ninety percent accuracy). Here's what they found.

In almost every country the researchers analyzed, white women appeared more in the search results for "beautiful," and Black and Brown women appeared more in the results for "ugly"—even in Nigeria, Angola, and Brazil, where Black and Brown populations are predominate. In the United States, the results for "beautiful" were eighty percent white and mostly in the age range of nineteen to twenty-eight, whereas the results for "ugly" dropped to sixty percent white—and rose to thirty percent Black—and the ages mostly ranged from thirty to fifty, according to the AI estimates. This form of racism and ageism was not invented by Google's algorithm, it originates in society itself—but the algorithm picks up on it and then harmfully presents it to the world as an established fact. Thankfully, Google seems to have found ways of improving its algorithm in this regard as image searches for beauty now yield a much more diverse range of individuals.

[15]Caitlin Dewey, "Study: Image results for the Google search 'ugly woman' are disproportionately black," *Washington Post*, August 10, 2016: https://www.washingtonpost.com/news/the-intersect/wp/2016/08/10/study-image-results-for-the-google-search-ugly-woman-are-disproportionately-black/.

Google Photos is a service introduced by Google in 2015 that allows users to store and share photos, and it uses machine learning to automatically recognize the content of the photos. It now has over one billion users who collectively upload more than one billion photos to the platform daily. But just one month after its initial launch, Google had to offer an official apology and declared itself "appalled and genuinely sorry" for a racist incident—an incident that Google's chief social architect responded[16] to on Twitter by writing "Holy fuck. […] This is 100% not OK." What had happened? A Black software engineer and social activist revealed on Twitter that Google Photos repeatedly tagged pictures of himself and his girlfriend as "gorillas." Google said that as an immediate fix it would simply discontinue using the label "gorilla" in any capacity, and the company would work on a better longer-term solution.

Two and a half years later, *Wired* conducted a follow-up investigation[17] to see what Google had done to solve this heinous mislabeling problem. It turns out Google hadn't gotten very far from its original slapdash workaround: in 2018, "gorilla," "chimp," "chimpanzee," and "monkey" were simply disallowed tags on Google Photos. Indeed, *Wired* provided Google Photos with a database of forty thousand images well stocked with animals and found the platform performed impressively well at returning photos of whatever animals were requested—except for those named above: when those words were searched, Google Photos said no results were found. For all the hype in AI, the highly lauded team at Google Brain, the heralded breakthroughs provided by deep learning, it seems one of the most advanced technology companies on the planet couldn't figure out how to stop its algorithms from tagging Black people as gorillas other than by explicitly removing gorilla as a possible tag.

The problem here is that even the best AI algorithms today don't form abstract conceptualizations or common sense the way a human brain does—they just find patterns when hoovering up reams of data. (You might object that when I discussed deep learning earlier in this book, I did say that it is able to form abstract conceptualizations, but that's more in the sense of patterns within patterns rather than the kind of anthropomorphic conceptualizations us humans are used to.) If Google's algorithms are trained on real-world data that contains real-world racism, such as Black people being referred to as gorillas, then the algorithms will learn and reproduce this same form of racism.

Let me quickly recap the very public racist Google incidents discussed so far to emphasize the timeline. In May 2015, the *Washington Post* reported the Google Maps White House story that the office of the first Black president in

[16]Loren Grush, "Google engineer apologizes after Photos app tags two black people as gorillas," *The Verge*, July 1, 2015: https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas.

[17]Tom Simonite, "When It Comes to Gorillas, Google Photos Remains Blind," *Wired*, January 11, 2018: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.

the history of the United States was labeled with the most offensive racial slur in existence. One week later, Google launched Google Photos and within a month had to apologize for tagging images of Black people as gorillas, a story covered by the *Wall Street Journal*, among others. Less than a year later, in April 2016, the *Guardian* reported that Google image searches for unprofessional hairstyles mostly showed photos of Black women. Just a few months after that, in August 2016, the *Washington Post* covered a research investigation that showed Google image search results correlated beauty with race. Oh, I almost forgot: two months earlier, in June 2016, it was reported in many news outlets, including *BBC News*,[18] that doing a Google image search for "three black teenagers" returned mostly police mugshots, whereas searching for "three white teenagers" just showed smiling groups of wholesome-looking kids. These documented racist incidents are just a sample of the dangers inherent in letting Google's data-hungry machine learning algorithms sort and share the world's library of photographs.

Google Autocomplete

Google's autocomplete feature suggests popular searches for users after they type in one or more words to the search box on Google's homepage or to the address bar in Google's web browser Chrome. It is supposed to be a simple efficiency tool, like the autocomplete on your phone that helps you save time by suggesting word completions while you are texting. But Google searches are a powerful instrument that billions of people use as their initial source of information on just about every topic imaginable, so the consequences can be quite dire when Google's autocompletes send people in dangerous directions.

Suggesting Hate

In December 2016, it was reported[19] in the *Guardian* that Google's suggested autocompletes for the phrase "are Jews" included "evil," and for "are Muslims" they included "bad." Several other examples of offensive autocompletes were also found. In response,[20] a Google representative said the following:

[18]Rozina Sini, "'Three black teenagers' Google search sparks Twitter row," *BBC News*, June 9, 2016: https://www.bbc.com/news/world-us-canada-36487495.

[19]Carole Cadwalladr, "Google, democracy and the truth about internet search," *Guardian*, December 4, 2016: https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook.

[20]Samuel Gibbs, "Google alters search autocomplete to remove 'are Jews evil' suggestion," *Guardian*, December 5, 2016: https://www.theguardian.com/technology/2016/dec/05/google-alters-search-autocomplete-remove-are-jews-evil-suggestion.

"Our search results are a reflection of the content across the web. This means that sometimes unpleasant portrayals of sensitive subject matter online can affect what search results appear for a given query. These results don't reflect Google's own opinions or beliefs." This refrain—that Google isn't racist, it's merely reflecting racism in society—has been a recurring defense throughout all these scandals. Google's response to this *Guardian* article went on to explain that its algorithmically generated autocomplete predictions "may be unexpected or unpleasant," but that "We do our best to prevent offensive terms, like porn and hate speech, from appearing."

On the official company blog, Google explains[21] that autocomplete is really providing "predictions" rather than "suggestions"—meaning it is using machine learning trained on the company's vast database of searches to estimate the words most likely to follow the words the user has typed so far.[22] In other words, Google is not trying to suggest what you should search for, it is just trying to figure out what is most probable that you will be searching for based on what you have typed so far. The Google blog explains that the autocomplete algorithm makes these statistical estimates based on what other users have searched for historically, what searches are currently trending, and also—if the user is logged in—your personal search history and geographic location.

Google is smart enough to moderate (both algorithmically and manually) the results of this machine learning prediction system. The company blog states that "Google removes predictions that are against our autocomplete policies, which bar: sexually explicit predictions that are not related to medical, scientific, or sex education topics; hateful predictions against groups and individuals on the basis of race, religion or several other demographics; violent predictions; dangerous and harmful activity in predictions." It says that a "guiding principle" here is to "not shock users with unexpected or unwanted predictions." In case you lost track, this means Google has said that its autocompletes may be "unexpected or unpleasant," but they aren't supposed to be "unexpected or unwanted." Confused yet? I know that I am.

A Google spokesperson said the company took action within hours of being notified of the offensive autocompletes uncovered by the *Guardian* article.

---

[21]Danny Sullivan, "How Google autocomplete works in Search," *Google blog*, April 20, 2020: https://www.blog.google/products/search/how-google-autocomplete-works-search/.

[22]This general idea of next-word prediction is somewhat similar to the GPT-3 system discussed in Chapter 2. However, GPT-3 was pre-trained on huge volumes of written text and then in real time considers only the prompt text. Google's autocomplete, on the other hand, uses pre-training data that is more focused on searches, and its real-time calculation uses not just the text typed into the prompt so far but also many other factors, as discussed shortly.

However, the *Guardian* found[23] that only some of the offensive examples listed in that article were removed; others remained. Evidently, Google's "guiding principle" is difficult to implement uniformly and incontrovertibly in practice. A little over a year later, in a February 2018 UK parliamentary hearing, Google's vice president of news admitted that "As much as I would like to believe our algorithms will be perfect, I don't believe they ever will be."

An investigation[24] was published in *Wired* just a few days after this UK hearing, finding that "almost a year after removing the 'are jews evil?' prompt, Google search still drags up a range of awful autocomplete suggestions for queries related to gender, race, religion, and Adolf Hitler." To avoid possibly misleading results, the searches for this *Wired* article were conducted in "incognito" mode, meaning Google's algorithm was only using general search history data rather than user-specific data. The top autocompletes for the prompt "Islamists are" were, in order of appearance, "not our friends," "coming," "evil," "nuts," "stupid," and "terrorists." The prompt "Hitler is" yielded several reasonable autocompletes as well as two cringeworthy ones: "my hero" and "god." The first autocomplete for "white supremacy is" was "good," whereas "black lives matter is" elicited the autocomplete "a hate group." Fortunately, at least, the top link for the search "black lives matter is a hate group" was to a Southern Poverty Law Center post explaining why BLM is not, in fact, a hate group. Sadly, however, one of the top links for the search "Hitler is my hero" was a headline proclaiming "10 Reasons Why Hitler Was One of the Good Guys."

Strikingly, the prompt "blacks are" had only one autocomplete, which was "not oppressed," and the prompt "feminists are" also only had a single autocomplete: "sexist." Google had clearly removed most of the autocompletes for these prompts but missed these ones which are still biased and a potentially harmful direction to send unwitting users toward. Some things did improve in the year between the original *Guardian* story and the *Wired* follow-up. For instance, the prompt "did the hol" earlier autocompleted to "did the Holocaust happen," and then the top link for this completed search was to the neo-Nazi propaganda/fake news website *Daily Stormer*, whereas afterward this autocomplete disappeared, and even if a user typed the full search phrase manually, the top search result was, reassuringly, the Holocaust Museum's page on combatting Holocaust denial.

It's difficult to tell how many of the autocomplete and search improvements that happen over time are due to specific ad hoc fixes and how many are due to overall systemic adjustments to the algorithms. On this matter, a Google spokesperson wrote: "I don't think anyone is ignorant enough to think,

[23]See Footnote 20.

[24]Issie Lapowsky, "Google Autocomplete Still Makes Vile Suggestions," *Wired*, February 12, 2018: https://www.wired.com/story/google-autocomplete-vile-suggestions/.

'We fixed this one thing. We can move on now'." It is important to remember that when it comes to hate, prejudice, and disinformation, Google—like many of the other tech giants—is up against a monumental and mercurial challenge.

One week after the *Wired* article, it was noted[25] that the top autocomplete for "white culture is" was "superior"; the top autocompletes for "black culture is" were "toxic," "bad," and "taking over America." Recall that in 2014, Ferguson, Missouri, was the site of a large protest movement responding to the fatal police shooting of an eighteen-year-old Black man named Michael Brown. In February 2018, the autocompletes for "Ferguson was" were, in order: "a lie," "staged," "not about race," "a thug," "planned," "he armed," "a hoax," "fake," "stupid," and "not racist"; the top autocompletes for "Michael Brown was" were "a thug," "no angel," and "a criminal."

While drafting this chapter in November 2020, I was curious to see how things had developed since the articles discussed here. After switching to incognito mode, I typed "why are black people" and Google provided the following autocompletes: "lactose intolerant," "'s eyes red," "faster," "called black," and "'s palms white." Somewhat strange, but not the most offensive statements. I was relieved to see that Google had indeed cleaned up its act. But then, before moving on, I decided to try modifying the prompt very slightly: "why do black people" (just changing the "are" to "do"). The autocompletes this produced absolutely shocked and appalled me. In order, the top five were "sag their pants," "wear white to funerals," "resist arrest," "wear durags," and "hate jews." How it is deemed even remotely acceptable to use an algorithm to broadcast such harmful vitriol and misinformation to any of the billions of people who naively seek information from Google is simply beyond me.

In addition to all these autocompletes that are wrong in a moral sense, many autocompletes are just plain wrong in a literal, factual sense, as I next discuss.[26]

Suggesting Fake News

When people use Google to search for information, they sometimes interpret the autocomplete suggestions as headlines, as statements of fact. So when autocompletes are incorrect or misleading assertions, this can be seen as another instance of fake news on Google.

---

[25]Barry Schwartz, "Google Defends False, Offensive & Fake Search Suggestions As People's Real Searches," *Search Engine Round Table*, February 23, 2018: https://www.seroundtable.com/google-defends-false-fake-search-suggestions-25294.html.

[26]Of course, most of the preceding examples are also factually wrong—but I am trying to separate the topic of hate speech from the topic of fake news in this treatment of Google autocompletes. The dividing line, however, is admittedly quite blurred.

A December 2016 investigation in *Business Insider* found[27] the following. The first autocomplete for "Hillary Clinton is" was "dead," and the top link that resulted from this search was an article on a fake news site asserting that she was indeed dead. The top autocomplete for "Tony Blair is" was also "dead." Same for Vladimir Putin. The February 2018 *Wired* investigation cited above noted that the top autocompletes for "climate change is" were, in order, "not real," "real," "a hoax," and "fake." Also in February 2018, it was noted[28] that the autocompletes for "mass shootings are" included "fake," "rare," "democrats," and "the price of freedom"; the top autocomplete for "David Hogg," the activist student survivor of the Stoneman Douglas High School mass shooting, was "actor."

In September 2020, Google announced[29] that it had updated the autocomplete policies related to elections: "We will remove predictions that could be interpreted as claims for or against any candidate or political party. We will also remove predictions that could be interpreted as a claim about participation in the election—like statements about voting methods, requirements, or the status of voting locations—or the integrity or legitimacy of electoral processes, such as the security of the election." After a week, *Wired* conducted experiments[30] to see how well Google's new policies were working. In short, while this policy change was well intentioned and mostly successful, the implementation was not perfect. Typing "donate" led to a variety of suggestions, none of which concerned the upcoming presidential election, but typing "donate bid" was autocompleted to "donate biden harris actblue," a leading Democratic political action committee. On the other hand, typing "donate" and then the first few letters of Trump's name didn't result in any political autocompletes—the only autocomplete was "donate trumpet."

A few weeks later, Google posted[31] an overview description of how the autocomplete feature works. It includes the following remark on fake news: "After a major news event, there can be any number of unconfirmed rumors or information spreading, which we would not want people to think

[27]Hannah Roberts, "How Google's 'autocomplete' search results spread fake news around the web," *Business Insider*, December 5, 2016: https://www.businessinsider.com/autocomplete-feature-influenced-by-fake-news-stories-misleads-users-2016-12.

[28]See Footnote 24.

[29]Pandu Nayak, "Our latest investments in information quality in Search and News," *Google blog*, September 10, 2020: https://blog.google/products/search/our-latest-investments-information-quality-search-and-news/.

[30]Tom Simonite, "Google's Autocomplete Ban on Politics Has Some Glitches," *Wired*, September 11, 2020: https://www.wired.com/story/googles-autocomplete-ban-politics-glitches/.

[31]Danny Sullivan, "How Google autocomplete predictions are generated," *Google blog*, October 8, 2020: https://blog.google/products/search/how-google-autocomplete-predictions-work/.

Autocomplete is somehow confirming. In these cases, our systems identify if there's likely to be reliable content on a particular topic for a particular search. If that likelihood is low, the systems might automatically prevent a prediction from appearing." You have to read that statement carefully: Google is not saying that it removes misinformative autocompletes, it is saying that it removes some autocompletes that would yield mostly fake news search results. I suppose the idea behind this is that if a user sees a false assertion as an autocomplete, the truth should be revealed when the user proceeds to search for that assertion—and only if the assertion is not readily debunkable this way should it be removed from autocomplete. But, to me at least, that doesn't really jibe with the first sentence in Google's statement, which makes it seem that the company is concerned about people seeing misinformative autocompletes regardless of the searches they lead to.

Do you remember Guillaume Chaslot, the former Google computer engineer you met in Chapter 4 who went from working on YouTube's recommendation on the inside to exposing the algorithm's ills from the outside? On November 3, 2020—election day—he found that the top autocompletes for "civil war is" were, in order: "coming," "here," "inevitable," "upon us," "what," "coming to the us," and "here 2020." On January 6, 2021—the day of the Capitol building insurrection—he tried the same phrase and found the top autocompletes were no less terrifying: "coming," "an example of which literary term," "inevitable," "here," "imminent," "upon us."

In a post[32] on *Medium*, Chaslot used Google Trends to look into how popular these searches were. Rather shockingly, he found that in the month leading up to the Capitol building event, the phrase "civil war is what" was searched seventeen times more than "civil war is coming," thirty-five times more than "civil war is here," fifty-two times more than "civil war is starting," one hundred thirteen times more than "civil war is inevitable," and one hundred seventy-five times more than "civil war is upon us." In other words, Google was suggesting extremely incendiary searches even though they were far less popular than a harmless informative query. Chaslot pointed out that this "demonstrates that Google autocomplete results can be completely uncorrelated to search volumes" and that "We don't even know which AI is used for Google autocomplete, neither what it tries to optimize for." He also found that one of the autocompletes for "we're hea" was "we're heading into civil war"—so that even people typing something completely unrelated might be dangerously drawn into this extremist propaganda.

In October 2020, just weeks before the election, Chaslot also found harmful misinformation about the COVID pandemic: the autocompletes for

[32]Guillaume Chaslot, "Google Autocomplete Pushed Civil War narrative, Covid Disinfo, and Global Warming Denial," *Medium*, February 9, 2021: https://guillaumechaslot.medium.com/google-autocomplete-pushed-civil-war-narrative-covid-disinfo-and-global-warming-denial-c1e7769ab191.

"coronavirus is" included "not that serious," "ending," "the common cold," "not airborne," and "over now." In fact, of the ten autocompletes for this search phrase, six were assertions that have been proven wrong. And once again the order of these autocompletes was unrelated to search popularity as measured by Google Trends. Chaslot found climate change denial/misinformation persisted as well: three of the top five autocompletes for "global warming is" were "not caused by humans," "good," and "natural." The phrase "global warming is bad" was searched three times as often as "global warming is good," and yet the latter was the number four autocomplete, while the former was not included as an autocomplete.

The popularity of a search depends on the window of time one is considering (the last hour? day? month? year?), so it's possible that Google's autocomplete algorithm was just using a different window than Chaslot was on Google Trends, but we don't really know. Chaslot concludes his *Medium* post with the following stark warning/critique: "The Google autocomplete is serving the commercial interests of Google, Inc. […] It tries to maximize a set of metrics, that are increasing Google's profit or its market share. They choose how they configure their AI."

Google News

The *News* section of Google provides links to articles that are algorithmically aggregated into collections in a few different ways: *For you* articles are individualized recommendations based on user data, *Top stories* are trending stories that are popular among a wide segment of the user population, and there are a variety of categories (such as business, technology, sports, etc.) that collect trending stories by topic. Google News also allows users to perform keyword searches that return only news articles rather than arbitrary website links. Needless to say, when stories appear in these aggregated collections or news article searches, it lends them an air of legitimacy, in addition to a large audience—even if the story is fake news.

The details of how these news gathering/ranking algorithms work are closely guarded corporate secrets. In May 2019, the VP of Google News wrote[33] that "The algorithms used for our news experiences analyze hundreds of different factors to identify and organize the stories journalists are covering, in order to elevate diverse, trustworthy information." While most of these hundreds of factors are not publicly known, it is understood that they include, among other things, the number of clicks articles get, estimates of the trustworthiness of the publishing organizations, geographic relevance, and freshness of the content.

[33]Richard Gingras, "A look at how news at Google works," *Google blog*, May 6, 2019: https://blog.google/products/news/look-how-news-google-works/.

The VP of Google News also stated that "Google does not make editorial decisions about which stories to show" and that "our primary approach is to use technology to reflect the news landscape, and leave editorial decisions to publishers." To prevent fake news from running rampant on the platform, Google says "Our algorithms are designed to elevate news from authoritative sources." Very little has been said publicly about what this really means and how it is accomplished. All I could find on Google's official website that supposedly describes how trustworthy news is elevated[34] is that the algorithms rely on signals that "can include whether other people value the source for similar queries or whether other prominent websites on the subject link to the story."

Alas, it doesn't seem that much progress has been made toward uncovering the state of fake news on Google News and the company's efforts to limit it. However, the official explanatory site for Google News also states that "Our ranking systems for news content across Google and YouTube News use the same web crawling and indexing technology as Google Search," so it seems the time is right to turn now to this chapter's lengthiest section: the role Google search plays in the dissemination of fake news.

Google Search

When we search for information on Google, the results that come up—and the order they are presented in—shape our views and beliefs. This means that for Google to limit the spread of misinformation, it must find ways of training its algorithms to lift quality sources to the top without taking subjective, biased perspectives on contentious issues and also without impinging on people's ability to scour the depths of the Web. There are many pieces of this story. In this section, I will present evidence backing up the assertion that search result rankings affect individuals' worldviews; look into what factors Google's search algorithm uses to decide how to rank links; illustrate how featuring highlights from top searches has led to problematic misinformation; show how authentic links have been removed from Google through deceptive means; introduce the deep learning language model Google recently launched to power its search and many other tools; and, finally, discuss Google's efforts to elevate quality journalism in its search rankings.

[34] https://newsinitiative.withgoogle.com/hownewsworks/mission.

Ranking Matters

In August 2015, a study[35] of the impact Google search rankings have on political outlook was published in the prestigious research journal *Proceedings of the National Academy of Sciences*. One of the main experiments in this study was the following. Participants were randomly placed in three different groups. The participants were all provided brief descriptions of two political candidates, call them A and B, and then asked how much they liked and trusted each candidate and whom they would vote for. They were then given fifteen minutes to look further into the candidates using a simulated version of Google that only had thirty search results—the same thirty for all participants—that linked to actual websites from a past election. After this fifteen-minute session, the participants were asked the same questions as before about the two candidates. The key was that one group had the search results ordered to return the results favorable to candidate A first, another group had results favorable to candidate B first, and for the third group the order was mixed.

The researchers found that on all measures, the participants' views of the candidates shifted in the direction favored by the simulated search rankings, by amounts ranging between thirty-seven and sixty-three percent. And this was just from a single fifteen-minute search session. The researchers also experimented with a real election—two thousand undecided votes in a 2014 election in India. The authors stated[36] that "Even here, with real voters who were highly familiar with the candidates and who were being bombarded with campaign rhetoric every day, we showed that search rankings could boost the proportion of people favoring any candidate by more than 20 percent—more than 60 percent in some demographic groups."

The researchers go on to boldly suggest that "Google's search algorithm, propelled by user activity, has been determining the outcomes of close elections worldwide for years, with increasing impact every year because of increasing Internet penetration." I find this assertion to be a stretch—at least, the evidence to really back it up isn't in their PNAS paper—but my focus in this book is not political bias and elections, it is fake news. And the researchers here did convincingly establish that search rankings matter and affect people's views, which means there are real consequences when Google places fake news links highly in its search rankings.

---

[35]Robert Epstein and Ronald Robertson, "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections," *Proceedings of the National Academy of Sciences (PNAS)*, August 18, 2015, 112 (33), E4512-E4521: https://doi.org/10.1073/pnas.1419828112.

[36]Robert Epstein, "How Google Could Rig the 2016 Election," *Politico*, August 19, 2015: https://www.politico.com/magazine/story/2015/08/how-google-could-rig-the-2016-election-121548/.

Dylann Roof, the neo-Nazi who in 2015 murdered nine Black people at a church in Charleston, South Carolina, wrote a manifesto[37] in which he claims to have been inspired by Google. He described how he typed "black on White crime" into Google, and he has "never been the same since that day." He says the first site this search produced was for an organization called the Council of Conservative Citizens (CCC). It contained numerous descriptions of "brutal black on White murders." He alleges that seeing this made him question the media's attention on the Trayvon Martin case, and it motivated him to pursue a self-education journey on Google on racial matters—a journey that, as we now know, ended in unfathomable tragedy.

While this is anecdotal evidence, it is nonetheless quite unsettling and shows one of the dangers of letting algorithms decide what information we should see, and in what order. As UCLA professor Safiya Noble pointed out in her book *Algorithms of Oppression* critiquing Google, the top result for Roof's search should have been an authoritative source such as FBI crime statistics—which shows that most violence against white Americans is committed by white Americans—rather than the CCC, an organization whose Statement of Principles says that it "opposes all efforts to mix the races of mankind."

Signals the Algorithm Uses

What factors determine how highly ranked pages are in Google searches? Once again, Google won't reveal much about its algorithmic trade secrets—in part to prevent competitors from copying them, but also to prevent people from gaming the algorithm—so we only know the broadest outlines. The official company website describing the search algorithm[38] states the following: "Search algorithms look at many factors, including the words of your query, relevance and usability of pages, expertise of sources, and your location and settings. The weight applied to each factor varies depending on the nature of your query—for example, the freshness of the content plays a bigger role in answering queries about current news topics than it does about dictionary definitions." These factors are largely about finding links that appear to be good matches to the search query. When it comes to ranking the results, Google says the algorithm attempts to "prioritize the most reliable sources available" by considering factors that "help determine which pages demonstrate expertise, authoritativeness, and trustworthiness on a given topic." This sounds good, but it's quite vague. The two examples Google gives are that a

[37]Daniel Strauss, "Racist manifesto linked to Dylann Roof emerges online," *Politico*, June 20, 2015: https://www.politico.com/story/2015/06/dylan-roofs-racist-manifesto-emerges-online-119254.

[38]https://www.google.com/search/howsearchworks/algorithms/.

site is bumped up in the rankings if other prominent sites link to it (this is the essence of the original PageRank algorithm Google first launched with in 1998) or if many users visit the site after doing closely related searches.

Earlier in the history of the algorithm, the PageRank method played a more prominent role, and less attention was given to assessing the quality of information by other means. Nefarious actors figured out how to use this narrow focus on link counting to manipulate the rankings. In December 2016, it was reported[39] that fake news and right-wing extremist sites "created a vast network of links to each other and mainstream sites that has enabled them to game Google's algorithm." This led to harmful bigotry and disinformation—for instance, eight of the top ten search results for "was Hitler bad?" were to Holocaust denial sites.

Several months later, in April 2017, Google admitted[40] that fake news had become a serious problem: "Today, in a world where tens of thousands of pages are coming online every minute of every day, there are new ways that people try to game the system. The most high profile of these issues is the phenomenon of 'fake news,' where content on the web has contributed to the spread of blatantly misleading, low quality, offensive or downright false information." One change Google implemented at the time was to provide more detailed guidance on "misleading information, unexpected offensive results, hoaxes and unsupported conspiracy theories" for the team of human evaluators the company uses to provide feedback on the search algorithms: "These guidelines will begin to help our algorithms in demoting such low-quality content and help us to make additional improvements over time." I'll return to these human moderators and the role they play in shaping Google's search algorithm at the end of this section. Google also said that the signals used in the algorithm were adjusted in order to "help surface more authoritative pages and demote low-quality content," but no details were provided.

One year later, as part of a multipronged effort to "elevate quality journalism" (including a three-hundred-million-dollar outreach initiative[41]), Google tweaked the algorithm again. This time, it revealed which signals were prioritized in this

[39]Carole Cadwalladr, "Google 'must review its search rankings because of rightwing manipulation'," *Guardian*, December 5, 2016: https://www.theguardian.com/technology/2016/dec/05/google-must-review-its-search-rankings-because-of-rightwing-manipulation.

[40]Ben Gomes, "Our latest quality improvements for Search," *Google blog*, April 25, 2017: https://blog.google/products/search/our-latest-quality-improvements-search/.

[41]Kevin Roose, "Google Pledges $300 Million to Clean Up False News," *New York Times*, March 20, 2018: https://www.nytimes.com/2018/03/20/business/media/google-false-news.html.

adjustment and why, at least in very broad strokes:[42] "To reduce the visibility of [harmful misinformation] during crisis or breaking news events, we've improved our systems to put more emphasis on authoritative results over factors like freshness or relevancy."

Featured Snippets

In 2014, Google introduced a tool called *featured snippets*: a box of text is placed at the top of the results page for certain searches containing a highlighted passage from a top link that the algorithm believes is relevant to the search. These are not included for all searches, just ones that Google thinks are asking for specific information that it can try to find on the Web. Featured snippets help people extract information quickly from the internet since you get the answers to your questions directly from a Google search without having to choose a link, click it, then find the relevant passage buried somewhere on that site; they are also useful for voice search on mobile devices and Google's Home Assistant because the user can ask a question verbally and then the device responds verbally by reading aloud the featured snippet resulting from the corresponding Google search.

But if a Google search turns up fake news, then the answer provided by Google in the featured snippet might be wrong. And the featured snippet format strips the answer of any context and presents it in an authoritative-sounding manner, leaving the reader/listener even less aware than in a typical Google search of how unreliable the information source might be. Sure enough, featured snippets eventually made headlines for providing disastrously misinformed answers to a variety of questions.

Indeed, in March 2017, it was found[43] that asking Google which US presidents were in the KKK resulted in a snippet falsely claiming that several were; asking "Is Obama planning a coup?" yielded a snippet that said "According to details exposed in Western Center for Journalism's exclusive video, not only could Obama be in bed with the Communist Chinese, but Obama may in fact be planning a Communist coup d'état at the end of his term in 2016!"; searching for a gun control measure called "Proposition 63" yielded a snippet falsely describing it as "a deceptive ballot initiative that will criminalize millions of law abiding Californians." In the case of the Obama coup snippet, the top search result was an article debunking this fake news story about an upcoming coup attempt, but when using Google's Home Assistant, there are no search results listed—all one gets is the featured snippet read aloud.

---

[42]Richard Gingras, "Elevating quality journalism on the open web," *Google blog*, March 20, 2018: https://blog.google/outreach-initiatives/google-news-initiative/elevating-quality-journalism/.

[43]Rory Cellan-Jones, "Google's fake news Snippets," *BBC News*, March 6, 2017: https://www.bbc.com/news/technology-39180855.

In a blog post[44] a year later describing how the snippets tool had improved over time, the company said "Last year, we took deserved criticism for featured snippets that said things like […] Obama was planning a coup. We failed in these cases because we didn't weigh the authoritativeness of results strongly enough for such rare and fringe queries." While this claim is superficially true, it conveniently sweeps under the rug that a prerequisite to *weighing* authoritativeness is *measuring* authoritativeness, which is a challenging, fraught issue—one I'll return to at the end of this section. Reducing misinformation by elevating quality sources is not nearly as simple as adjusting a lever labeled "authoritativeness" the way Google suggests here in this remark.

Blocking Search Results

Some fake news publishers scrape articles from the Web and repost them as their own in an effort to give the stories a wider platform and to collect ad revenue in the process. While most fake news on the Web doesn't violate any Google policies that would prevent it from being eligible to show up on search results, if an article is found to be in violation of copyright, then Google will expunge it from search listings—so these spammy fake news sites indeed get delisted.

But an extensive *Wall Street Journal* investigation[45] found an unsettling twist here: people have been gaming Google's copyright infringement request system in order to delist content that is unflattering or financially impactful to certain parties. One of the techniques used is *backdating*: someone copies a published article and posts it on their blog but with a misleading time stamp to make it appear that it predates the published article; then they tell Google that the published article is violating their blog's copyright, and the published article is removed from Google search results. When this happens, Google is delisting an actual news article on the basis of a false copyright infringement notification. Daphne Keller, a former Google lawyer and currently a program director at Stanford University's Cyber Policy Center, said that "If people can manipulate the gatekeepers to make important and lawful information disappear, that's a big deal." The *Wall Street Journal* found that not only can people indeed do this, but they have been doing so in surprisingly large numbers.

---

[44]Danny Sullivan, "A reintroduction to Google's featured snippets," *Google blog*, January 30, 2018: https://blog.google/products/search/reintroduction-googles-featured-snippets/.

[45]Andrea Fuller, Kirsten Grind, and Joe Palazzolo, "Google Hides News, Tricked by Fake Claims," *Wall Street Journal*, May 15, 2020: https://www.wsj.com/articles/google-dmca-copyright-claims-takedown-online-reputation-11589557001.

Google received copyright removal requests for fewer than one hundred thousand links between 2002 and 2012, whereas it now routinely handles more than a million requests in a single day. In order to scale up to this magnitude, a company spokesperson said that Google has automated much of the process so that human review is mostly not needed. In 2019, eighty percent of the nearly one-quarter billion links flagged for copyright infringement were removed from Google's search listings. However, after the *Wall Street Journal* uncovered numerous cases of fraudulent violation notifications, Google restored more than fifty thousand links that had been removed.

One of the clusters of fraudulent requests the *Wall Street Journal* found concerned Russian-language news articles critical of politicians and business leaders in Ukraine. These articles were taken off Google after various organizations including a Russian edition of *Newsweek* filed a copyright violation request—but it turned out these organizations were all fake, the supposed Russian *Newsweek* had nothing to do with *Newsweek*, it was just using *Newsweek*'s logo to deceive Google into thinking the copyright violation notification was legitimate.

There is a secondary harm to these deceptive methods for tricking Google into delisting real news articles: in Google's recent efforts to elevate quality journalism, one factor the ranking algorithm considers is the number of copyright violations sites have received. A Google spokesperson said that "If a website receives a large number of valid takedown notices, the site might appear lower overall in search results." But the *Wall Street Journal* investigation established that many of the takedowns that Google thinks are valid are actually invalid and the result of deliberate disinformation aimed at Google's automated system. This opens the door for more gaming of Google's rankings by dishonest actors.

BERT

A humorous cultural trend emerging in the AI community over the past few years has been to make as many *Sesame Street* allusions as possible when naming deep learning text processing algorithms. You saw *Grover* in Chapter 2, the system developed by the Allen Institute for AI to generate text like GPT-2 with the ultimate goal of being able to detect such deep learning generated text. Currently, the two most impressive and powerful deep learning systems for text are GPT-3, which was discussed extensively in Chapter 2, and a system developed by Google called *BERT* (which stands for *Bidirectional Encoder Representations from Transformers*, but don't worry about that just yet). BERT builds on an earlier system developed by the Allen Institute for AI named *ELMo* (standing for *Embeddings from Language Models*). Alas, there is not yet an Ernie or a Snuffleupagus.

While text generation is immensely useful, it turns out that for many applications one needs something that is essentially a by-product of the inner workings of a deep neural net that occurs automatically while training for tasks like text generation: a *vector representation* of words (sometimes called a *word embedding*). This means a way of representing each word as a vector—that is, a list of numerical coordinates—in such a way that the geometry of the distribution of word vectors reflects important semantic and syntactic information. Roughly speaking, we want words that frequently appear in close proximity to each other in written text to have vector representations that are geometrically in close proximity to each other. Vector embeddings translate data in messy formats like text into the standard numerical formats that machine learning algorithms know and love.

Earlier word embeddings (one of the most popular, called *Word2vec*, was developed by Google in 2013) produced a fixed, static vector for each word. This opened the door to many breakthroughs: for example, analyzing the sentiment of words and sentences (how positive or negative they are) turns into a more familiar geometric analysis of vectors in Euclidean space, where, for instance, one looks for a plane that separates positive word vectors from negative word vectors. One of the key drawbacks in these early static approaches was that a single word can have multiple meanings (such as "stick" a landing, "stick" from a tree, and "stick" to a plan), and all the different meanings got conflated when the word was represented as a vector. In contrast, ELMo and BERT are *contextual* word embeddings, which means the vector representations are not fixed and static—they depend on the surrounding text. If you feed these systems the sentence "I hope the gymnast sticks the landing" and the sentence "the toddler sticks out her tongue," the word "sticks" will have different vector representations in each case. This allows for much more flexibility in language modeling and understanding.

BERT learns its contextual word embeddings through a massive self-supervised pre-training process somewhat similar to that of GPT-3. As you may recall from Chapter 2, GPT-3 was fed huge volumes of text, and as it read through this, it used the preceding words to try to predict the next word. BERT's self-supervised training process also involved reading massive volumes of text, but in this case a percentage of the words were randomly masked (hidden from the algorithm). BERT learned how to "predict" these missing words by guessing what they were and then unmasking them and using the difference between the guess and the actual unmasked word as an error to propagate through the neural net and adjust all the parameters so that over time the guesses become more accurate. In this way, BERT was trained to predict missing words. BERT's ultimate goal is not to predict words, but being able to predict words well is considered a good proxy for understanding them; or, from a more technical perspective, the contextual vector embeddings BERT develops internally while training on hidden word prediction turn out to be very useful for a wide range of linguistic tasks.

Actually, this masked word task helps BERT learn about words in the context of each sentence, but to get a more global perspective, it is simultaneously trained on a self-supervised sentence prediction task: it is shown pairs of sentences from the training text and learns to estimate the probability that one sentence immediately precedes the other. This task helps the word embeddings encode larger-scale meaning that extends beyond individual sentences. In case you are curious, the *Bidirectional* in BERT's name refers to the fact that it reads training text both left-to-right and right-to-left in order to get both past and future context for each word. This is a reasonable thing to do precisely because, unlike GPT-3, BERT is not aiming to predict future words—it is aiming to produce word embeddings that draw in as much context as possible. The *Encoder Representations* part of the name just indicates that words are encoded with vector representations, and the *Transformer* in the name refers to a specific deep learning architecture that is used.

When you type a search phrase into Google, you are providing more than just a list of keywords to match—often you are providing a grammatical snippet of text that Google's search algorithm needs to understand. BERT is the intermediary service that translates your search phrase into a collection of vectors that the search algorithm can then process quantitatively. In an October 2019 company blog post announcing the absorption of BERT into Google's search algorithm, Google's vice president of Search said[46] that "Search is about understanding language," and BERT has indeed been one of the most successful steps forward in allowing computers to better understand human language.

This Google blog post went on to say that "when it comes to ranking results, BERT will help Search better understand one in 10 searches in the U.S. in English" and that "Particularly for longer, more conversational queries, or searches where prepositions like 'for' and 'to' matter a lot to the meaning, Search will be able to understand the context of the words in your query." To illustrate the types of improvements users should expect, the blog post included an example of a user searching "Can you get medicine for someone pharmacy." Previously, the top search result was an article that included each of these individual words, but it didn't answer the question the user was attempting to ask; with the new BERT-powered search, the top result was an article specifically addressing when and how people can pick up medications for others at the pharmacy.

Google evidently underestimated itself with the "one in 10" figure, because almost exactly one year later in another company blog post[47] Google declared

[46]Pandu Nayak, "Understanding searches better than ever before," *Google blog*, October 25, 2019: https://blog.google/products/search/search-language-understanding-bert/.

[47]Prabhakar Raghavan, "How AI is powering a more helpful Google," *Google blog*, October 15, 2020: https://www.blog.google/products/search/search-on/.

that "BERT is now used in almost every query in English." And in September 2020, Google announced[48] that BERT was also being used "to improve the matching between news stories and available fact checks." In Chapter 9, I'll cover fact-checking tools in depth; for now, what's relevant here is that BERT is used to automatically scan through lists of human fact-check reports and figure out which ones pertain to a given news article.

But BERT also featured prominently in a recent debacle that landed Google in the news in an unflattering light. Stanford-trained computer scientist Timnit Gebru is one of the world's leading experts on ethics in AI and algorithmic bias; she is a cofounder of the organization *Black in AI*; and, until recently, she was one of the leaders of Google's Ethical Artificial Intelligence Team. But she was abruptly fired from Google in December 2020.[49] She was working on a research paper with several other Google employees when a request from the higher-ups came in asking her to either withdraw the paper or remove the names of all Google employees from it. She refused and demanded to know who was responsible for this bizarre request and their reasoning behind it, but Google leadership rebuffed her demand and instead fired her.

This move was shocking, not just to Gebru but to the entire AI community. Gebru is widely respected and recognized for important pioneering work; Google's iron-fisted handling of this incident did not sit well with most people. The optics of Google censoring and then firing a prominent and beloved Black woman AI researcher from a leadership role on an ethics team were very poor, to say the least. And what was the topic of Gebru's research paper that stirred up all this controversy in the first place? The paper was titled "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", and it was on the dangers—ranging from environmental costs to inscrutability to harmful biases—involved in large deep learning language models like BERT.

We now come to the final topic of this lengthy section, and of the entire chapter, which is how Google has been trying to adjust its algorithm so that accurate information rises to the top of search rankings and fake news is relegated to later pages of search results.

---

[48]Pandu Nayak, "Our latest investments in information quality in Search and News," *Google blog*, September 2020: https://blog.google/products/search/our-latest-investments-information-quality-search-and-news.

[49]Bobby Allyn, "Google AI Team Demands Ousted Black Researcher Be Rehired And Promoted," *NPR*, December 17, 2020: https://www.npr.org/2020/12/17/947413170/google-ai-team-demands-ousted-black-researcher-be-rehired-and-promoted.

Elevating Quality Journalism

In February 2019, at an international security conference in Munich, Google released a white paper[50] on the company's efforts "to tackle the intentional spread of misinformation—across Google Search, Google News, YouTube and our advertising systems." While mostly repeating the philosophies and general approaches that were already sprinkled across various Google blog posts and corporate documents (and already mentioned in this chapter), this white paper does include a few remarks and insights that helpfully shine some additional light on certain details.

In attempting to elevate quality journalism, Google's search ranking algorithm needs to assess the trustworthiness of news sites. The white paper clarifies that these assessments are not just overall measures, they depend specifically on the scope of the search phrase: "For instance, a national news outlet's articles might be deemed authoritative in response to searches relating to current events, but less reliable for searches related to gardening." It also clarifies that the "ranking system does not identify the intent or factual accuracy of any given piece of content." In other words, everything—true and false—is allowed to show up on Google, and Google's search algorithm does not try to determine which particular links contain valid information versus misinformation; instead, it just tries to rank more highly the sites it deems more generally trustworthy in the context of the present search.

In June 2020, the Parliament of the United Kingdom published a policy report[51] with numerous recommendations aimed at helping the government fight against the "pandemic of misinformation" powered by internet technology. The report is rather forceful on the conclusions it reaches:

> The Government must make sure that online platforms bear ultimate responsibility for the content that their algorithms promote. […] Transparency of online platforms is essential if democracy is to flourish. Platforms like Facebook and Google seek to hide behind 'black box' algorithms which choose what content users are shown. They take the position that their decisions are not responsible for harms that may result from online activity. This is plain wrong. The decisions platforms make in designing and training these algorithmic systems shape the conversations that happen online.

[50] Kristie Canegallo, "Fighting disinformation across our products," *Google blog*, February 16, 2019: https://www.blog.google/around-the-globe/google-europe/fighting-disinformation-across-our-products/.

[51] "Digital Technology and the Resurrection of Trust," *House of Lords*, Select Committee on Democracy and Digital Technologies, Report of Session 2019–21: https://committees.parliament.uk/publications/1634/documents/17731/default/.

While preparing this report, Parliament collected oral evidence from a variety of key figures. One of these was Vint Cerf, Vice President and Chief Internet Evangelist at Google. He was asked: "Can you give us any evidence that the high-quality information, as you describe it, that you promote is more likely to be true or in the category, 'The earth is not flat', rather than the category, 'The earth is flat'?" His intriguing response provided a sliver of daylight in the tightly sealed backrooms of Google:

> The amount of information on the world wide web is extraordinarily large. There are billions of pages. We have no ability to manually evaluate all that content, but we have about 10,000 people, as part of our Google family, who evaluate websites. We have perhaps as many as nine opinions of selected pages. In the case of search, we have a 168-page document given over to how you determine the quality of a website. […] Once we have samples of webpages that have been evaluated by those evaluators, we can take what they have done and the webpages their evaluations apply to, and make a machine-learning neural network that reflects the quality they have been able to assert for the webpages. Those webpages become the training set for a machine-learning system. The machine-learning system is then applied to all the webpages we index in the world wide web. Once that application has been done, we use that information and other indicators to rank-order the responses that come back from a web search.

He summarized this as follows: "There is a two-step process. There is a manual process to establish criteria and a good-quality training set, and then a machine-learning system to scale up to the size of the world wide web, which we index." Many of Google's blog posts and official statements concerning the company's efforts to elevate quality journalism come back to this team of ten thousand human evaluators, so to dig deeper into Cerf's dense statement here, it would be helpful to better understand what these people do and how their work impacts the algorithm. Fortunately, an inside look at the job of the Google evaluator was provided in a *Wall Street Journal* investigation[52] from November 2019.

[52]Kirsten Grind, Sam Schechner, Robert McMillan, and John West, "How Google Interferes With Its Search Algorithms and Changes Your Results," *Wall Street Journal*, November 15, 2019: https://www.wsj.com/articles/how-google-interferes-with-its-search-algorithms-and-changes-your-results-11573823753.

While Google employees are very well compensated financially, these ten thousand evaluators are hourly contract workers who work from home and earn around $13.50 per hour. One such worker profiled in the *Wall Street Journal* article said he was required to sign a nondisclosure agreement, that he had zero contact with anyone at Google, and that he was never told what his work would be used for (and remember these are the people Cerf referred to as "part of our Google family"). He said he was "given hundreds of real search results and told to use his judgment to rate them according to quality, reputation and usefulness, among other factors." The main task these workers perform, it seems, is rating individual sites as well as evaluating the rankings for various searches returned by Google. These tasks are closely guided by the hundred-sixty-eight-page document these workers are provided. Sometimes, the workers also received notes, through their contract work agencies, from Google telling them the "correct" results for certain searches. For instance, at one point, the search phrase "best way to kill myself" was turning up how-to manuals, and the contract workers were sent a note saying that all searches related to suicide should return the National Suicide Prevention Lifeline as the top result.

This window into the work of the evaluators, brief though it is, helps us unpack Cerf's testimony. Google employees—presumably high-level ones—make far-reaching decisions about how the search algorithm should perform on various topics and in various situations. But rather than trying to directly implement these in the computer code for the search algorithm, they codify these decisions in the instruction manual that is sent to the evaluators. The evaluators then manually rate sites and search rankings according to this manual, but even with this army of ten thousand evaluators, there are far too many sites and searches to go through by hand—so as Cerf explained, these manual evaluations provide the training data for supervised learning algorithms whose job is essentially to extrapolate these evaluations so that hopefully all searches, not just the ones that have been manually evaluated, behave as the Google leadership intends.

While some of the notable updates to the Google search algorithm have been publicly announced[53] by the company (several were mentioned in this chapter), Google actually tweaks its algorithm extremely often. In fact, the same *Wall Street Journal* investigation just mentioned also found that Google modified its algorithm over thirty-two hundred times in 2018. And the number of algorithm adjustments has been increasing rapidly: in 2017, there were around twenty-four hundred, and back in 2010 there were only around five hundred. Google

[53]These announcements typically appear in Google blog posts, but a convenient list and description of the substantial ones has been collected by a third-party organization called the *Search Engine Journal*: https://www.searchenginejournal.com/google-algorithm-history/.

has developed an extensive process for approving all these algorithm adjustments that includes having evaluators experiment and report on the impact to search rankings. This gives Google a sense of how the adjustments will work in practice before turning them loose on Google's massive user base. For instance, if certain adjustments are intended to demote the rankings of fake news sites, the evaluators can see if that actually happens in the searches they try.

Let me return now to Vint Cerf. Shortly after the question that led to his description of Google's "two-step" process that I quoted above, the chair of the committee asked Cerf another important, and rather pointed, question: "Your algorithm took inaccurate information, that Muslims do not pay council tax, which went straight to the top of your search results and was echoed by your voice assistant. That is catastrophic; a thing like that can set off a riot. Obviously, 99 percent of what you do is not likely to do that. How sensitised are your algorithms to that type of error?" Once again, Cerf's frank answer was quite intriguing. He said that neural networks (which, as you recall, are the framework for deep learning) are "brittle," meaning sometimes tiny changes in input can lead to surprisingly bad outputs. Cerf elaborated further:

> Your reaction to this is, "WTF? How could that possibly happen?" The answer is that these systems do not recognise things in the same way we do. We abstract from images. We recognise cats as having little triangular ears, fur and a tail, and we are pretty sure that fire engines do not. But the mechanical system of recognition in machine-learning systems does not work in the same way our brains do. We know they can be brittle, and you just cited a very good example of that kind of brittleness. We are working to remove those problems or identify where they could occur, but it is still an area of significant research. To your primary question, are we conscious of the sensitivity and the potential failure modes? Yes. Do we know how to prevent all those failure modes? No, not yet.

In short, we trust Google's algorithms to provide society with the answers to all its questions—even though it sometimes fans the flames of hate and fake news and we don't entirely know how to stop it from doing so.

Recall the quote I included earlier from Google's white paper: "[Google's] ranking system does not identify the intent or factual accuracy of any given piece of content." The *Wall Street Journal* investigation discussed in this section noted that Facebook has taken a more aggressive approach to removing misinformation and said Google publicly attributes this difference in approach

to the fact that Facebook actually hosts content whereas Google merely indexes it. But in private a Google search executive told the *Wall Street Journal* that the problem of defining misinformation is incredibly hard and Google "didn't want to go down the path of figuring it out."

Summary

Fake news and harmful misinformation appearing at or near the top of Google search results became a widely discussed topic after Trump unexpectedly won the 2016 election. Many people started to blame Google and the other tech giants for their role in the election and in eroding the very notion of truth. Google responded by making a series of adjustments to its ranking algorithm—often with the assistance of an army of low-paid contract workers—over the ensuing years aimed at bringing trustworthy links to the top of searches and pushing less reliable ones lower down. In this chapter, I presented a variety of examples where this played out, gathered what technical details I could about the closely guarded search algorithm, and looked into the public statements and general strategies Google has employed in this effort to elevate quality journalism. I also discussed instances of misinformation, deception, hateful stereotyping, and blatant racism that surfaced on other corners of Google such as maps, image search, and autocomplete.

In the next chapter, I tackle another aspect of Google: how its advertising platform provides the revenue stream for a huge fraction of the fake news industry. Facebook is also brought into the fray, though its advertising platform fans the flames of fake news in a rather different way, as you will soon see.

CHAPTER

7

Avarice of Advertising

How Algorithmic Ad Distribution Funds Fake News and Reinforces Racism

One of the incentives for a good portion of fake news is money.

—Fil Menczer, Professor of Informatics, Indiana University

When we think of Google supporting the fake news industry, the first thing that comes to mind is how, as described in the previous chapter, it serves up an audience with its various search products. However, there is an entirely separate way—less obvious but extremely influential—that Google supports the fake news industry: financially through ad revenue. The first half of this chapter focuses on the mechanics and scale of Google's algorithmic ad distribution system, the extent to which it funds fake news organizations, and the reluctant steps Google has taken over the years to curtail this dangerous flow of funds.

© Noah Giansiracusa 2021
N. Giansiracusa, *How Algorithms Create and Prevent Fake News*,
https://doi.org/10.1007/978-1-4842-7155-1_7

The second half of the chapter turns attention to Facebook. Here, the issue is not that the company is funding fake news, the issue is that Facebook profits from fake news in the form of political advertisements and in the process exposes a massive audience to fake news. The chapter also takes a deep dive into Facebook's algorithmically powered ad distribution system and details multiple dimensions of racism and discrimination that the system engages in. Here, too, the sequence of reluctant steps the company has taken to mitigate these problems is discussed.

Google Ads and Fake News

Peddlers of fake news surely hope to profit from their nefarious endeavor in some way. For a portion of them, the aim is primarily political—they could be part of a foreign government's disinformation campaign intended to sow chaos and weaken a sovereign nation's democratic pillars, or part of a domestic movement resorting to deceptive means in order to sway popular opinion on certain issues. For many others, however, the aim is quite simply money, and peddling fake news is just a business.

But how does one make money by peddling fake news? Usually, through pageview advertising of the kind discussed in Chapter 1. This is where Google enters the story: while we typically view Google as a search engine or a technology company, from a revenue perspective it is predominantly an advertising company—and it is the biggest one in the world. A report from 2017 found[1] that online advertising had surpassed television to become the largest ad medium, that Google's ad revenue in 2016 (nearly eighty billion dollars) was the biggest in the advertising industry and triple that of the next competitor, Facebook, and that Google and Facebook combined for nearly twenty percent of the earnings of the entire global advertising industry. In 2020, Google's annual ad revenue had grown[2] to nearly one hundred fifty billion dollars, and it accounted for eighty percent of the company's combined revenue. In other words, four out of every five dollars that Google earns comes by way of advertising. And one out of every three dollars made in the US digital advertising industry goes to Google.[3]

---

[1]Julia Kollewe, "Google and Facebook bring in one-fifth of global ad revenue," *Guardian*, May 1, 2017: https://www.theguardian.com/media/2017/may/02/google-and-facebook-bring-in-one-fifth-of-global-ad-revenue.

[2]Joseph Johnson, "Advertising revenue of Google from 2001 to 2020," *Statista*, February 5, 2021: https://www.statista.com/statistics/266249/advertising-revenue-of-google/.

[3]Brad Adgate, "In A First, Google Ad Revenue Expected To Drop In 2020 Despite Growing Digital Ad Market," *Forbes*, June 22, 2020: https://www.forbes.com/sites/bradadgate/2020/06/22/in-a-first-google-ad-revenue-expected-to-drop-in-2020-despite-growing-digital-ad-market/.

There are two ways that Google earns revenue through advertising; to understand them, it helps to think in terms of real estate. The most direct method is that organizations pay Google to display their ads on Google's site. For instance, if you want to advertise a particular product, you can pay Google to place your ad at the top of Google searches that include keywords of your choosing. This is like renting a small piece of property directly from Google where you can place your ad. The second method is more indirect—here again organizations pay Google to place their ads, but this time they are placed on third-party properties rather than Google's own. In this approach, Google is just a middleman, in essence a virtual realtor helping facilitate transactions between different clients. If you want to advertise a product, you tell Google what kind of websites you'd like your ad shown on, and Google will find ones that are willing to host ads and then place yours there for a fee. On the other hand, if you run a popular website and would like to make money from it, just tell Google that a piece of your web property is for rent, and Google will, for a fee, find an advertiser happy to take up your offer. The websites hosting ads placed by Google are officially called the *Google Display Network* (since the sites "display" ads). Due to the massive size of this network and the literally billions[4] of ads placed in it every single day, these transactions are all automated and powered by algorithms.

This brings us back to fake news. When Google places ads on a website that publishes misinformation, Google is providing a revenue stream—often the only one—for this website. But how can Google's ad placement algorithm know if the information on a website is accurate? I'll turn to this topic in the last two chapters of this book; the brief answer is that it is *very* hard. Even if this were doable, there are no laws preventing Google from profiting by placing ads on fake news or dangerously provocative sites. Moreover, popular backlash for this manifestly unethical conduct is limited because users on an ad hosting site see no indication of which company was responsible for the ad placement; the invisible hand of Google's algorithm is hidden from all except those directly involved in the transaction.

In short, Google is financially incentivized to provide ads on as many sites as possible, not just the trustworthy ones. And in doing so, Google is providing a financial incentive for these sites to bring in as much web traffic as possible, even if it comes by way of manipulative disinformation. In this manner, unscrupulous bloggers and journalists profit from sensational attention-grabbing fake news—and so does Google. This connects the main topic of this chapter with the main topic of Chapter 1, the pageview economic corruption of journalism.

[4]John Koetsier, "30 billion times a day, Google runs an ad (13 million times, it works)," *VentureBeat*, October 25, 2012: https://venturebeat.com/2012/10/25/30-billion-times-a-day-google-runs-an-ad-13-million-times-it-works/.

It is time now to take a closer look at Google's role in funding fake news through algorithmic ad placement.

2017 Report

In October 2017, the *Campaign for Accountability (CfA)*—a nonpartisan, nonprofit, evidence-based watchdog organization—released a detailed report[5] on Google's algorithmic placement of ads on hyperpartisan, misleading, and misinformative websites. The report noted that Google's terms of service did not prohibit sites in the Google Display Network from publishing fake news. Additionally, Google's dashboard where advertisers choose what kind of content their ads should be placed on allowed users to select left-wing or right-wing political content, but it did not include any options concerning hyperpartisan content—nor did it make any distinctions between quality journalism and fake news. While Google had discontinued Display Network membership for some particularly egregious sites, the CfA found that Google continued to include many hyperpartisan sites that frequently post disinformation.

Moreover, Google provided an "anonymous" option that allowed ad hosting websites to conceal their identity from the advertising organizations. Advertisers could choose not to place ads on any anonymous sites, but doing so would cut out a significant revenue source since the anonymous sites were disproportionately lucrative. Indeed, the CfA report analyzed a sample of over a thousand partisan news sites in the Google Display Network and found that only fifteen percent were anonymous—but these anonymous ones contributed eight times as much revenue per site compared with the non-anonymous ones, and this anonymous fifteen percent of the sampled sites provided an estimated sixty percent of the total revenue in the sample. As the report authors described it, "advertisers are forced to choose between safe audiences and large audiences, and may unwittingly end up funding groups and activities that are antithetical to their values."

After the 2016 US election, there were some complaints from the public and from a handful of major advertisers over Google's role in supporting hateful and factually misleading sites with ad revenue. Google responded[6] with a statement that included the following assertion: "Moving forward, we will restrict ad serving on pages that misrepresent, misstate, or conceal information

---

[5]Daniel Stevens, "Report: Google Makes Millions from Fake News," *Campaign for Accountability*, October 30, 2017: https://campaignforaccountability.org/report-google-makes-millions-from-fake-news/.

[6]Julia Love and Kristina Cooke, "Google, Facebook Move to Restrict Ads on Fake News Sites," *Reuters*, November 4, 2016: https://www.reuters.com/article/us-alphabet-advertising/google-facebook-move-to-restrict-ads-on-fake-news-sites-idUSKBN1392MM.

about the publisher, the publisher's content, or the primary purpose of the web property." This is a pretty subtle statement: if you read it closely, you notice that it is not about prohibiting sites from publishing misleading content (as a number of news headlines at the time interpreted it), it is just about prohibiting sites in the Display Network from misleading the advertising organizations that they hope to secure and profit from. But even this mild form of restriction seems to have been, shall we say, fake news, because Google evidently continued to allow ad hosts to remain anonymous—the very definition of "conceal information about the publisher," in my opinion.

It's hard to fathom how this blatant contradiction between Google's public statement and the actual inner workings of the ad distribution system is anything but perfidy on the part of Google's corporate leadership. There was no explanation for why anonymity continued to be allowed after this vapid public declaration; an earlier Q&A post[7] simply said "some publishers who've partnered with Google via the AdSense program, and provide us with the slots for placing your ads on the Display Network, choose to offer these placements anonymously and not disclose their site names to advertisers for various reasons." Thanks to the CfA report, we now know the real reason Google continued to allow this: substantial profits.

Complaints about harmful ad placements continued, and in March 2017 Google tried to step up its game:[8] "Starting today, we're taking a tougher stance on hateful, offensive and derogatory content. This includes removing ads more effectively from content that is attacking or harassing people based on their race, religion, gender or similar categories." But the CfA report found that content of this type remained in the Google Display Network. For instance, *Breitbart* was included in the network. Many advertisers explicitly pulled their ads from *Breitbart* once they realized their ads were being placed there, but this had to happen on an ad hoc basis since Google's dashboard lacked the refined controls for preventing such ad placement in the first place. As the CfA report describes: "Under Google's system, it is incumbent upon advertisers to identify and blacklist specific domains that they find objectionable. But Google doesn't make this easy: its ad platforms don't allow advertisers to block fake news sites as a category. […] Even if advertisers could identify specific extreme websites, Google offers these publishers a way to circumvent advertiser exclusions by making their sites anonymous."

---

[7] https://web.archive.org/web/20170905210605/https://www.en.advertisercommunity.com/t5/AdWords-Tracking-and-Reporting/What-does-quot-anonymous-google-quot-mean-in-my-Placement/td-p/473414?nobounce.

[8] Philipp Schindler, "Expanded safeguards for advertisers," *Google blog*, March 21, 2017: https://blog.google/technology/ads/expanded-safeguards-for-advertisers/.

The CfA report found numerous examples of Google ads placed on fake news articles—for instance, ones concerning the Las Vegas mass shooting that took place just weeks before the CfA report was published. Strikingly, in the sample of hosts analyzed by the CfA, right-wing content producers generated sixty-eight percent of the revenue, whereas left-wing content producers generated only four percent. Moreover, the report found that "Hyper-partisan, right-wing websites like *Breitbart*, *Drudge Report* and the *Daily Mail*, which commonly post highly dubious and conspiracy-minded content, were the top revenue-generating publishers in the sample."

While much of the public outcry facing Google in the period after the 2016 election centered on overt disinformation and bias in Google's search results and its suggested news articles—the topics of the previous chapter—it seems that Google's ad placement service, while much less visible, was playing a substantial role in funding politically dangerous organizations. Let us next see if Google's ad placement policies and patterns improved in the years that followed.

2019 Report

In September 2019, a UK nonprofit called the *Global Disinformation Index (GDI)* released a report[9] analyzing ad placement on fake news sites. The researchers behind this report started by collecting a list of seventeen hundred websites that had been flagged by fact-checking organizations such as *PolitiFact* for publishing content that included fake news. They found that Google was serving up ads on a whopping seventy percent of these dubious sites; the second largest contributor was AppNexus, with ads on eight percent of the sites; coming in at third place was Amazon, with four percent.

By estimating the number of monthly pageviews these sites received and then applying a market average figure for ad rates, the GDI researchers produced a ballpark estimate for the annual ad revenue these seventeen hundred sites took home. These sites are believed to be representative of a much larger collection of twenty thousand sites known to publish fake news. By extrapolating the revenue estimate for the seventeen hundred sites to this larger collection, the GDI researchers came to the conclusion that the fake news industry brought in nearly a quarter billion dollars in the year 2019—of which Google was responsible for an estimated eighty-seven million dollars, the largest amount of any ad exchange. This means Google was responsible

[9] "The Quarter Billion Dollar Question: How is Disinformation Gaming Ad Tech?" *Global Disinformation Index*, September 22, 2019: https://disinformationindex.org/wp-content/uploads/2019/09/GDI_Ad-tech_Report_Screen_AW16.pdf.

for almost forty percent of the fake news industry's revenue[10] that year; curiously, this is almost exactly the same share of ad revenue that Google was responsible for among well-respected, factual news sites. In other words, Google had equally dominant shares of the advertising market on both real and fake news sites.

This bears repeating: fake news is a *big* business, worth nearly a quarter billion dollars in 2019 in online ad revenue—at least according to the GDI estimates—and Google was responsible for a larger share of this revenue than any other advertising company. If Google really has been attempting to stanch the flow of funds to fake news sites since the 2016 election, evidently by 2019 there was still quite a lot of room for improvement in this regard.

2021 Report

The news rating company *NewsGuard Technologies* released a report[11] in January 2021 studying ad placement and revenue on sites publishing fake news concerning the 2020 US presidential election. NewsGuard flagged one hundred sixty sites "for publishing falsehoods and conspiracy theories about the election," and it found that between October 1, 2020, and January 14, 2021, these sites ran over eight thousand unique ads from over sixteen hundred different brands; around forty percent of these brands ran ads on more than one of the flagged sites. It is believed that many of the brands did this unknowingly and likely would have been apprehensive to be associated with, and to financially support, electoral disinformation that ultimately played a role in the Capitol building uprising. More than eighty percent of these flagged sites received their algorithmic ad placements from Google.

The two brands that ran ads on the largest number of these flagged sites were Progressive Insurance (which ran nearly three hundred ads across twenty-five of the sites) and Planned Parenthood (which ran seventy-one ads across eighteen of the sites). The AARP, the American Cancer Society, and Sloan Kettering Cancer Center each ran dozens of ads across a handful of these sites. The sites included some of Trump's favorites for peddling his bogus claims of election fraud, such as *One America News Network* and the *Gateway Pundit*. A particularly sad irony is that several well-respected medical organizations inadvertently provided funding, via ad revenue, to fake news sites that published harmful medical disinformation. Family-friendly Disney ran ads on a site that "published claims that COVID-19 was a hoax and promoted

[10]This forty percent is less than the seventy percent mentioned above because revenue is based on traffic—and while Google served ads on seventy percent of the sites, some of the highly trafficked ones were served ads by other providers.

[11]Matt Skibinski, "Special Report: Advertising on Election Misinformation," *NewsGuard*, January 14, 2021: https://www.newsguardtech.com/special-report-advertising-on-election-misinformation/.

false cures for the virus." American Express advertised on *Sputnik News*, an organization controlled by the Russian government that is known for targeting disinformation at American audiences. Even the Department of Veterans Affairs and the Department of Homeland Security were found to have placed a few ads on flagged sites, as was the BBC.

A Google spokesperson said[12] that "Claims that voter fraud was widespread or that the election was stolen are all prohibited by our policies. When we find content that violates our policies we remove its ability to monetize." Google first demonetizes individual stories that violate its policies; it only resorts to sitewide demonetization in cases of persistent, egregious violation. Most of the sites tracked by NewsGuard posted content that quite clearly violates Google's policies—evidently, though, they hadn't crossed Google's threshold for sitewide prohibition.

In August 2020, the leaders of more than a dozen large philanthropic organizations wrote a public letter[13] to Sundar Pichai, CEO of Google's parent company, Alphabet, after it was discovered that ads for Red Cross and others were placed alongside COVID-19 misinformation. They urged Google to institute a new system that "does not put [advertisers] into unwanted and damaging associations that undermine their good works and values." Google seems to have responded by demonetizing individual articles but not repeat offender sites other than in extremely rare instances.

To sum up, the problem of Google funding fake news through algorithmically placed ads—which first came to public awareness around the 2016 election—is evidently still ongoing today. And this is despite multiple public declarations and various tweaks to the algorithm and to company policies allegedly aimed at addressing the issue. Before moving on to the ills of Facebook's algorithmic advertising system, I'd like to take a moment to look at another problematic issue with Google's algorithmic advertisements: racism. All the discussion so far in this section has been about Google's placement of ads on external sites in the Display Network; the following discussion concerns the other form of Google advertising, where ads are placed directly on Google's site and aligned with user-specified keyword searches.

---

[12]Issie Lapowsky, "Google says it's fighting election lies, but its programmatic ads are funding them," *Protocol*, January 14, 2021: https://www.protocol.com/google-programmatic-ads-misinformation.

[13]Issie Lapowsky, "In a letter to Pichai, top philanthropists slam Google for placing charity ads on disinfo sites," *Protocol*, August 13, 2020: https://www.protocol.com/google-ads-charities-disinformation-sites.

Racism in Google Advertising

One of the first scholars to recognize and carefully investigate the harmful impact algorithms can have on society is Harvard professor Latanya Sweeney. In a 2013 research paper,[14] she studied how Google ads for arrest records were appearing more frequently when users searched for names typically associated with Black people than with white people. In fact, she found that Googling her own name, "Latanya Sweeney," resulted in a Google search ad from a company called *Instant Checkmate* suggesting that she had an arrest record (which she did not), while searching for several more stereotypically white names yielded no such ads.

In the conclusion of the paper, Sweeney ponders the causes of this unseemly algorithmic advertising behavior: "Why is this discrimination occurring? Is this Instant Checkmate, Google, or society's fault? Answering […] requires further information about the inner workings of Google AdSense." She goes on to explain that Google allows advertisers to provide multiple different ads for the same keyword search, and Google displays these in a probabilistic manner: initially, they are given equal odds of appearing on the search, but as people click the different versions at different rates, Google's algorithm updates the probabilities so that the more popular ones are then shown more frequently. Instant Checkmate said that the different messages in its ads were grouped by last name, not first name, so the fact that its arrest records ads were appearing more frequently for typically Black names than white means that people were clicking the arrest record ads for Black first names more often than for white first names. Thus, the racism here originated with society at large—but it was enabled and reinforced by Google's algorithm.

More recently, a July 2020 investigation[15] by two investigative reporters for the technology-oriented news site the *Markup* looked into Google's *Keyword Planner*, which suggests terms for advertisers to associate with their ads so that the ads show up on relevant Google searches. They found that the majority of the keywords suggested for the phrases "Black girls," "Latina girls," and "Asian girls" were pornographic, and so were the suggestions for boys of these ethnicities, whereas for "white girls" and "white boys," no keywords were returned at all. The investigators insightfully summarized the situation as follows: "Google's systems contained a racial bias that equated people of color with objectified sexualization while exempting White people from any associations whatsoever. […] By not offering a significant number of non-

[14]Latanya Sweeney, "Discrimination in Online Ad Delivery," *Communications of the Association of Computing Machinery*, Vol. 56 No. 5, 44–54: https://dl.acm.org/doi/10.1145/2460276.2460278.

[15]Leon Yin and Aaron Sankin, "Google Ad Portal Equated 'Black Girls' with Porn," *Markup*, July 23, 2020: https://themarkup.org/google-the-giant/2020/07/23/google-advertising-keywords-black-girls.

pornographic suggestions, this system made it more difficult for marketers attempting to reach young Black, Latinx, and Asian people with products and services relating to other aspects of their lives."

An even more recent investigation by the *Markup* uncovered racist aspects of Google's placement of ads on YouTube (which, as you recall, is owned by Google). In June 2020, just weeks after the police murder of George Floyd, the CEO of YouTube wrote[16] that "We're committed to doing better as a platform to center and amplify Black voices and perspectives. […] At YouTube, we believe Black lives matter and we all need to do more to dismantle systemic racism." But ten months later, in April 2021, the *Markup* found[17] that Google was blocking "Black Lives Matter" as a search phrase for advertisers to find videos and channels to place ads on. This makes it harder for people in the movement to monetize their videos, thereby denying them a valuable revenue source, and it also prevents people who want to place ads supporting the movement from being able to reach an appropriate audience.

When a potential advertiser did a search on the Google Ads platform for the phrase "White Lives Matter" (which the Southern Poverty Law Center describes as a "racist response to the civil rights movements Black Lives Matter"), over thirty million YouTube videos were returned as possible places to place an ad; in contrast, searching "Black Lives Matter" returned zero videos. Over one hundred million videos were returned for the search "White Power," whereas zero videos were returned for "Black Power." After the *Markup* journalists contacted Google with these findings, Google blocked phrases associated with white supremacy like "White Lives Matter" and "White Power" from the ad search, but it did not unblock the corresponding Black phrases—even though it is widely considered that the white phrases are part of hate movements, while the Black phrases are part of legitimate social justice and antiracist movements. Strangely, the *Markup* found that Google even started to block more search phrases like "Black in Tech" and "antiracism." The only conceivable excuse I can imagine for blocking phrases like these and "Black Lives Matter" is that, as the *Markup* reporters point out, it does make it harder for critics to monetize their anti-BLM videos—but to me at least, that doesn't justify the unequal treatment caused by these rather surreptitious algorithmic adjustments.

---

[16]Susan Wojcicki, "Susan Wojcicki: My mid-year update to the YouTube community," *YouTube blog*, June 11, 2020: https://blog.youtube/inside-youtube/susan-wojcicki-my-mid-year-update-youtube-community.

[17]Leon Yin and Aaron Sankin, "Google Blocks Advertisers from Targeting Black Lives Matter YouTube Videos," *Markup*, April 9, 2021: https://themarkup.org/google-the-giant/2021/04/09/google-blocks-advertisers-from-targeting-black-lives-matter-youtube-videos.

And back in September 2017, a *BuzzFeed News* investigation[18] found that when advertisers tried to target audiences with certain bigoted phrases, Google not only allowed it but automatically suggested further offensive search phrases the advertiser should consider using. For instance, when the advertiser used "Why do Jews ruin everything" as the search phrase for targeting an ad, Google suggested also using "the evil Jew" and "Jewish control of banks." These suggestions did not come from the minds of humans at Google—they were based on statistical associations that Google's data-hungry algorithms absorbed from the massive amounts of search data that the company collects. *BuzzFeed* tested the system by placing an ad with these targeted search phrases, and indeed the ad went live and came up when someone searched Google for any of these phrases.

Facebook Ads and Racism

Facebook allows advertisers to pick from among a vast list of categories to target. For example, if you are selling pet products, you can likely find a "dog lovers" category and pay for a *promoted post*, which means your ad will be placed on the newsfeeds of users in this category. In traditional media advertising, the different target audiences an advertiser can aim for are organized manually and are rather broad—with Facebook, the ad categories are created and curated algorithmically, allowing advertisers to really focus their campaigns on niche target audiences. Facebook's algorithm that creates the list of possible ad categories and decides which users are in which ones relies, unsurprisingly, on sophisticated machine learning—but, also unsurprisingly, the details are veiled in corporate secrecy, so we only know the broad outlines.

The algorithm considers content users have explicitly written in their Facebook profiles (for instance, if you list dogs in the *interests* field, then you'll probably get tagged in the dog lovers category), but it also harvests your interests from more implicit information in your online activity, such as your posts (if you share an article about best dog breeds, the algorithm will probably figure out that you are a dog lover), the posts by other users that you like or comment on, who you follow and who is in your friend network, etc. The key to remember here is that Facebook isn't just deciding which advertising groups its nearly three billion users belong in—it is also algorithmically creating and naming these groups. And that's where the problems with bigotry start to arise.

[18]Alex Kantrowitz, "Google Allowed Advertisers To Target People Searching Racist Phrases," *BuzzFeed News*, September 15, 2017: https://www.buzzfeednews.com/article/alexkantrowitz/google-allowed-advertisers-to-target-jewish-parasite-black.

Offensive Ad Categories

In September 2017, journalists at *ProPublica* found[19] a Facebook ad category called "Jew hater" and tried to place a sponsored post in it. Facebook's system responded that the category was too small, it only accepts ads whose target audience is above a minimum cutoff size, so Facebook offered an algorithmically generated suggestion for a second category to jointly target: "Second Amendment." Evidently, Facebook's algorithm had, rather frighteningly, correlated anti-Semites with gun enthusiasts. The journalists decided instead to search for more anti-Semitic categories just to see what was available. And indeed they found others, such as "How to burn jews" and "History of 'why jews ruin the world'." The journalists noted that these would be excellent category choices if you wanted to, say, market Nazi memorabilia or recruit marchers for a far-right rally.

The memberships in these anti-Semitic categories were still too small, so the *ProPublica* journalists added another category: The National Democratic Party of Germany—a far-right, ultranationalist political party. With a combined membership of close to two hundred thousand, they finally hit the mark. They paid Facebook thirty dollars to place a few sponsored posts in these categories, and the ads were approved within fifteen minutes. A week later, they received a report that their ads reached five thousand eight hundred ninety-seven people, generating one hundred one clicks and thirteen *engagements* (a *like* or *share* or comment on the post). After contacting Facebook about this experiment, the particular anti-Semitic categories the journalists had found were removed from Facebook's list of ad categories, but of course this was just a band-aid on a festering wound—the larger problem the journalists were getting at with their investigation remained.

And this wasn't the first serious problem *ProPublica* had exposed in Facebook's algorithmic advertising system.

Racist Exclusionary Advertising

Almost one year earlier, in October 2016, *ProPublica* found[20] that Facebook's ad system allowed advertisers to exclude users by race. When purchasing an ad on Facebook, in addition to choosing target audience categories to reach, a screen came up with an option to "Narrow Audience" that had a prompt for the user to select categories of "demographics, interests, or behaviors" to

[19]Julia Angwin, Madeleine Varner, and Ariana Tobin, "Facebook Enabled Advertisers to Reach 'Jew Haters'," *ProPublica*, September 14, 2017: https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters.

[20]Julia Angwin and Terry Parris Jr., "Facebook Lets Advertisers Exclude Users by Race," *ProPublica*, October 28, 2016: https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race.

exclude from the ad's audience. Among the possible demographics the advertiser could select to exclude were "African American," "Asian American," and "Hispanic." The journalists smartly contextualize this as follows: "Imagine if, during the Jim Crow era, a newspaper offered advertisers the option of placing ads only in copies that went to white readers. That's basically what Facebook is doing nowadays." Such racial exclusion in advertising is prohibited by federal law, at least for employment and housing and credit advertising.

Maybe Facebook thought that if an algorithm comes up with these illegal exclusionary categories instead of a person (though we don't even know if that's what happened), then it's OK. Or maybe Facebook's flimsy reasoning was that it described these categories as "ethnic affinities" rather than ethnicities (despite nesting them under the "demographics" category)—implying that Facebook wasn't letting you exclude Black people from advertisements, just people interested in Black topics. When the *ProPublica* journalists showed these ethnic exclusionary categories to the prominent civil rights lawyer John Relman, he responded: "This is horrifying. This is massively illegal. This is about as blatant a violation of the federal Fair Housing Act as one can find.'' How did Facebook respond, and did Facebook end up in the courtroom for this? I'll come back to these questions momentarily, but first please allow me a quick digression.

Algorithms Could Help Instead of Hurt

For context, and to show that algorithms are capable of both good and bad, it helps to compare Facebook's algorithmically generated racist ad categories—both the targeted groups and the excluded groups—with an example where algorithms are used to help human moderators *reduce* bias in advertising.

The *New York Times* runs an automated filter[21] to catch ads that contain discriminatory phrases such as "whites only," and also to bring to the attention of human moderators any ads with known potentially discriminatory coded phrases such as "near churches" and "close to a country club." The *Times* also rejects housing ads with photos that are too disproportionately white, which is something that an algorithm could help detect. It's not that the algorithms at the *Times* are without flaw, it's just that they are used primarily as tools to assist human moderators—whereas at Facebook the algorithms seem to be the primary arbiters. To be fair, however, Facebook is dealing with a scale that is many orders of magnitude larger than the *Times*, so it isn't really a fair comparison. That said, one should keep in mind that there are many different ways of using algorithms.

[21] See Footnote 20.

Illegal Exclusionary Advertising Continued

Fast-forward to November 2017. This is just over a year since *ProPublica* brought attention to Facebook's option of illegally excluding various ethnic demographics from advertisements, and two months after *ProPublica* brought attention to Facebook providing advertisers with the ability to target offensive and hateful interest groups. It is also nine months after Facebook announced[22] that it had taken several steps to strengthen the procedures it uses to prevent discriminatory advertising, especially in the areas of housing, employment, and credit—the three areas in which federal law prohibits discriminatory ads. The *Washington Post* headline reporting this company announcement read "Facebook cracks down on ads that discriminate." But *ProPublica* conducted another investigation at this time and found that Facebook's supposed improvements fell far short of the bold proclamation.

Indeed, *ProPublica*'s November 2017 investigation[23] showed that it was still possible for ad purchasers to select excluded audience categories such as "African Americans, mothers of high school kids, people interested in wheelchair ramps, Jews, expats from Argentina, and Spanish speakers." All of these groups are protected under the federal Fair Housing Act, so Facebook was still quite clearly in violation of federal law. Facebook's response this time? "This was a failure in our enforcement and we're disappointed that we fell short of our commitments. The […] ads purchased by *ProPublica* should have but did not trigger the extra review and certifications we put in place due to a technical failure." A remarkably vague excuse/reassurance.

The *ProPublica* journalists reported that from their experience as ad purchasers, the only difference they noticed when placing these illegal ads compared to one year earlier is that the "Ethnic Affinity" category had been renamed "Multicultural Affinity" and it had been moved from the "Demographics" section to the "Behaviors" section. In other words, Facebook still seemed to be implicitly hiding behind the flimsy excuse that advertisers were not excluding people based on their race but based on their… racial behavior? Good grief. The *ProPublica* journalists also had no trouble *redlining*—which is to say, targeting their Facebook ads at specific ZIP codes as a way of targeting specific racial groups. This is also prohibited by federal law. And yet, Facebook undauntedly and unabashedly continued this manifestly malfeasant activity for another sixteen months, until…

---

[22]"Improving Enforcement and Promoting Diversity: Updates to Ads Policies and Tools," *Facebook newsroom*, February 8, 2017: https://about.fb.com/news/2017/02/improving-enforcement-and-promoting-diversity-updates-to-ads-policies-and-tools/.

[23]Julia Angwin, Ariana Tobin, and Madeleine Varner, "Facebook (Still) Letting Housing Advertisers Exclude Users by Race," *ProPublica*, November 21, 2017: https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.

Legal Action

After years of pressure by civil rights advocates and legal organizations, in March 2019 the ACLU announced that Facebook finally agreed to make sweeping changes to its advertising platform as the result of a settlement arising from multiple legal cases. The ACLU announcement[24] lauded this as a "historic civil rights settlement" and detailed some of the "major changes" Facebook would undertake: "In the first-of-its-kind settlement announced today, Facebook has agreed to create a separate place on its platform for advertisers to create ads for jobs, housing, and credit. Within the separate space, Facebook will eliminate age- and gender-based targeting as well as options for targeting associated with protected characteristics or groups. Targeting based on ZIP code or a geographic area that is less than a 15-mile radius will not be allowed." Facebook said it would enact these changes by the end of the year—which is to say, it allowed itself to continue violating federal law for another nine months.

Then, just one week after this ACLU settlement was announced, Facebook was charged in federal court by the US Department of Housing and Urban Development (HUD) for violating the Fair Housing Act. It's rather strange that the federal government didn't act upon the clear violations of the Fair Housing Act that were uncovered by *ProPublica* two and a half years earlier—and which were left in place this entire time—and that the charges were finally brought literally days after the ACLU's landmark settlement that was intended to once and for all halt Facebook's habitual violations. Political motivations may have played a role, as the Trump administration was chafing at some of the actions of Facebook and the other tech giants that some felt were stifling right-wing discourse on the platforms. When reporting the news of this HUD charge, the *Washington Post* declared[25] "The Trump administration delivered its first sanction of a tech giant" and revealed that HUD had already alerted Twitter and Google that it is scrutinizing their practices for similar violations. Whatever the motivation, I strongly believe this was the right, albeit long overdue, course of action from HUD.

Unfortunately, however, it would soon be found that Facebook's problems with racism run deeper than initially believed.

---

[24]Galen Sherwin and Esha Bhandari, "Facebook Settles Civil Rights Cases by Making Sweeping Changes to Its Online Ad Platform," *ACLU blog*, March 19, 2019: https://www.aclu.org/blog/womens-rights/womens-rights-workplace/facebook-settles-civil-rights-cases-making-sweeping.

[25]Tracy Jan and Elizabeth Dwoskin, "HUD is reviewing Twitter's and Google's ad practices as part of housing discrimination probe," *Washington Post*, March 28, 2019: https://www.washingtonpost.com/business/2019/03/28/hud-charges-facebook-with-housing-discrimination/.

Algorithmic Bias

Eight months after the HUD suit was announced, a research paper was published[26] by a group of scholars led by two professors at Northeastern University showing how eliminating prohibited discrimination in Facebook's advertising system is much more difficult than simply removing excluded categories from the interface. The reason is that societal bias is baked into the data that drive the machine learning algorithm Facebook uses to determine which users to show which ads.

Facebook lets each ad purchaser choose one of three different metrics for the algorithm to maximize: the number of views an ad gets, the number of clicks and amount of engagement it receives, or the quantity of sales it generates. This sets up ad distribution as a supervised learning task: based on the ad's content and the selected audience categories, the algorithm predicts which Facebook users will result in the highest value of the selected metric. The ad is then displayed in the newsfeeds of users according to these predictions. Since this is supervised learning, the predictions are based on data of prior user behavior, and that's where the problem of bias comes in: the algorithm will hunt for whatever patterns it can find in the data that might give an edge in optimizing the selected metric—even if these patterns reflect historical bias. In short, Facebook's ad algorithm learns racism, sexism, and other forms of bias from the training data it is fed (which reflects these forms of bias in society); then it uses this bias to try to boost the chosen ad metric, which leads to biased outcomes that in turn push society further down the road of racism and sexism and discrimination. This is yet another pernicious data-driven algorithmic feedback loop, similar to the one briefly discussed in Chapter 5 in the context of bias in algorithmic lie detection for employment screening.

The specific findings of this research study illustrate the severity and scope of bias in Facebook's ad algorithm. The researchers created an employment ad for doctors and found that Facebook provided it with an audience that was forty-four percent white, whereas a nearly identical ad for janitors was given an audience that was only thirty-six percent white. The percentage was even higher for jobs in AI (fifty-three) and even lower for taxi driver ads (twenty-nine). The ad purchasers here did not choose different targeted audiences, they just changed the words in the ad content and let the algorithm do its thing. Similarly, ads for AI jobs were shown to an audience that was fifty-four percent male, whereas for nursing jobs it was only thirty-seven percent, for secretaries it was twenty-six percent, and for jobs in a preschool it was only

[26]Muhammad Ali et al., "Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes," *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3, CSCW, Article 199 (November 2019). ACM, New York, NY: https://dl.acm.org/doi/10.1145/3359301.

twenty-four percent male. The researchers also found that Facebook's algorithm provided ads for home sales with an audience that was about three-quarters white, whereas the audience for home rental ads was only about half white. The enormity of Facebook's advertising platform (it controls twenty-two percent of the US market share for digital ads, second only to Google[27]) means that even small differences in percentages here lead to huge differences in the total number of people viewing these different ads.

In sum, this Northeastern University study found that Facebook's algorithm detected and utilized—which in turn means perpetuated and amplified—existing biases that put white men in higher-paying and more prestigious jobs than minority women, and it encouraged homeownership for white people more than for non-white people. How might a Facebook ad purchaser attempt to compensate for this kind of automated machine learning bias? Other than waiting for Facebook to drastically redesign its system from within, essentially the only option one had was to explicitly target the underrepresented audiences, for instance, choosing Black audiences for tech jobs and home sales advertisements. But, in somewhat of a Catch 22 irony, the option for such targeting is, as you recall, precisely what was removed by Facebook in an effort to eliminate bias in its ad distribution system.

Corporate Progress

In July 2020, the *Wall Street Journal* reported[28] that Facebook was "creating new teams dedicated to studying and addressing potential racial bias on its core platform and Instagram unit, in a departure from the company's prior reluctance to explore the way its products affect different minority groups." This followed—and likely resulted from—sustained pressure from civil rights groups (including the Anti-Defamation League, Color of Change, and the NAACP), increasingly vocal employee unrest, and a months-long advertising boycott (including large, prominent clients such as Coca-Cola, Disney, McDonald's, Starbucks, and Walmart) that cost Facebook advertising revenue and global reputation. These new equity and inclusion teams at Facebook aim to study how all of the company's products and algorithms—not just its targeted advertising platform—impact minority users and how the user experience for Black, Hispanic, and other ethnic groups differs from that of white users. Let's hope these teams are given the resources and respect they need to make a positive impact at the company.

---

[27]Felix Richter, "Amazon Challenges Ad Duopoly," *Statista*, February 21, 2019: https://www.statista.com/chart/17109/us-digital-advertising-market-share/.

[28]Deepa Seetharaman and Jeff Horwitz, "Facebook Creates Teams to Study Racial Bias, After Previously Limiting Such Efforts," *Wall Street Journal*, July 21, 2020: https://www.wsj.com/articles/facebook-creates-teams-to-study-racial-bias-on-its-platforms-11595362939.

What of Fake News?

Since Facebook, unlike Google, only places ads on its own site, there is no risk of funding fake news organizations the way Google does. The only risk here is that fake news organizations might post ads on Facebook that users see while scrolling through their newsfeed, and the ads could be misinterpreted as actual news, despite a "sponsored" label on the ads. This is a significant concern—it is estimated[29] that during the 2018 midterm elections, approximately four hundred million dollars was spent on political advertising on Facebook in the United States and users were receiving on average twelve political ads per day—but for the most part, this boils down to company policies rather than algorithmic behavior. As such, this aspect of fake news falls outside the purview of this book—other than that, as I will explore in the next chapter, Facebook uses algorithms to help moderate its platform and enforce its policies. That said, it's worth taking a moment to highlight some of the significant developments and events concerning Facebook's stance on fake news in advertising:

- One week after the 2016 election, Facebook updated its advertising policies to clarify that its prohibition of deceptive and misleading content includes a ban on fake news:[30] "We do not integrate or display ads in apps or sites containing content that is illegal, misleading or deceptive, which includes fake news." An academic study[31] came out later that aimed to measure the impact of this policy update. The authors used Twitter as a control group since it did not implement any advertising policy changes at the time, and they found that shares on Facebook of fake news articles about childhood vaccines dropped by seventy-five percent after this update compared to Twitter.

[29]Kim et al., "The Stealth Media? Groups and Targets behind Divisive Issue Campaigns on Facebook," *Political Communication* Vol. 35 Issue 4, July 2018: https://doi.org/10.1080/10584609.2018.1476425.

[30]Julia Love and Kristina Cooke, "Google, Facebook move to restrict ads on fake news sites," *Reuters*, November 14, 2016: https://www.reuters.com/article/us-alphabet-advertising/google-facebook-move-to-restrict-ads-on-fake-news-sites-idUSKBN1392MM.

[31]Lesley Chiou and Catherine Tucker, "Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement," *SSRN*, July 6, 2018: https://doi.org/10.2139/ssrn.3209929.

- In August 2017, Facebook announced[32] that it would try to limit the spread of fake news by banning pages that have repeatedly shared disinformation from advertising on Facebook: "We've found instances of Pages using Facebook ads to build their audiences in order to distribute false news more broadly. Now, if a Page repeatedly shares stories that have been marked as false by third-party fact-checkers, they will no longer be able to buy ads on Facebook."
- In October 2019, Facebook denied a request from the Biden campaign to take down a video ad from the Trump campaign that falsely claimed Biden offered Ukraine a billion dollars in foreign aid if it halted the investigation of a company tied to his son.[33] The ad racked up five million views within a couple weeks. In a written response to the Biden campaign's denied request, Facebook's head of global elections policy hinted that the company has chosen to follow a more laissez-faire approach to moderating political advertising: "Our approach is grounded in Facebook's fundamental belief in free expression, respect for the democratic process, and the belief that, in mature democracies with a free press, political speech is already arguably the most scrutinized speech there is." I don't see how to reconcile this statement with the company's post-2016 election claim that it doesn't allow fake news in ads, yet I haven't been able to find any public statements overtly stating that that earlier policy has been revoked.

[32]Satwik Shukla and Tessa Lyons, "Blocking Ads From Pages that Repeatedly Share False News," *Facebook newsroom*, August 28, 2017: https://about.fb.com/news/2017/08/blocking-ads-from-pages-that-repeatedly-share-false-news/.

[33]Cecilia Kang, "Facebook's Hands-Off Approach to Political Speech Gets Impeachment Test," *New York Times*, October 8, 2019: https://www.nytimes.com/2019/10/08/technology/facebook-trump-biden-ad.html.

- Days later, Elizabeth Warren's presidential campaign ran a deliberately, and provocatively, false political ad on Facebook to illustrate that the platform was not doing enough to limit fake news in political advertising.[34] The ad shockingly read: "Breaking news: Mark Zuckerberg and Facebook just endorsed Donald Trump for re-election. You're probably shocked, and you might be thinking, 'how could this possibly be true?' Well it's not. (Sorry.) But what Zuckerberg *has* done is given Donald Trump free rein to lie on his platform—and then to pay Facebook gobs of money to push out their lies to American voters." In a subsequent tweet, Warren wrote "We intentionally made a Facebook ad with false claims and submitted it to Facebook's ad platform to see if it'd be approved. It got approved quickly and the ad is now running on Facebook." In another tweet, she wrote "Facebook changed their ads policy to allow politicians to run ads with known lies—explicitly turning the platform into a disinformation-for-profit machine. This week, we decided to see just how far it goes."
- On January 9, 2020, Facebook announced[35] the introduction of additional controls on the platform that allow users to reduce the amount of political content they see, if so desired. The announcement also included the following statement on the company's political ad policy: "In the absence of regulation, Facebook and other companies are left to design their own policies. We have based ours on the principle that people should be able to hear from those who wish to lead them, warts and all, and that what they say should be scrutinized and debated in public." The post practically begs for governmental regulation so that the tech giants don't have to each chart their own course on political censorship and fact-checking: "Frankly, we believe the sooner Facebook and other companies are subject to democratically accountable rules on this the better."

[34]Elizabeth Culliford, "Warren campaign challenges Facebook ad policy with 'false' Zuckerberg ad," *Reuters*, October 12, 2019: https://www.reuters.com/article/us-usa-election-facebook/warren-campaign-challenges-facebook-ad-policy-with-false-zuckerberg-ad-idUSKBN1WRONU.

[35]Rob Leathern, "Expanded Transparency and More Controls for Political Ads," *Facebook newsroom*, January 9, 2020: https://about.fb.com/news/2020/01/political-ads/.

- On January 24, 2020, Donald Trump's reelection campaign ran political ads on Facebook claiming that the "Fake news media" will block his upcoming Super Bowl ad.[36] This was patently false—in fact, for a network TV station to block his ad would be a violation of FCC regulations. The Facebook ad also encouraged users to "DEMAND THAT THE LIBERAL MEDIA AIRS OUR AD," which is particularly ironic since the Super Bowl, and hence Trump's ad, was broadcast on Murdoch-owned Fox.
- In June 2020, Facebook announced[37] that it would disallow ads from media organizations that are "wholly or partially under the editorial control" of foreign governments: "later this summer we will begin blocking ads from these outlets in the US out of an abundance of caution to provide an extra layer of protection against various types of foreign influence in the public debate ahead of the November 2020 election in the US."

Concluding Thoughts

In November 2019, Google announced in a blog post[38] that it would be "making a few changes to how we handle political ads on our platforms globally" in order to "help promote confidence in digital political advertising and trust in electoral processes worldwide." After explaining how the new policy would further limit microtargeting in political advertising, the post has a section titled "Clarifying our ads policies" that includes the following: "It's against our policies for any advertiser to make a false claim—whether it's a claim about the price of a chair or a claim that you can vote by text message, that election day is postponed, or that a candidate has died." This sounds great, but what Google failed to mention here—and it continues to sweep this under the rug—is that even when the ads do not make false claims, they might be placed by Google's algorithms on a fake news site. If Google really wants to promote trust in electoral processes, it is not enough to prevent false advertising—Google needs to stop funneling nearly a hundred million dollars a year to fake news publishers.

---

[36]Brian Fung, "Trump campaign runs hundreds of misleading Facebook ads warning of Super Bowl censorship," *CNN*, January 24, 2020: https://www.cnn.com/2020/01/24/media/trump-super-bowl-facebook-ad/index.html.

[37]Nathaniel Gleicher, "Labeling State-Controlled Media On Facebook," *Facebook newsroom*, June 4, 2020: https://about.fb.com/news/2020/06/labeling-state-controlled-media/.

[38]Scott Spencer, "An update on our political ads policy," *Google blog*, November 20, 2019: https://blog.google/technology/ads/update-our-political-ads-policy/.

In the immediate aftermath of the 2016 election, Facebook publicly asserted that fake news is prohibited from the platform's political advertisements. As you have seen, this stance seems to have dissolved in the lead-up to the 2020 election as numerous blatantly false political accusations were allowed in political ads—and in the case of the Trump campaign's video ad about Biden's interactions with Ukraine, a formal request to take down the blatantly untrue ad was formally rebuffed by Facebook. From 2019 onward, Facebook issued a series of policy statements that, while somewhat vaguely worded, seem to suggest the company no longer believes in disallowing disinformation in political ads. One gets the impression from Google and Facebook that they will not take a stronger stance against profiting from fake news until the government steps in and develops regulation in this sector. And indeed, why should they walk away from this source of revenue when there is no real incentive to do so?

One particular challenge with fake news showing up in social media advertisements comes from the way that the distribution algorithms allow for extreme microtargeting. In a remarkable article[39] from 2012, the techno-sociologist Zeynep Tufekci (whom you previously encountered in Chapter 4 with a critique of YouTube's recommendation algorithm) presciently wrote that "Misleading TV ads can be countered and fact-checked," but a misleading microtargeted ad "remains hidden from challenge by the other campaign or the media." In other words, the algorithmic distribution system for ads used by Facebook and other platforms shields misleading ads from public scrutiny because there is no public record of all the ads that are shown—there is only the algorithmically tailored experience of each individual user. This means it is important for each user to take an active role in fact-checking; I'll discuss some tools and approaches for this in Chapter 9.

Summary

Google is the largest advertising company in the world, and it serves ads in two ways: by placing them on its own site—for instance, to appear when users do keyword searches—and by placing them on external sites in the so-called Google Display Network. Many sites in this network have a proven track record of publishing fake news, and yet they remain in the network. The result is that Google is funneling huge sums of money to fake news

[39] Zeynep Tufekci, "Beware the Smart Campaign," *New York Times*, November 16, 2012: https://www.nytimes.com/2012/11/17/opinion/beware-the-big-data-campaign.html.

publishers—and it is profiting tremendously in the process. Facebook, on the other hand, only services ads on its own properties, but numerous investigations have demonstrated that its algorithmically powered ad distribution system has a years-long track record of engaging in federally prohibited discriminatory practices. Meanwhile, the company's policies on fake news in political advertisements have proven rather mercurial and seem to have recently tended toward a laissez-faire loosening under the guise of free speech adherence. In the next chapter, I take a direct look at how Facebook and Twitter have dealt with the spread of fake news on their platforms.

CHAPTER

8

Social Spread

Moderating Misinformation on Facebook and Twitter

> *At a moment of rampant disinformation and conspiracy theories juiced by algorithms, we can no longer turn a blind eye to a theory of technology that says all engagement is good engagement.*[1]
>
> —Apple CEO Tim Cook

In this chapter, I explore several ways in which algorithms interact with the complex dynamics of social media when it comes to fake news. First, I set the stage with some context, issues, and examples that help us better understand what has happened and what's at stake. Next, I look at how algorithms have been used to scrape data from social media platforms to provide remarkable quantitative insight into how fake news spreads—both organically and when part of deliberate disinformation campaigns. Along the way, the role in this spread played by the social media platforms' own content recommendation algorithms is explored. Attention is then turned to the algorithmic tools that social media companies—primarily Facebook and Twitter—have used and

[1]Stephen Nellis, "Apple's Tim Cook criticizes social media practices, intensifying Facebook conflict," *Reuters*, January 28, 2021: https://www.reuters.com/article/us-apple-facebook/apples-tim-cook-criticizes-social-media-practices-intensifying-facebook-conflict-idUSKBN29X2NB.

© Noah Giansiracusa 2021
N. Giansiracusa, *How Algorithms Create and Prevent Fake News*,
https://doi.org/10.1007/978-1-4842-7155-1_8

could potentially use in their battle against harmful misinformation, as well as the limitations and challenges of taking algorithmic approaches to this thorny, multifaceted problem.

Setting the Stage

In this section, I collect some background information that will help frame and inform the discussions in the following sections. To start out, let me take a look at how one's media diet correlates with one's interest and knowledge in current events and with one's exposure to fake news.

Those Who Rely Primarily on Social Media

Pew surveys conducted between October 2019 and June 2020 found[2] that nearly one in five US adults said they turn most to social media for political and election news—and among those under thirty years old, this figure was nearly one in two. Fewer than one in ten of those who relied primarily on social media said they were following news about the 2020 election very closely, whereas for those who relied primarily on cable TV or print news, around one in three said they were following it closely. The proportion of people who said they were following the coronavirus outbreak closely was twice as high among people who got their news primarily from cable TV or national network TV or news websites and apps as it was for people who relied primarily on social media.

Not only were social media–oriented participants less engaged in political and medical news as measured by self-assessment, they were factually less knowledgeable as well: the Pew researchers gave a brief quiz and found that only seventeen percent of those in the primarily social media–informed group scored highly on basic political knowledge, whereas for those who relied primarily on cable TV or national network TV, this figure was around one-third, and for those who relied primarily on print, radio, or news sites/apps, this figure was over forty percent. Meanwhile, exposure to false conspiracy theories related to the pandemic was higher for people who relied primarily on social media for news than it was for all the other media categories[3]—yet

[2]Amy Mitchell et al., "Americans Who Mainly Get Their News on Social Media Are Less Engaged, Less Knowledgeable," *Pew Research Center*, July 30, 2020: https://www.journalism.org/2020/07/30/americans-who-mainly-get-their-news-on-social-media-are-less-engaged-less-knowledgeable/.

[3]A separate study in the UK found that half of parents of small children had been exposed on social media to misinformation about vaccines: Sarah Boseley, "A report from January 2019 in the U.K. found that half of parents with small children had been exposed to misinformation about vaccines," *Guardian*, January 24, 2019: https://www.theguardian.com/society/2019/jan/24/anti-vaxxers-spread-misinformation-on-social-media-report.

the percentage of people who said they were concerned about the impact of misinformation on the 2020 election was lower for the social media group than it was for all others except for people who relied primarily on local news. Being exposed to conspiracy theories is not the same as believing them—to some extent, the fact that the social media group saw more misinformation but was less concerned by it might mean that this group was more adept at telling fact from fiction. But the lower performance on the knowledge quiz undercuts this sanguine interpretation.

A difficult question here is that of causation versus correlation: does relying primarily on social media, and hence being exposed to more misinformation, cause people to lose touch with reality, or is it instead that people who are less interested in, and knowledgeable about, the world are inherently more drawn to social media consumption? While an important question for some considerations, the answer is largely immaterial for the present discussion: the bottom line here is that a lot of people—especially young people—look to social media to understand what is happening in the world, and not all of what they see is true.

Next, let me turn to some background and context on the role social media algorithms play in spreading fake news.

Social Media Algorithms

For the first decade of Twitter's existence, algorithms suggested who to follow but were otherwise not involved in the newsfeed: what you saw on the platform was simply a chronological listing of all the posts by people you followed. Then, in 2016, the company took a more active hand in shaping your personal experience by creating machine learning algorithms to decide which posts should be shown to you first, ordering them based on popularity and your estimated interest in them rather than just ordering them based on their time stamps. As recently as early 2019, Twitter took this even further when it started showing you not just tweets from people you follow but also certain tweets (selected algorithmically) from users followed by the accounts you follow.

These both may at first seem like innocuous steps, but it is important to recognize that any time an algorithm decides what content to show you, there is a risk that misinformation will get algorithmically amplified beyond the confines in which it would normally prosper organically. Social media algorithms are typically designed to maximize various user engagement metrics, so if a harmful conspiracy theory—such as COVID-19 being a government-developed biological weapon or the 2020 election being stolen—generates a lot of engagement, then the algorithms pick up on this and display the misinformation more prominently and broadcast it to a wider audience. Needless to say, this applies not just to Twitter but also to Facebook's newsfeed and to any other ranking and recommendation algorithms in social media.

An illustrative example where you can really see this in action concerns the *Epoch Times*—the dubious news organization you first encountered in Chapter 2 in the context of algorithmically mass-produced low-quality news and fake news, then again in Chapter 4 where a network of fake news channels covertly connected to the *Epoch Times* sprang up on YouTube peddling the lie that the 2020 election victory was stolen from Trump. An investigation[4] by the *New York Times* in October 2020 detailed the development of the *Epoch Times* from "a small, low-budget newspaper with an anti-China slant that was handed out free on New York street corners" to "one of the country's most powerful digital publishers" and "a leading purveyor of right-wing misinformation."

The *Epoch Times* experienced a slow but steady growth from its inception in 2000 through 2014, at which point the organization's finances were solid and the newspaper even had some journalism awards under its belt. But in 2015 its traffic and ad revenue suddenly slipped, allegedly as the result of one of Facebook's modifications to the newsfeed algorithm. The *Epoch Times* responded by having its reporters churn out as many as five articles a day in search of viral hits, much of it lowbrow clickbait. Then as the 2016 election neared, the coverage took a sharp turn to the right: the organization started fervently supporting Donald Trump and various conspiracy theories in his orbit. Outwardly, this was painted as a natural political connection: Donald Trump was openly critical of the Chinese government, and this nicely aligned with the *Epoch Times*' roots in the Falun Gong movement that had long been persecuted by the Chinese government. But the *New York Times* investigation uncovered a hidden motive to this as well: the *Epoch Times* developed a secret strategy to leverage Facebook's platform and algorithms in order to rapidly grow the organization's following and finances, and the choice to align with Trump and his sea of popular disinformation was not just political—it was also a Machiavellian way to tap into virality.

The *Epoch Times*' Facebook strategy began as an experiment in Vietnam, where the organization's local division created a network of Facebook pages filled with entertaining viral videos as well as pro-Trump content—much of it directly copied from other sites—then they used bots to artificially inflate the likes and shares on this material, in direct violation of Facebook's policies against inauthentic accounts. This artificial popularity was then unknowingly picked up by Facebook's algorithms and transformed into actual popularity: the rapid growth in these pages led them to be recommended to a wide user base, and the combination of lighthearted videos and far-right content (much of it sensational fake news) found an enthusiastic audience awaiting. Internal emails obtained by the *New York Times* showed that the *Epoch Times*' leadership praised this as the way forward, and soon the Vietnam model was exported internationally to the rest of the organization's operations. This "Facebook

[4]Kevin Roose, "How The Epoch Times Created a Giant Influence Machine," *New York Times*, October 24, 2020: https://www.nytimes.com/2020/10/24/technology/epoch-times-influence-falun-gong.html.

strategy" was successful: by incubating dozens of Facebook pages the way it had in Vietnam, the *Epoch Times* has now grown to have tens of millions of followers across its social media presence. And the Trump connection continued to work and reached new heights: articles by the *Epoch Times* were shared to massive audiences by Trump and members of his family during his presidency, and the *Epoch Times* even had members of the Trump administration sit for interviews with its reporters.

In short, the *Epoch Times* strategically used Facebook's recommendation algorithms to build itself into a massive media empire. It did this in part through prohibited methods—using bots to generate artificial engagement—and in part through allowed but arguably very unethical and dangerous methods: peddling far-right fake news, often in the guise of balanced authentic journalism, in order to build a popular following. That said, it is important to recognize that not all of the societal problems frequently discussed today with social media platforms are caused by their machine learning ranking/recommendation algorithms, and yet the popular vilification of social media often focuses on these algorithms without providing much evidence that they are indeed the root of the problem. The *New York Times* investigation into the *Epoch Times* relied on internal emails and discussions with current and former employees to detail the organization's deliberate approach centered on Facebook's algorithms. We don't know if the algorithms really had the impact ascribed to them there, but at least we know that the corporate leaders of the *Epoch Times* believed they did, and that their strategy worked.

A January 2021 article[5] in the *New York Times* claims that "Facebook's algorithms have coaxed many people into sharing more extreme views on the platform—rewarding them with likes and shares for posts on subjects like election fraud conspiracies, Covid-19 denialism and anti-vaccination rhetoric." To back up this assertion, the article profiled several individuals who have been posting this kind of content, and the article says that "a journey through their feeds offers a glimpse of how Facebook rewards exaggerations and lies." These particular individuals all followed a similar path: they tried in vain for years to amass social media followings by posting nonpolitical content; then eventually they tried posting pro-Trump content and far-right disinformation and found this yielded a sharp increase in their number of followers. But the details in the article, in my opinion, don't adequately justify the article's strong narrative of algorithmic blame.

The first individual profiled in the article is said to have daily written comments on Trump's Facebook posts in order to generate interest in the individual's own Facebook page. I fail to see how this is a matter of an algorithm gone awry—it's just a user going to a popular location to find people with sympathetic

[5]Stuart Thompson and Charlie Warzel, "They Used to Post Selfies. Now They're Trying to Reverse the Election." *New York Times*, January 14, 2021: https://www.nytimes.com/2021/01/14/opinion/facebook-far-right.html.

views where he can syphon off some of the attention. The article goes on to explain that "Most realized that the same post on a personal page generated only scant attention compared with the likes, shares and comments it could get on a group page. Facebook groups for like-minded people are where lies begin to snowball, building momentum, gaining backers and becoming lore." This is a critique of Facebook as a platform for allowing people with similar views to come together, not a critique of recommendation algorithms. While recommendation algorithms certainly play a role in augmenting the popularity of Facebook groups, the article's inveighing against algorithms seems rather unsubstantiated based on the evidence presented in the article. In fact, somewhat ironically for an article about people espousing views for the sake of popularity, this strikes me as the authors jumping on the anti-algorithm bandwagon because they know that's what readers want to see.

But just because many mainstream news articles do not provide the evidence to support their narrative of algorithmic culpability does not mean the evidence doesn't exist. The evidence does exist and is generally quite damning, as you'll see in this chapter. For instance, a recent study[6] of the 2020 election found that far-right content generated more engagement than any other partisan group, and far-right misinformation generated sixty-five percent more engagement than far-right factual content. This means that any recommendation/ranking algorithm with engagement as the metric it aims to maximize will prioritize far-right misinformation above all else. And it is no secret that Facebook and most other social media platforms are built around the goal of maximizing engagement, and they bake this priority into all of their algorithms.

Later in this chapter, I'll detail the algorithmic approaches Facebook has used in the battle against misinformation spreading on its platform—but first, while still setting the stage here, it helps to preview the topic by both stepping back to see the larger context of fake news on Facebook and also zooming in on a few concrete examples that illustrate the issues involved.

Facebook's Problems and Reactions

The *Wall Street Journal* reported[7] that from the morning of January 6, 2021 (the day of the Capitol building insurrection) to the afternoon, Facebook's internal team of data scientists noted a tenfold increase in user-reported violent content on the platform; user reports of fake news surged to forty

[6]Laura Edelson et al., "Far-right news sources on Facebook more engaging," *Medium*, March 3, 2021: https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging-e04a01efae90.

[7]Jeff Horwitz and Deepa Seetharaman, "Facebook Turned on Trump After Warnings That 'Business as Usual Isn't Working'," *Wall Street Journal*, January 13, 2021: https://www.wsj.com/articles/facebook-turned-on-trump-after-warnings-that-business-as-usual-isnt-working-11610578907.

thousand per hour, which was quadruple the peak from prior days. Leadership at Facebook feared a dangerous feedback loop in which incendiary material online inspires more violent real-world action which then leads to even more attention on social media and so on. We all know that Facebook responded within hours by taking down posts by President Trump and announcing that he was suspended from the platform indefinitely. But Facebook also privately designated the United States a "temporary high-risk location" for political violence, which "triggered emergency measures to limit potentially dangerous discourse" on the platform, though it was not revealed exactly what this meant. Facebook tried to dodge the blame for the events that transpired on January 6 when Sheryl Sandberg, the Chief Operating Officer, brazenly said that "these events were largely organized on platforms that don't have our abilities to stop hate, don't have our standards and don't have our transparency."

In the days after the Capitol building insurrection, it was reported[8] that Facebook was showing ads for weapons accessories and body armor in "patriot" and militia-themed Facebook groups alongside pro-Trump disinformation about the election being rigged and stolen. Two days later, three US Senators co-authored a public letter to Facebook founder and CEO Mark Zuckerberg urging him to take immediate action to halt these ads that they described as "designed to equip white nationalists, neo-Nazis and other domestic extremist organizations." The next day, Facebook acquiesced and announced it was immediately halting all such ads for a week, only allowing them to return after the upcoming inauguration. But the day after this announcement, it was found that many of these dangerous ads had not been taken down. A few weeks later, Facebook began experimentally reducing the amount of political content for a sample of users[9] in an effort to "turn down the temperature and discourage divisive conversations and communities," with the aid of a machine learning algorithm trained to identify political content.

Flashback to November 19, 2016, less than two weeks after Donald Trump's surprising election victory. Mark Zuckerberg posts[10] a message on his Facebook account that begins: "A lot of you have asked what we're doing about misinformation, so I wanted to give an update." After saying that "we know people want accurate information," he goes on to admit that "The problems here are complex, both technically and philosophically" and that Facebook is taking an indirect approach: "We do not want to be arbiters of

[8]"How Facebook Profits from the Insurrection," *Tech Transparency Project*, January 18, 2021: https://www.techtransparencyproject.org/articles/how-facebook-profits-insurrection.

[9]Taylor Telford, "Facebook moves to scale down political content," *Washington Post*, February 10, 2021: https://www.washingtonpost.com/business/2021/02/10/facebook-political-content/.

[10]Mark Zuckerberg, Facebook status update, November 19, 2016: https://www.facebook.com/zuck/posts/10103269806149061.

truth ourselves, but instead rely on our community and trusted third parties." He lists several projects underway in the fight against misinformation, the first and "most important" being stronger detection: "This means better technical systems to detect what people will flag as false before they do it themselves." Zuckerberg is implying here that out of all the approaches to limit the spread of misinformation on his platform, the primary one is developing predictive machine learning algorithms.

A curious subtlety to note here is that what he has in mind is not the supervised learning task of classifying posts as true or false, but instead a somewhat less epistemological classification into posts likely to be flagged by users versus posts unlikely to be flagged. This nuanced distinction matters: whether or not a post in a private group gets flagged as misleading depends a lot on what the focus of the group is. Regardless, the main takeaway here is that in the wake of the 2016 election, Zuckerberg was advocating a technical—and specifically, machine learning—approach to moderating his platform. In numerous public statements before and after this, Zuckerberg has maintained this philosophy that AI will be the company's panacea when it comes to thorny societal problems like the spread of harmful misinformation. But a lot transpired between that presidential election and the next, and one of the main goals of this chapter is to discuss what algorithmic approaches Facebook actually tried and how well they worked and what other algorithmic methods might be possible.

Facebook continually tweaks its platform—especially the enormously influential newsfeed algorithm that decides what posts we are all shown and in what order—usually through controlled experiments in which a random group of users is provided with the modified platform and all the other users serve as the control group. This is a scientific approach to product development inspired by the randomized controlled trial that revolutionized medical research in the 20th century (where the control group is the one given placebos); it has swept through Silicon Valley, where it goes by the name *A/B testing* and is made possible by the tremendous amount of data that tech companies, especially the tech giants, are able to quickly and cheaply obtain.

Just like in medical science, not all experimental treatments are successful. Shortly after Trump's 2016 election victory, Facebook trialed a quite simple algorithmic approach to mitigating fake news that didn't rely on machine learning or professional human moderators. The idea was that when a shared article is fake news, very often astute readers recognize this and post comments calling out the article as fake, sometimes even explaining why it is fake and providing evidence supporting the assertion. But these helpful comments get buried under the deluge of other comments, so Facebook's experiment was just to automatically promote any comment containing the word "fake" to the top of the comments section. This seemed like a fairly straightforward way to help inform readers about potential fake news.

What was the result of this experiment? People in the group where this comment promotion method was implemented were for the most part angry, confused, less able to tell what was fake versus real, and less confident in Facebook's ability to stem the flow of misinformation. Why? Because suddenly the first thing these users saw under nearly every article about politics and politicized topics like the economy and climate change from the most reputable news organizations like *BBC News*, the *New York Times*, the *Economist*, etc. were comments proclaiming the story to be fake. Rather than helping people spot actual fake news, this made real news look fake and left readers in a world where nothing, and hence everything, was believable. Jen Roberts, a freelance PR consultant, captured it well when she said[11] that "to question the veracity of every single story is preposterous" because this "blurs the lines between what is real and what isn't" and turns Facebook newsfeeds into "some awful Orwellian doublethink experiment." Suffice it to say, this was one A/B test that when the experiment concluded Facebook decided to trash the modification and go back to the way things were, imperfect as they were.

Sophisticated algorithms play a role in the spread of fake news on Facebook not just through the ranking of newsfeed content and recommendations for groups to join and pages to follow (the main topics of this chapter), and in political advertising (the topic of the previous chapter), but also in Facebook's search and autocomplete features where the problems are similar to the ones with Google described in Chapter 6. A February 2019 report[12] by the *Guardian* found that when logged in to a new user account, with no friends or other activity, typing "vaccine" into Facebook's search bar produced autocompletes such as "vaccine re-education," "vaccine truth movement," and "vaccine resistance movement" that push people into the world of anti-vax misinformation. Even if the user resisted these autocomplete temptations and simply searched for "vaccination," the top twelve Facebook groups that came up were all anti-vaccination organizations, and eight of the top twelve Facebook pages that came up were anti-vaccination pages suffused with misinformation. Several months earlier, Facebook had launched a policy of deleting misinformation designed to provoke "violence or physical harm," but it stated that anti-vax content does not violate this or any other Facebook policy. However, that changed in February 2021 when Facebook revised its policies and announced[13] that it would start removing essentially all false claims about vaccines.

---

[11] Jane Wakefield, "Facebook's fake news experiment backfires," *BBC News*, November 7, 2017: https://www.bbc.com/news/technology-41900877.

[12] Julia Wong, "How Facebook and YouTube help spread anti-vaxxer propaganda," *Guardian*, February 1, 2019: https://www.theguardian.com/media/2019/feb/01/facebook-youtube-anti-vaccination-misinformation-social-media.

[13] Guy Rosen, "An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19," *Facebook newsroom*, April 16, 2020: https://about.fb.com/news/2020/04/covid-19-misinfo-update/#removing-more-false-claims.

Better late than never: in August 2020, the progressive nonprofit organization Avaaz released a report[14] on the state of global health misinformation, including anti-vax propaganda, on Facebook throughout the preceding year—and the picture it painted was not pretty. The report estimated that content on groups and pages sharing global health misinformation received nearly four billion views during that year. Within the timeframe of the report, views of this misinformative content peaked in April when the coronavirus pandemic was spiraling out of control, despite Facebook's concerted efforts to fight COVID-19 misinformation. The report estimated that during this April peak, content from the top ten most popular health misinformation sites collected four times as many views as did content from the top ten leading authoritative sources such as the WHO and the CDC. It also found that certain "super spreader" pages were responsible for a large fraction of the misinformation, and that many of these super spreaders had origins in the anti-vax movement. One particular article falsely claiming that the American Medical Association was encouraging doctors to overcount COVID-19 deaths received over six million comments or likes and was viewed an estimated one hundred and sixty million times. In response to this report, Facebook said[15] that from April to June it applied fact-checker warning labels to nearly one hundred million COVID-19 misinformation posts and removed seven million others that the company believed risked imminent harm.

Another problem on Facebook involving algorithms and related to misinformation has been the use of bots—fake accounts that can be commanded to behave in certain ways. Bots are often used to artificially seed initial popularity in specified Facebook groups or pages through likes and shares and comments; Facebook's recommendation and newsfeed algorithms detect this high level of engagement and mistake it for authentic activity, causing the algorithms to promote the groups/pages to real users, which in turn drives their actual popularity. As you recall, the *Epoch Times* used this technique in its highly successful "Facebook strategy," even though it manifestly violates the platform's policies against inauthentic account ownership and activity. In September 2020, a data scientist named Sophie Zhang who had been working at Facebook on a team dedicated to catching and blocking bot accounts was fired. On her last day in the office, she posted[16] a lengthy internal memo to the entire company describing the terrifying scope of the platform's

[14]"Facebook's Algorithm: A Major Threat to Public Health," *Avaaz*, August 19, 2020: https://secure.avaaz.org/campaign/en/facebook_threat_health/.

[15]Elizabeth Dwoskin, "Misinformation about the coronavirus is thwarting Facebook's best efforts to catch it," *Washington Post*, August 19, 2020: https://www.washingtonpost.com/technology/2020/08/19/facebook-misinformation-coronavirus-avaaz/.

[16]Craig Silverman, Ryan Mac, and Pranav Dixit, "'I Have Blood on My Hands': A Whistleblower Says Facebook Ignored Global Political Manipulation," *BuzzFeed News*, September 14, 2020: https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo.

bot problem and the corporate leadership's reluctance to heed her repeated warnings to promptly and properly respond to this problem.

Zhang's memo is filled with concrete examples of coordinated bot activities around the world aimed to sway public opinion and influence election outcomes—sometimes with the bots traced to heads of government and political parties. "In the three years I've spent at Facebook, I've found multiple blatant attempts by foreign national governments to abuse our platform on vast scales to mislead their own citizenry, and caused international news on multiple occasions," she wrote. She said that in the 2018 elections in the United States and in Brazil, her team took down over ten million fake reactions and likes of high-profile politicians. She was shocked at how slow Facebook leadership was to respond to many of the bot campaigns her team uncovered—sometimes taking months to act—and also shocked at the unchecked power she and her team had as moderators on the site: "I have personally made decisions that affected national presidents without oversight, and taken action to enforce against so many prominent politicians globally that I've lost count." When Zhang was fired, she was offered a sixty-four-thousand-dollar severance package, but one requirement it included was that she sign a non-disparagement agreement. She turned down this severance package specifically so that she could post her internal company-wide memo—in the hope that it would lead to some real change within the company.

Just a few months before Zhang's departure, the *Wall Street Journal* published[17] a glimpse into the closely guarded backrooms of Facebook's private research on, and response to, some of the negative societal impacts of the company's platform and algorithms over the preceding few years. It turns out that an internal presentation to company leadership in 2016 showed that extremist content—much of it racist, conspiracy-minded, and pro-Russian—was widely found on over a third of large political Facebook groups in Germany, and a relatively small number of very active users were responsible for a large amount of this content. Quite disturbingly, the presentation asserted that "64% of all extremist group joins are due to our recommendation tools," especially the algorithmically driven "Groups You Should Join" and "Discover" suggestions. Quite bluntly, the presentation stated that "Our recommendation systems grow the problem." And Facebook employees told the *Wall Street Journal* that, unsurprisingly, this problem was in no way special to Germany. Two years later, another internal presentation to company leadership stated that "Our algorithms exploit the human brain's attraction to divisiveness" and that if left unchecked they would select "more and more divisive content in an effort to gain user attention & increase time on the platform."

[17]Jeff Horwitz and Deepa Seetharaman, "Facebook Executives Shut Down Efforts to Make the Site Less Divisive," *Wall Street Journal*, May 26, 2020: https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499.

These internal presentations were part of company efforts—some initiating at the very top with Mark Zuckerberg—to understand how Facebook's algorithms influenced user behavior in potentially harmful ways. But the *Wall Street Journal* revealed that the findings from these efforts were to a large extent ignored, and the proposals for addressing the problems were mostly dismissed or greatly reduced in scope. The main team looking at these issues wrote in a mid-2018 internal document that many of its proposed remediations were "antigrowth" and required the company to "take a moral stance." A particularly delicate issue was that the team found that problematic behavior such as fake news, spam, clickbait, and inauthentic users came disproportionately from hyperpartisan users—and that there were larger networks aligned with the far right than the far left. This meant that even politically neutral efforts to reduce problematic behavior would, overall, affect conservative content more than liberal content. Facebook leadership did not want to alienate conservative users with actions that appeared to project a liberal bias, so the company's handling of these matters has been highly constrained.

One specific remediation proposed by this internal Facebook research team was the following. Since Facebook's algorithms were designed to maximize various user engagement metrics (likes, shares, comments, time spent logged on, etc.), users who are very active on the platform have a greater impact on the algorithms than do less active users. The team suggested a "Sparing Sharing" algorithmic adjustment to reduce the spread of content that was disproportionately driven by hyperactive users. The team believed this would help protect Facebook from coordinated manipulation efforts, but the *Wall Street Journal* revealed that senior Facebook executives were apprehensive, claiming this would unfairly hurt the platform's most dedicated users. When the team and the senior executives couldn't agree about this, the debate over it was eventually elevated all the way up to Zuckerberg who evidently said to implement it but only after cutting the proposed mitigation weighting by eighty percent—and he reportedly "signaled he was losing interest in the effort to recalibrate the platform in the name of social good" and asked the team "that they not bring him something like that again."

Why QAnon Is So Tricky

One of the challenges that has emerged with controlling the spread of misinformation on social media is that much of it lately comes by way of diffuse, all-encompassing, and constantly evolving conspiracy theories like the QAnon movement. Rather than being based on a well-defined and falsifiable central tenant, these conspiracy theories weave together many unrelated assertions in a web of deceit so that debunking any particular aspects of the theory—or labeling individual posts on social media as false—does little to slow down the movement as a whole. A further hurdle to moderation is that actions by social media platforms to rein in these movements are frequently

absorbed into the conspiracy theories as coordinated efforts to keep people from learning the truth—which further galvanizes support for the movements. Rather than being static, these theories are more like viruses that constantly adapt and reconfigure themselves in order to persist and spread more rampantly. The supporters of these movements actively look for messaging that allows them to escape policy violations; often while doing so, they land on softer and more moderate ways to frame their ideology—and in the long run, this allows them to reach and convince an even wider audience. It's almost like a bacterial infection that becomes more insidious and difficult to treat after resisting a partial course of antibiotics.

You can see all these factors at play in the 2020 election. An October 2020 report[18] from the Election Integrity Partnership—a self-described "coalition of research entities" supporting "information exchange between the research community, election officials, government agencies, civil society organizations, and social media platforms"—looked into preemptive efforts to delegitimize the 2020 election. It noted that the rumor spreading across social media of a deep state coup attempt to steal the election from Trump was "worth examining […] to understand how it weaves together a wide swath of discrete events into an overarching meta-narrative," and how this "meta-narrative becomes a scaffolding on which any future event can be hung: any new protest, or newly-discovered discarded ballot, is processed as further confirmatory evidence […] that there is a vast conspiracy to steal the election, and that the results will be illegitimate." The report goes on to explain the psychological impact, and social media dynamics, of this framework: "What may previously have been isolated incidents with minimal social media traction may gain significant new weight when they are processed as additional evidence of an underlying conspiracy." This is strikingly similar to what you saw in Chapter 4 with YouTube where an impressionable viewer feels that all signs point to the same hidden truth when the recommendation algorithm naively strings together videos from different users on the same conspiratorial themes.

One of the most bizarre yet influential sprawling meta-narrative conspiracy movements in recent years—even reaching the heights of the US House of Representatives in Marjorie Taylor Greene—is QAnon; let me now take a look at the evolving attempts to control its presence on social media. While some specific QAnon material and accounts had been banned from social media platforms for violating certain company policies, until the summer of 2020 there was nothing prohibited about QAnon itself despite the vast landscape of misinformation it is rooted in and the obvious potential for real-world harm it could lead to. Twitter made the first official move directly

[18]Renée DiResta and Isabella Garcia-Camargo, "Laying the Groundwork: Meta-Narratives and Delegitimization Over Time," *Election Integrity Partnership*, October 19, 2020: https://www.eipartnership.net/rapid-response/election-delegitimization-meta-narratives.

against QAnon when in July 2020 it announced[19] that it would "(1) No longer serve content and accounts associated with QAnon in Trends and recommendations, (2) Work to ensure we're not highlighting this activity in search and conversations, and (3) Block URLs associated with QAnon from being shared on Twitter." Two months later, Twitter said[20] that views of QAnon content on the platform had dropped by more than fifty percent.

One month after Twitter announced its anti-QAnon efforts, Facebook followed suit with an announcement[21] that it would start "taking action against Facebook Pages, Groups and Instagram accounts tied to offline anarchist groups that support violent acts amidst protests, US-based militia organizations and QAnon." The announcement admitted that while content directly advocating violence was already banned on the platform, there had been a growth of movements threatening public safety in slightly more oblique manners such as celebrating violent acts or harboring members who show themselves carrying weapons with the suggestion that they will use them. It went on to explain that Facebook would still "allow people to post content that supports these movements and groups," but it would start to "restrict their ability to organize on our platform." In other words, QAnon content would not be prohibited from individual Facebook users, but QAnon groups and pages on Facebook would face a new collection of restrictions. These restrictions included no longer suggesting QAnon groups and pages as recommendations for users to join/follow; decreasing the newsfeed ranking for posts from QAnon groups and pages; decreasing the search ranking for QAnon groups and pages and removing their names and QAnon hashtags from the autocomplete feature in Facebook's search function; prohibiting paid ads and Facebook's fundraising tools for QAnon; and removing QAnon groups and pages that discuss violence—even if the discussion relies on "veiled language and symbols particular to the movement."

Two months later, Facebook posted an update to this announcement declaring that "we believe these efforts need to be strengthened when addressing QAnon." As of October 6, 2020, Facebook would start removing all groups and pages "representing QAnon, even if they contain no violent content." As justification for ramping up its actions against QAnon, Facebook noted examples such as the movement spreading misinformation about the west coast wildfires that did not fall under the umbrella of inciting or even discussing violence and yet caused real public harm by impeding local officials' ability to

[19]Twitter company tweet, July 21, 2020: https://twitter.com/TwitterSafety/status/1285726277719199746.

[20]Twitter company tweet, September 17, 2020: https://twitter.com/TwitterSupport/status/1306641045413822465.

[21]"An Update to How We Address Movements and Organizations Tied to Violence," *Facebook newsroom*, August 19, 2020: https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/.

fight the fires. This update to the announcement admitted that enforcing this new ban would not be a trivial matter due to how quickly QAnon pivots its messaging and that Facebook expects "renewed attempts to evade our detection, both in behavior and content shared on our platform." A few weeks later, the announcement was updated again to say that now when people search for terms related to QAnon, they would be directed to a counter-terrorism organization. Then in January 2021 the announcement was updated once again, this time to provide current tallies for the QAnon takedown effort: over three thousand Facebook QAnon pages, ten thousand groups, five hundred events, and eighteen thousand profiles had been removed.

What happened in the two months between Facebook's initial announcement that it would reduce QAnon's presence on the platform, primarily through algorithmic adjustments, and its ensuing announcement of an outright ban on the movement? For one thing, the *New York Times* published a scathing article[22] showing that Facebook's first crackdown attempt was, to put it mildly, insufficient. The journalists found that Facebook's recommendation algorithm continued to suggest QAnon groups; one particular QAnon group gained hundreds of followers after the initial crackdown despite openly pushing against important public health advice on the pandemic such as wearing masks; a militia movement on Facebook affiliated with QAnon that was calling for armed conflict on American soil gained thousands of new followers; and hundreds of thousands of Facebook users were pushed toward conspiracy theory groups and pages in the general QAnon orbit under the false pretense of an online campaign against human trafficking. The journalists tracked a hundred QAnon groups on Facebook and calculated that in sum they averaged just over thirteen thousand new followers per week after the crackdown, only a modest decrease from the roughly twenty thousand combined new followers they were averaging prior to it. Meanwhile, these groups became slightly more active, averaging a combined six hundred thousand weekly engagements after the crackdown compared to just over five hundred thousand prior to it.

The journalists found that many QAnon groups/pages were able to avoid Facebook's crackdown simply by changing the letter Q in their name to "cue" or "17" (referencing the fact that Q is the 17th letter of the alphabet), despite Facebook's earlier assertion that it would be on the lookout for veiled and symbolic references. They also found that Facebook groups that had nothing to do with politics, such as parenting and yoga groups, were suddenly suffused with QAnon content—but of a toned-down form, emphasizing the

[22]Sheera Frenkel and Tiffany Hsu, "Facebook Tried to Limit QAnon. It Failed." *New York Times*, September 18, 2020: https://www.nytimes.com/2020/09/18/technology/facebook-tried-to-limit-qanon-it-failed.html.

conspiratorial child trafficking aspects of the movement. Many Facebook groups that were branded as anti-trafficking organizations but really were QAnon propaganda groups actually saw their growth rates spike after the Facebook crackdown. In short, the QAnon movement adapted to Facebook's efforts in both technological and psychological ways to ensure it continued to prosper and spread in the new Facebook environment. With the 2020 election on the horizon and QAnon spreading pro-Trump misinformation about rampant voter fraud, Facebook evidently felt the need to ditch its original strategy and take a much stronger stance against the spiraling and spreading QAnon movement.

Now that the stage has been properly set, I'll first look in more detail at how fake news spreads on social media, then this will inform the ensuing discussion of methods for curbing this spread.

Quantifying the Spread of Fake News

In this section, I look at a handful of academic studies on the intricate propagation dynamics of fake news. Facebook's data is less publicly available, so most of these studies focus on Twitter. That said, one study[23] looked at five hundred fake news sites and ten thousand fake news stories on Facebook and Twitter between January 2015 and July 2018 and found that user interactions with fake news steadily rose on both platforms through the end of 2016 and then sharply declined on Facebook while continuing to rise on Twitter. But I urge the reader caution when interpreting that particular finding: it was only one relatively small study, much of Facebook's fake news problem occurs in private groups that cannot be tracked so easily, and a lot has happened since 2018—especially with the ramp-up of QAnon and the flood of coronavirus misinformation. For instance, data scientists within Facebook found[24] in the months before the 2020 election that seventy of the top one hundred most active groups oriented toward US civics had been flagged for repeated issues such as hate speech, misinformation, bullying, and harassment. One of these top groups claimed it was run by fans of Donald Trump, but it was actually run by "financially motivated Albanians" and generated millions of daily views on fake news and other harmful content.

[23]Hunt Allcott, Matthew Gentzkow, and Chuan Yu, "Trends in the diffusion of misinformation on social media," *Research & Politics* **6** no. 2 (2019): https://journals.sagepub.com/doi/full/10.1177/2053168019848554.

[24]Jeff Horowitz, "Facebook Knew Calls for Violence Plagued 'Groups,' Now Plans Overhaul," January 31, 2021: https://www.wsj.com/articles/facebook-knew-calls-for-violence-plagued-groups-now-plans-overhaul-11612131374.

Twitter and the 2016 Election

In January 2019, a research paper was published[25] that took a deep data dive into the dynamics of fake news on Twitter in the 2016 election. This is a convenient place to start our discussion of how fake news spreads on social media. The researchers collected every tweet they could find concerning Donald Trump and Hillary Clinton in the five months leading up to the election, which ended up being a whopping one hundred and seventy million tweets sent by eleven million users. Of these, thirty million tweets by just over two million users contained links to news articles or organizations. Using a characterization of news outlets by communications scholars into those that publish fake or unsubstantiated conspiratorial news versus those that are more traditional and fact based, the researchers labeled each of these thirty million tweets as factual or fake. Important to note here is that they didn't try to judge the accuracy of each individual news story shared on Twitter, which would be extremely difficult, they only assessed whether the news outlet linked to in each tweet was typically factual or typically fake. The researchers also recorded a political orientation and extremity score for each news outlet that again came from a consensus of communications scholars.

They found that among the thirty million tweets with news links, ten percent were to fake news organizations and another fifteen percent were to extremely biased news. Nearly one in five of the fake news links were tweeted from non-official Twitter apps (and among the tweets that came from non-official Twitter apps, links to news were four times as likely to be to fake news than to traditional news); while there are some legitimate and semi-legitimate uses of non-official apps, this was interpreted as evidence of substantial bot activity in the spread of fake news during the lead-up to the election. By looking at retweets, the researchers were able to study the network flow of information and identify the key influencers. They found that the top spreaders of traditional news were journalists and public figures with verified Twitter accounts, whereas a large number of the top spreaders of fake and extremely biased news were unknown users and accounts that were subsequently deleted. The tweeting/retweeting network was more connected and homogeneous for fake news than it was for traditional news; the fake news network was also more tightly entwined with the network of right-wing news tweets than it was for center and left-wing news.

Perhaps most remarkably, the researchers uncovered two crucial differences between Clinton supporters and Trump supporters in terms of their interactions with news on Twitter. First, for center and left-leaning news, there was a top-down effect in which the top spreaders—which, as you recall, were journalists and public figures—strongly influenced the Twitter activity of

[25]Alexandre Bovet and Hernán Makse, "Influence of fake news in Twitter during the 2016 US presidential election," *Nature Communications* **10** no. 7 (2019): https://www.nature.com/articles/s41467-018-07761-2.

Clinton supporters, whereas for Trump supporters the direction of influence was reversed: in a bottom-up manner, it was the activity of Trump supporters that influenced top spreaders of fake news, meaning the top fake news spreaders were primarily conduits and amplifiers for the stories circulating among the masses. Second, it was found that Clinton supporters mostly interacted with center and left-wing news sources, whereas Trump supporters were more inclined to interact with news sources across the gamut—from right-wing to left-wing and from factual to fake.

Another pair of researchers looked more directly[26] into the scope of bot activity on Twitter in the weeks leading up to the 2016 election. By using political hashtag and keyword searches, they collected twenty million tweets from nearly three million users between September 16 and October 21. They then applied a machine learning algorithm to classify these tweets as human versus bot, and the algorithm reported that nearly twenty percent of the tweets were likely from bots. Keep in mind, however, that determining what social media activity is bot-driven has proven challenging, and any attempt to do so will surely have a margin of error. The algorithm used in this study was a supervised learning method relying on one thousand predictors "spanning content and network structure, temporal activity, user profile data, and sentiment analysis" that was trained on known bot and human activity on Twitter. In previous tests, it scored an accuracy rate of ninety-five percent, but there's no guarantee that this high accuracy rate would be sustained as new bot algorithms, tricks, and behaviors develop.

After any supervised learning algorithm has been trained, it is possible to inspect it to determine which predictors are most influential (the technical term for this is *feature importance*). For the particular bot detection algorithm used in this study, it turned out the strongest indicators of bot accounts were:

- That the user profile looked like the Twitter default, rather than being customized with individual information; the username had signs of randomness in its creation; and the account was created recently.
- The absence of geolocation metadata (which for humans is recorded when tweeting from mobile devices).
- Various activity statistics: Bots often post large numbers of tweets in short spans of time that would be impossible or nearly impossible for humans to accomplish, and unsurprisingly their ratio of retweets to original tweets tends to be much higher than for humans and so is their ratio of accounts followed to followers.

[26]Alessandro Bessi and Emilio Ferrara, "Social bots distort the 2016 U.S. Presidential election online discussion," *First Monday* **21** no. 11 (2016): https://firstmonday.org/ojs/index.php/fm/article/view/7090/5653.

When comparing the four million tweets that their algorithm labeled as bot activity with the sixteen million tweets labeled as human activity, the researchers found a couple intriguing although somewhat predictable contrasts. First, the volume of human tweets tended to respond to political events—for instance, there were large numbers of tweets immediately following the presidential debates—whereas the volume for bots was less closely tied to political events. Second, while bots generated replies at a lower level than human users overall, they got retweeted at the same rate as humans. A different research team studied[27] bots on Twitter during a ten-month period beginning six months before the 2016 election. They found that bots were particularly active and successful in the very early stages of virality for misinformation—often sharing fake news stories within seconds of them being published—and in this way would broadcast fake news until it caught on and spread organically through humans on Twitter. They also found that bots frequently targeted influential human users with high volumes of replies and mentions that tended to eventually draw in the human users.

Yet another team of researchers, this one based at Oxford University, produced a paper[28] that looked at the geographic distribution in the United States of low-quality political information on Twitter in a ten-day period around the 2016 election. In this study, a link to a news story was deemed low quality if the news organization was among a list of certain Russian, WikiLeaks, and junk/fake news sites, and the location of a tweet was obtained from information on the user's account profile. The location was not available for all users, and even among the users for which it was available, the location listed in a profile is not necessarily accurate; but in the aggregate, this still gives a reasonable sense of geographic distribution. The researchers found that, when weighted by the number of tweets coming from each state, the influential swing states in the election on average saw a higher fraction of tweets with low-quality news links than did the uncontested states. The margin here was relatively small, but this was an election where small margins could have had large impacts due to tightness of the race and the structure of the electoral college. One of the authors of the study said[29] that "We know the Russians have literally invested in social media. Swing states would be the ones you would want to target."

---

[27]Shao et al., "The spread of low-credibility content by social bots," *Nature Communications* **9** no. 4787 (2018): https://www.nature.com/articles/s41467-018-06930-7.

[28]Philip Howard et al., "Social Media, News and Political Information during the US Election: Was Polarizing Content Concentrated in Swing States?" Oxford internal publication, September 28, 2017: https://comprop.oii.ox.ac.uk/research/posts/social-media-news-and-political-information-during-the-us-election-was-polarizing-content-concentrated-in-swing-states/.

[29]Denise Clifton, "Fake News on Twitter Flooded Swing States That Helped Trump Win," *Mother Jones*, September 28, 2017: https://www.motherjones.com/politics/2017/09/fake-news-including-from-russian-sources-saturated-battleground-states-trump-barely-won/.

The studies discussed so far provide an informative view of the propagation and network structure of fake news on Twitter during the 2016 election, but what's missing is how this translates to the individual-level experiences and activities of registered voters on Twitter. Fortunately, a paper[30] was published in 2019 in the top academic journal *Science* that attempts to fill in this crucial missing piece of the story. The researchers linked a sample of public voter registration records to sixteen thousand Twitter accounts and collected the tweets from these users—let's call them the "voters"—between August and December 2016. They collected lists of all the users following and followed by the voters, and by sampling the tweets posted by the latter—called "exposures" since these are the tweets the voters were potentially exposed to in their newsfeeds—the researchers were able to estimate the composition of the voters' newsfeeds. They limited their investigation to exposures containing links to political content outside of Twitter, and they used these links to provide a discrete estimate of the left-right ideology of each voter. Here's what they found.

First, the statistics in terms of exposures and shares. Five percent of all political link exposures were to fake news, and more than half of these fake news exposures came from the same seven fake news sources. But these fake news exposures were not distributed evenly among the voters—quite the opposite, in fact: the newsfeeds of just one percent of the voters accounted for eighty percent of the fake news exposures. Posting of fake news was even more uneven: eighty percent of the fake news links shared by the voters came from just one-tenth of a percent of the voters. In other words, while there was a lot of fake news being seen and shared by the voters, the seeing of it was quite concentrated into the newsfeeds of certain voters (called "superconsumers of fake news"), and the sharing of it was even more concentrated into an even smaller number of the voters (called "supersharers of fake news"). The supersharers were extremely active: on average, a typical supersharer of fake news tweeted seventy times per day, while overall a typical voter tweeted only once every ten days. A typical superconsumer of fake news had almost forty-seven hundred exposures to political links per day, while overall a typical voter had fifty. The supersharers of fake news were much more active than even the highly prolific sharers on Twitter in general, and the researchers suspect that many of them were so-called *cyborg* accounts, which means a partially automated account (i.e., a hybrid of a bot and a human).

Next, the statistics in terms of the voters. The voters averaged two hundred exposures to fake news links during the last month of the election. On average, only one percent of the political links the voters saw were to fake

[30]Nir Grinberg et al., "Fake news on Twitter during the 2016 U.S. presidential election," *Science* **363** no. 6425 (2019): https://science.sciencemag.org/content/363/6425/374.

news. The voters with higher concentrations of fake news in their newsfeed were far more likely to be conservative than liberal: people seeing at least five percent fake political links made up only two and a half percent of the liberal voters but over sixteen percent of the conservative voters. The older a voter was, the higher was the proportion of fake news they saw in their newsfeed. Voters in swing states had slightly higher proportions of fake news (corroborating studies discussed earlier), as did men and whites, but the size of these effects was quite small. Among the voters classified politically as extreme left, just under five percent ever shared a fake news link; the rate for left and center users was also just under five percent, whereas for politically right users it jumped up to just under twelve percent, and more than one in five extreme right users shared a fake news link during the five months of the study period.

Lies Spread Faster and Deeper Than Truth

The studies discussed above investigated various data sets of tweets by classifying links to news organizations as either traditional media outlets or fake news outlets. This organization-based approach provides a wealth of valuable information, but it is by design blind to the actual stories linked to in these tweets. Some outlets publish both real and fake news, which renders this approach problematic. Moreover, in some sense what matters most from a societal perspective are the particular stories that spread virally across social media—a spread that usually involves hopping across different news organizations (recall the vertical and horizontal propagation you saw in Chapter 1)—and the people these stories reach along the way. A study[31] widely considered the pinnacle of fake news research, published by a trio of academics from MIT in 2018 in *Science*, took a drastically new approach by tracking the trajectory of individual stories—both true and false ones—across Twitter. Their findings are fascinating and transformative; in short, they showed that lies spread faster and deeper than the truth. To understand the meaning, impact, and limitations of this study, it is worth really getting into the weeds here.

The researchers started by collating a list of essentially all the stories/assertions/claims that have appeared on at least one of six popular and well-respected fact-checking websites.[32] This yielded a collection of "rumors," each labeled on a true-to-false scale by combining the scores from the individual fact-checking organizations (all of which rely on expert human opinion rather

[31]Soroush Vosoughi, Deb Roy, and Sinan Aral, "The spread of true and false news online," *Science* **359** no. 6380 (2018): https://science.sciencemag.org/content/359/6380/1146.

[32]These were snopes.com, politifact.com, factcheck.org, truthorfiction.com, hoax-slayer.com, and urbanlegends.about.com.

than algorithmic analysis of the type discussed in the next chapter of this book). The researchers were granted remarkable and nearly unprecedented access to the full historical archive of all tweets ever posted on Twitter, going back to the very first tweet in 2006. They searched through all the English-language tweets, up through December 2016, and extracted all of those with a link to an item in one of the six fact-checking websites. Among these extracted tweets, they further extracted the ones that were in the form of a reply to a top-level tweet, and then they used a combination of machine learning and manual methods to check whether the headline in the fact-checking link closely matched the content of the top-level tweet being replied to. For each of these top-level tweets with a relevant fact-checking reply, they extracted the full network of replies and retweets, which they call a "rumor cascade" since the top-level tweet concerns one of the rumors in the collated list and the network of replies and retweets conveys how this particular instance of the rumor cascaded through Twitter.

Next, machine learning was used to process the text, links, and images in the top-level tweets in these rumor cascades to group them according to the rumor they concerned. The researchers also collected additional rumor cascades by searching Twitter for top-level tweets that contained similar text/links/images to any of the top-level tweets in the rumor cascades and extracting their network of replies and retweets even if none of the replies contained a fact-checking link. In the end, they collected over one hundred and twenty-five thousand rumor cascades that were mapped onto about twenty-five hundred distinct rumors. The rumors represent any piece of contested information that was fact-checked by at least one of the six organizations, and the rumor cascades represent their various trajectories on Twitter. Of the rumors, seventy percent were false, twenty percent were true, and the remaining ten percent were mixed; of the rumor cascades, sixty-five percent concerned a false rumor, twenty percent concerned a true rumor, and the remaining fifteen percent concerned a mixed one.

After analyzing these cascades with tools from network science, the main findings in this study were the following: false rumors spread faster, farther, deeper, and more broadly than true rumors; this spread was more pronounced for false political rumors than for false rumors about topics like science, finance, and disasters; false rumors tended to be more novel and to involve fear, disgust, and surprise, whereas true rumors tended to be more similar to other content and to involve anticipation, sadness, joy, and trust; finally, bots appear to have accelerated the spread of true and false rumors at roughly equal rates. Putting the first and last findings together implies that it was humans, not bots, that were more likely to spread false rumors than true rumors. That said, it is important to keep in mind that this study concerns aggregate behavior over a ten-year span—it is quite plausible that in certain specific instances bots played a much larger role.

Here are some more details on these findings. On average, false rumors reached fifteen hundred people six times faster than true rumors did. Even when controlling for various differences between the users that originated rumor cascades, such as their number of followers and whether the account was verified by Twitter, false rumors were seventy percent more likely to get retweeted than true rumors. The largest rumor cascades reached around fifty thousand users when the rumor was false but only around two thousand users when it was true. There are two very different ways that information can spread and reach a large number of users on Twitter: a prominent influencer could tweet a story that many followers will directly retweet, or a less prominent user could tweet a story that gets retweeted by a small number of followers who then get it retweeted by some of their followers, etc. Even if a story reaches the same number of retweets in these two scenarios, the first is considered a shallow spread and the second a deep spread since it penetrates more deeply into the social network. It was found in this study that not only did false rumors ultimately reach larger audiences, but they did so with much greater depth: true rumors seldom chained together more than ten layers of retweets, whereas the most viral false rumors reached twenty layers of retweets—and they did so ten times as quickly as the true rumors reached their ten.

The main caveat—I'm tempted to call it a weakness or even a serious flaw—with this study is that it is not really about fake versus real news: it is about contested information. The rumors studied here are in no way a representative sample of all news since they were drawn from items that appeared on fact-checking websites. Most news stories are not even remotely controversial and so will have no presence on fact-checking sites. Go to the homepage of an online newspaper, even a highly partisan one, and ask yourself how many of the stories there are likely to appear on a fact-checking site—my guess is very few. Put another way, it could actually be the case that valid news spreads more on Twitter than fake news; we really don't know. In fact, for many years, the story on Twitter that had the record for receiving the most retweets was the (undeniably true and therefore not fact-checked) news of President Obama's victorious reelection in 2012. What this *Science* paper reveals—and it is unequivocally a startling and influential revelation—is that among the news events/stories where there is a sufficient degree of dispute to warrant a fact-check, the events/stories that fail this fact-check spread more in every conceivable metric than do the ones that pass it.

The 2020 Election

We don't yet have as detailed an understanding of the role of social media in the 2020 election as we do with the 2016 election (note that some of the most important studies of the latter were only published two or three years after the election), but we do have some preliminary analysis of the situation.

During the week of the 2020 election, there were three and a half million engagements (likes, shares, comments) on public Facebook posts referencing the phrase "Stop the Steal," a slogan for the pro-Trump false claims of voter fraud.[33] Six percent of these engagements occurred on the pages of four prominent influencers—Eric Trump and three conservative social media personalities—but the biggest super spreader here was Donald Trump: the twenty most engaged Facebook posts containing the word "election" were all from him, and they were all found to be false or misleading. In a four-week period surrounding November 3, President Trump and the other top twenty-five spreaders of voter fraud misinformation generated more than a quarter of the engagements on public voter fraud misinformation on Facebook that was indexed by an Avaaz investigation. Concerning the false claim that Dominion voting software deleted votes for Trump, more than a tenth of all engagements came from just seven posts. Many of the top spreaders of pro-Trump election misinformation on Facebook were also top spreaders of this same misinformation on Twitter—most notably, of course, President Trump himself.

The most detailed look so far at fake news on social media in the 2020 election is a lengthy report published[34] in March 2021 by the Election Integrity Partnership that carefully tracked over six hundred pieces of election misinformation. One challenge the authors noted is that the social media app Parler is believed to have harbored a lot of election misinformation, but it does not make its data readily available, and so it is challenging for researchers to study content on Parler; similarly, Facebook's private groups were hotbeds for misinformation, and the limited access they grant renders them difficult to study. Overall, the authors of this report found that misinformation in the 2020 election built up over a long period of time on all the social media platforms—despite efforts by the big ones to limit it—and very much exhibited the evolving meta-narrative structure discussed earlier in the context of QAnon. There were so many different forms and instances of false information about the election all pointing to a general—yet incorrect—feeling that the election would be, then was, stolen from President Trump that debunking any particular claim did little to slow down the movement, and sometimes doing so even brought more attention to the claim and further generated conspiratorial mistrust of the social media platform and/or fact-checking organization involved.

To see how this unfolded, it helps to look at one of the specific items of misinformation evolution tracked in this Election Integrity Partnership report.

---

[33]Sheera Frenkel, "How Misinformation 'Superspreaders' Seed False Election Theories," *New York Times*, November 23, 2020: https://www.nytimes.com/2020/11/23/technology/election-misinformation-facebook-twitter.html.

[34]"The Long Fuse: Misinformation and the 2020 Election," *Election Integrity Partnership*, March 3, 2021: https://purl.stanford.edu/tr171zs0069.

Recall that one part of the stolen election narrative was the false claim that Dominion voting machines deleted votes for Trump. Here's what the report found:

> *"Dominion narratives […] began with claims of poll glitches in online conversations on websites and Twitter, then spread through YouTube videos and the use of hashtags […] on Twitter and other platforms, such as Parler and Reddit. From there, high-profile accounts drew further attention to the incidents, as did hyperpartisan news websites like The Gateway Pundit, which used Twitter to promote its article discussing the incident. This collective Dominion narrative spread has since grown, having been subsequently promoted by the Proud Boys, The Western Journal, and Mike Huckabee across a number of platforms, including Facebook, Twitter, Instagram, Telegram, Parler, and Gab."*

This narrative trajectory gives a decent qualitative sense of how so much of the 2020 election fake news started and then snowballed, often passing through multiple platforms along the way, even though it lacks the quantitative details that the large-scale data-driven studies of the 2016 election offered.

Incidentally, while the spread of the Dominion voting machine false narrative is typical of many recent false narratives, what happened afterward is much less typical—and could potentially have far-reaching ramifications for the spread of disinformation in general. In December 2020, an employee of Dominion Voting Systems filed a defamation lawsuit against the Trump campaign, several Trump-affiliated lawyers (including Rudy Giuliani and Sidney Powell), and several conservative news organizations (including *Newsmax*, *One America News*, and *Gateway Pundit*) for their role in the false conspiracy theory concerning Dominion. This lawsuit, together with the threat of a related one by the voting company Smartmatic, led *Newsmax* to quickly retract its earlier claims of voting machine election fraud—and on April 30, 2021, *Newsmax* settled the Dominion case for an undisclosed amount and provided an official apology for broadcasting the story without any evidence that it was true (the rest of the Dominion case is still ongoing).[35] Meanwhile, Dominion and Smartmatic are both suing *Fox News*, seeking over four billion dollars in damages.[36] Finally, it seems, there is some accountability for broadcasting fake news—at least in this specific situation, though this may indeed set a new precedent. But it is important to note that none of the social media companies

[35]Joe Walsh, "Newsmax Apologizes To Dominion Exec As It Settles Lawsuit Over False Voter Fraud Claims," *Forbes*, April 30, 2021: https://www.forbes.com/sites/joewalsh/2021/04/30/newsmax-apologizes-to-dominion-exec-as-it-settles-lawsuit-over-false-voter-fraud-claims/.

[36]Michael Grynbaum and Jonah Bromwich, "Fox News Faces Second Defamation Suit Over Election Coverage," *New York Times*, March 26, 2021: https://www.nytimes.com/2021/03/26/business/media/fox-news-defamation-suit-dominion.html.

involved in the spread of this dangerous conspiracy theory have suffered any financial or legal consequences; the reason for this is a legal protection called Section 230 that I'll come to later in this chapter.

Now it is time to turn from studying the spread of fake news to the design and implementation of algorithms to help curb it. I'll start first with what Facebook and Twitter have actually done—based on the limited information they have made publicly available—then in the subsequent section, I'll turn to algorithmic approaches that have been suggested by various researchers outside of the social media companies.

How Algorithms Have Helped

Some of the technical solutions social media companies have implemented to reduce the spread of fake news and harmful content are quite straightforward. Facebook owns the messaging service WhatsApp which has proven yet another way that misinformation spreads virally, especially in countries outside the United States; you saw this with the malicious deepfake of a journalist in India in Chapter 3, and while I didn't discuss it in Chapter 4, WhatsApp was also a significant vector for far-right and conspiracy theory content in Brazil's 2018 election. To help counteract this, in 2018 Facebook reduced the number of people a WhatsApp message could be forwarded to from two hundred fifty to twenty, and this number has since been reduced to five; in anticipation of the 2020 election, it also reduced the forwarding limit on Facebook's Messenger from one hundred fifty to five.[37]

Other technical solutions are on the surface relatively simple but difficult to implement. In January 2020, Facebook announced[38] a seemingly straightforward move: it would start banning deepfakes on its platform. More precisely, this was a ban on "misleading manipulated media," meaning any video that "has been edited or synthesized—beyond adjustments for clarity or quality—in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say" and that also "is the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic." This lengthy description really is intended to spell out that the ban is aimed at deepfakes that are used in a deceptive manner, but you saw in Chapter 3 that detection of deepfakes—whether algorithmic or manual—is, and likely always will be, quite challenging, so this ban will not

[37]Mike Isaac, "Facebook Moves to Limit Election Chaos in November," *New York Times*, September 3, 2020: https://www.nytimes.com/2020/09/03/technology/facebook-election-chaos-november.html.

[38]Monika Bickert, "Enforcing Against Manipulated Media," *Facebook newsroom*, January 6, 2020: https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/.

be easy to implement in practice. Also, note that this prohibition doesn't cover shallowfakes of any kind, and for deepfakes it only covers words, not actions—so fake videos of public figures engaged in adulterous activity, for instance, are still allowed. Here's another example of a policy decision that was far simpler to state than to implement: in October 2019, Twitter announced that it would start banning political advertising. This requires deciding exactly which ads count as political, which is no simple matter.

Many of the technical approaches that have been implemented, however, are quite sophisticated and rely extensively on machine learning. This is particularly true of Facebook, and it's the topic I turn to next.

Facebook's Machine Learning Moderation

In November 2019, Facebook's Chief Technology Officer wrote a blog post[39] announcing and contextualizing the company's latest "Community Standards Enforcement Report." Conveniently, this blog post highlighted and outlined some of the company's recent developments in AI for content moderation. It stated that the biggest recent improvements were driven by a partial transition from supervised learning, where data needs to be manually labeled in what is typically a time-consuming and laborious process, to self-supervised learning where the required labels are automatically drawn directly from the data. You saw an example of self-supervised learning in Chapter 6 with BERT: the training data there is text with some words randomly hidden that the algorithm learns to predict, and the data labels for these hidden words are simply the words themselves. Self-supervision allows for much larger data sets to be processed so that one can train much larger, hence more accurate, predictive algorithms than one could when using traditional supervised learning methods. Since Facebook has access to truly enormous data sets, and it has the computational resources to train and support very large algorithms, this certainly is a logical path for the company to pursue. There are still plenty of situations where one really needs manually curated and labeled data, but often these situations can be combined with self-supervision in a hybrid approach that is, in a sense, the best of both worlds.

One of the specific instances mentioned in the blog post of self-supervised learning used to great effect is a language processing system called RoBERTa that was Facebook's answer to Google's BERT. Recall from Chapter 6 that BERT is very similar to GPT-3 except that BERT's goal is to produce vector embeddings of words since this opens the door to a wide range of machine learning operations on text. Like Google's use of BERT, Facebook uses RoBERTa to power its search algorithm—but more than that, Facebook also

[39]Mike Schroepfer, "Community Standards report," *Facebook blog*, November 13, 2019: https://ai.facebook.com/blog/community-standards-report/.

uses RoBERTa to help with identifying things like hate speech. This is a good example of a hybrid situation: Facebook trains a traditional supervised learning classifier on a collection of posts that have been manually labeled as hateful or not hateful, but rather than having the algorithm work directly with the text in these posts, it instead works with the text after the self-supervised RoBERTa has transformed the words in these posts into numerical vectors. Wherever there's a situation in which the tech giants need a computer to be able to deal with the meaning of language, especially potentially subtle uses of language, you can assume that they now use a massive self-supervised language model like BERT or RoBERTa and that doing so has drastically improved performance over past approaches. That said, hate speech is generally much more self-apparent, even to a computer, than something like misinformation that depends heavily on context and background knowledge.

Another important development highlighted in the blog post was a new "holistic" approach to content moderation on Facebook. Previously, Facebook's detection systems looked separately at the words in a post, the images in a post, and the comments on a post—and the systems also considered each violation activity separately. For example, before the holistic approach, one classifier would look for nudity in a posted photo, a separate classifier would look for violence in the photo, a separate classifier would look for hate speech in the photo's caption, etc. Facebook replaced this with a pre-trained machine learning algorithm called *Whole Post Integrity Embeddings* (*WPIE*) that converts each post (photos, text, comments, and even user data) into a vector—a sequence of numbers—so that any type of violation classifier need only work with these vectors rather than breaking the post data into disjoint pieces. In other words, this is like RoBERTa, but instead of just text, it reads in entire Facebook posts with comments and images and awareness of the users responsible for the posts and comments.

In general, the way algorithmic detection for policy violations works is the algorithm assigns a score to each piece of content, and if that score is above a certain threshold, then the content is automatically removed; if the score is below a certain threshold, then the content is left alone; and if the score is between these two thresholds, then the content is flagged for human moderators to look at and evaluate. Facebook says the holistic WPIE framework helped the company remove over four million pieces of drug sale content in the third quarter of 2019, more than ninety-seven percent of which was scored above the automatic removal threshold; this was a substantial increase over the pre-WPIE first quarter when the number removed was less than a million, and less than eighty-five percent of these were scored over the automatic removal threshold. However, when the coronavirus pandemic came around a few months later and detecting health misinformation became an urgent challenge, Facebook found itself relying largely on human moderators and external fact-checking organizations—although, as I'll discuss later in this

chapter and more extensively in the next chapter, machine learning still plays multiple important roles in that process.

As you recall from earlier in this chapter, bot activity on social media has long been a significant problem. As you also recall from this chapter, bot behavior tends to have quantitatively distinct patterns from human behavior—such as bursts of activity that far outpace what a human could achieve. When it comes to algorithmic detection, this powerful efficiency of bots is their own undoing: not only can detection algorithms look for direct signs of bots in metadata, but the algorithms can also look for behavioral differences. It is relatively easy to program a bot to share articles and even to write simple human-sounding posts and comments, but to fly under the radar, one needs to ensure that the bot does this at the approximate frequency and scope of a human user. And in the case of Facebook, there's another important factor involved: the friendship network. What has proven most challenging when it comes to creating bots that simulate human behavior is developing a Facebook account with a realistic-looking network of friends[40]—and doing this in large enough numbers for an army of bots to have a significant impact.

In November 2020, Facebook released[41] some details on its latest deep learning bot detection algorithm. It relies on over twenty thousand predictor variables that look not just at the user in question but also at all users in that user's network of friends. The predictors include demographic information such as the distribution of ages and gender in the friend network, information on the connectivity properties of the friend network, and many other pieces of information that Facebook did not disclose. The algorithm is trained in a two-tier process: first, it is trained on a large data set that has been labeled automatically, to get a coarse understanding of the task, then it undergoes fine-tuning training on a small data set that has been labeled manually so the algorithm can learn more nuanced distinctions. Facebook estimated[42] in the fourth quarter of 2020 that approximately one in twenty of its active users were fake accounts. Throughout that year, it used this new deep learning system to remove over five billion accounts that were believed to be fake and actively engaging in abusive behavior—and that number does not include the millions of blocked attempts to create fake accounts each day.

---

[40]With Twitter, an analogous challenge is developing a realistic-looking network of followers and accounts followed, although organic Twitter networks seem more varied—and therefore easier to spoof—than organic Facebook networks, perhaps because Facebook friendships tend to reflect real-life relationships whereas Twitter relationships do not.

[41]Teng Xu et al., "Deep Entity Classification: Abusive Account Detection for Online Social Networks," *Facebook research*, November 11, 2020: https://research.fb.com/publications/deep-entity-classification-abusive-account-detection-for-online-social-networks/.

[42]"Community Standards Enforcement Report," *Facebook*, February 2021: https://transparency.facebook.com/community-standards-enforcement#fake-accounts.

In November 2020, Facebook also announced[43] a new machine learning approach for ordering the queue of posts that are flagged for review by human moderators. Previously, posts flagged for human review for potentially violating Facebook's policies (which includes both posts flagged by users and posts that triggered the algorithmic detection system but didn't score above the threshold for automatic removal) were reviewed by human moderators mostly in the order in which they were flagged. The new approach uses machine learning to determine the priority of posts in the queue so that the most urgent and damaging ones are addressed first. The main factors the algorithm considers are virality, severity, and likelihood of violating a policy—but the ways these are measured and weighed against each other were not revealed, outside of saying that real-world harm is considered the most important.

Twitter's Bot Detection

In a September 2017 blog post,[44] Twitter released some details on its efforts following the 2016 election to rein in bot activity on the platform. At the time, the company's automated systems were catching around three million suspicious accounts per week—twice the rate from a year earlier in the months leading up to the election. Twitter also uses machine learning to identify suspicious login attempts—blocking half a million per day when the blog post was written—by looking for signs that the login is scripted or automated, though no indication of what predictors this actually involves was given. Additionally, Twitter uses clustering algorithms to look for large groups of accounts that were created and/or controlled by a single entity, but we don't know what variables are involved in this clustering process.

The 2020 Election and Its Aftermath

Remember from Chapter 6 that Google's main approach to reducing the impact of misinformation is to elevate quality journalism in its search rankings—and it does this by assigning an algorithmically determined score to sources that is based on data-driven measures like PageRank as well as human-driven quality assessment measures. Facebook similarly assigns a score to news publishers, called a *news ecosystem quality* (*NEQ*) score, that is relevant

[43]James Vincent, "Facebook is now using AI to sort content for quicker moderation," *The Verge*, November 13, 2020: https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation.

[44]"Update: Russian interference in the 2016 US presidential election," *Twitter blog*, September 28, 2017: https://blog.twitter.com/official/en_us/topics/company/2017/Update-Russian-Interference-in-2016--Election-Bots-and-Misinformation.html.

when posts contain links to news articles. Ordinarily, NEQ scores play only a small role in the newsfeed ranking algorithm, but several days after the 2020 election, Mark Zuckerberg acceded to the demands of a team of employees to significantly increase the algorithm's weighting of NEQ scores in order to reduce the spread of dangerous misinformation.[45] While this change was a temporary measure, some employees later requested that it become permanent; they were rebuffed by senior leadership, but it was decided that the impact of this temporary increase to NEQ scores would be studied and could inform future decisions.

What Else?

What has been described so far in this section on algorithmic moderation surely only scratches the surface of the myriad and multifaceted efforts behind closed doors at Facebook and Twitter to reduce fake news through the development and adjustment of algorithms. To get a sense of what else is possible, not just on these two platforms but in any social media setting, I turn next to approaches that have been developed by academic researchers outside of the tech giants. Some of these approaches might, in one form or another, already be absorbed into Facebook's or Twitter's internal approach—we just don't know, due to the veil of secrecy the companies keep around their methods. Others almost certainly are not, because they are either more primitive than what the companies actually use or they involve too massive of a design overhaul and/or they impinge too strongly upon the companies' bottom line: user engagement. Moreover, at the end of the day, social media companies will only use algorithms to remove misinformation in situations where they have an explicit policy against it. Nonetheless, it is helpful to look carefully at the methods in the next section. As opposed to the internal Facebook and Twitter approaches sketched above, the following academic approaches don't hide any technical details—so in addition to showing what's possible, they give a more concrete sense of how algorithmic moderation really works under the hood.

How Algorithms Could Help

In this section, I'll start first with fake news mitigation methods based on broader structural ideas for reengineering social media networks; then I'll turn to more down-to-earth methods that rely on the way fake news spreads through social networks as they currently operate.

[45]Kevin Roose, Mike Isaac, and Sheera Frenkel, "Facebook Struggles to Balance Civility and Growth," *New York Times*, November 24, 2020: https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html.

Structural Approaches

One idea proposed[46] in August 2020 was to introduce "circuit breakers" into social media platforms. Similar to how the New York Stock Exchange automatically closes when trading reaches dangerously volatile levels, social media platforms could pause or at least significantly slow the viral spread of content at key moments in order to provide a cool-off period and avail fact-checkers of more time to do their job. This is probably not very far from what Facebook actually did when it enacted emergency measures in the days after the 2020 election, but we don't know for sure because, as I mentioned earlier, the company has provided very few public indications of what those emergency measures really were. Friction could be added to platforms overall in moments of crisis, or it could be applied to individual pieces of controversial content to slow their virality without entirely censoring them. Facebook did say that it is internally testing this idea of deliberately slowing the spread of viral posts. The idea of an overall circuit breaker or individual virality speed bumps to help fight misinformation has even been likened to "flattening the curve," the mantra for lockdowns during the first months of the COVID-19 pandemic.[47]

A frequently discussed issue with social media is "filter bubbles," the idea that people get funneled into homogeneous networks of like-minded users sharing content that tends to reinforce preexisting viewpoints; this can create a more divisive, polarized society, and in extreme cases it may even lead to people living in different perceived realities from each other. Some researchers have proposed algorithmic methods for bursting, or at least mitigating, these social media filter bubbles. One particular approach[48] is to first assign a vector of numbers to each user and to each social media post (or at least each news link shared) that provides a quantitative measure of various dimensions such as political alignment. A filter bubble is reflected by a collection of users whose vectors are similar to each other and who tend to post content with similar vectors as well. With this setup, researchers designed an algorithm to prioritize diverse content to a select group of users who are deemed likely to share it and help it spread across the social network, penetrating filter bubbles as it goes. The researchers' particular strategy for selecting the users to seed this spread was found to be three times more effective at increasing the overall diversity of newsfeeds in the network than a simpler approach of just

[46]Shannon Bond, "Can Circuit Breakers Stop Viral Rumors On Facebook, Twitter?" *NPR*, September 22, 2020: https://www.npr.org/2020/09/22/915676948/can-circuit-breakers-stop-viral-rumors-on-facebook-twitter.

[47]Ellen Goodman and Karen Kornbluh, "Social Media Platforms Need to Flatten the Curve of Dangerous Misinformation," *Slate*, August 21, 2020: https://slate.com/technology/2020/08/facebook-twitter-youtube-misinformation-virality-speed-bump.html.

[48]Michelle Hampson, "Smart Algorithm Bursts Social Networks' 'Filter Bubbles'," *IEEE Spectrum*, January 21, 2021: https://spectrum.ieee.org/tech-talk/computing/networks/finally-a-means-for-bursting-social-media-bubbles.

targeting diverse content to the most well-connected users. The researchers do admit that this diversity-oriented newsfeed algorithm would not maximize engagement the way the current social media algorithms do—and I cannot help but wondering how challenging it would be to assign these numerical ideology vectors in the real world.

Another method researchers have suggested—which, like the previous method, involves a fairly significant reenvisioning of how social networks should operate in order to reduce polarization—is based around the idea of decentralization. Many social networks naturally form a collection of hubs around highly influential users. If you think of each of these influencers as the center of a bicycle wheel shape with spokes emanating out to the followers, the network will look like a bunch of bicycle wheels with some but not many connections between the different wheels. A consequence of this network structure is that ideas and viewpoints tend to emerge from the limited number of central influencers and percolate outward, but they have difficulty crossing from one bicycle wheel to the next. In the case of Donald Trump on Twitter, there was a massive bicycle wheel with him at the center that encompassed much of the Republican user base on the platform and possibly impeded the flow of diverse perspectives within this user base. The general idea of decentralizing a social network is to reengineer it so that these bicycle-shaped circles of influence are less likely to emerge. This can also be seen as creating *egalitarian* networks in which users have less heavily imbalanced influence on each other.

One leading scholar advocating for decentralized networks claims[49] that in an egalitarian network "new ideas and opinions can emerge from anywhere in the community," and their spread is "based on their quality, and not the person touting them." This is in contrast to a centralized network, where it is claimed that "if the influencer at the middle shows even a small amount of partisan bias, it can become amplified throughout the entire group." This scholar goes so far as to assert that the centralized nature of social media is "one of the main reasons why misinformation and fake news has become so pervasive," because it provides "biased influencers a disproportionate impact on their community—enabling small rumors and suppositions to become amplified into widespread misconceptions and false beliefs." It is difficult to back up these bold claims, but the fact that President Trump was found[50] to be the single largest driver of coronavirus misinformation does at least modestly point in this direction. Unfortunately, however, research on the ills of

[49]Damon Centola, "Why Social Media Makes Us More Polarized and How to Fix It," *Scientific American*, October 15, 2020: https://www.scientificamerican.com/article/why-social-media-makes-us-more-polarized-and-how-to-fix-it/.

[50]Sheryl Gay Stolberg and Noah Weiland, "Study Finds 'Single Largest Driver' of Coronavirus Misinformation: Trump," September 30, 2020: https://www.nytimes.com/2020/09/30/us/politics/trump-coronavirus-misinformation.html.

centralized networks at present far outpaces research on how to prevent social media networks from becoming centralized—without explicitly deciding who people should be friends with and/or limiting the number of followers they can have.

Fake News Detection

One thing that should be apparent from the earlier section in this chapter on quantifying the spread of fake news is that, simply put, fake news typically propagates through social media in a somewhat different manner than traditional news. This is the basic principle behind a fake news detection algorithm[51] developed by a London-based AI startup called *Fabula* that was acquired by Twitter shortly after the algorithm was demonstrated. Fabula's approach uses a version of deep learning that is custom-tailored to networks; this allows the developers to train a supervised learning classifier on the task of distinguishing between real and fake news based on a vast number of predictors concerning the flow through a social media platform. The algorithm is blind to the content of social media posts—it only sees, and bases its classification on, the intricate web of user interactions that propagate content through the network (as well as the profiles of the users involved in these interactions). While this cannot possibly yield a perfect result in every individual case, as long as it is accurate in the bulk of cases, then the borderline ones can always be recommended for human inspection. One of the main challenges with this approach is that, generally speaking, we want to catch and remove fake news *before* it goes viral rather than recognizing it *after* it has already done so—hence, in order for Fabula's algorithm to be useful in practice, it needs to learn how to detect fake news very early based only on initial propagation data.

A fake news detection algorithm introduced[52] in April 2020 by a team of researchers from Microsoft Research and Arizona State University aims particularly at detecting fake news early on by combining different types of predictors. Unlike Fabula's content-blind system, this team's supervised learning classifier does draw some information from actual content, but it only uses measures that are fairly easy to extract (that is to say, it doesn't try to really read and understand and estimate the factual accuracy of content). Their system assigns sentiment scores to posts and articles in links (measuring how strongly positively or negatively worded they are) because fake news

[51] Natasha Lomas, "Fabula AI is using social spread to spot 'fake news'," *Tech Crunch*, February 6, 2019: https://techcrunch.com/2019/02/06/fabula-ai-is-using-social-spread-to-spot-fake-news/.

[52] Kyle Wiggers, "Microsoft claims its AI framework spots fake news better than state-of-the-art baselines," *VentureBeat*, April 7, 2020: https://venturebeat.com/2020/04/07/microsoft-ai-fake-news-better-than-state-of-the-art-baselines/.

tends to have a wider sentiment range than real news; it estimates how biased each user is by comparing them to a database of users with labeled bias scores, because biased users are more likely to share fake news; and the system clusters users based on their metadata (profile information and account activity statistics) and assigns users who are part of a large homogeneous cluster a lower credibility score because they are more likely to be bots or human users in extremely strong echo chambers. These measures and other related ones form the predictors in this supervised learning algorithm—and all the text processing involved is handled by Facebook's RoBERTa system that was discussed earlier. Because the predictors do involve content and not just network flow, there is a reasonable chance of catching fake news immediately before it has spread.

Just a few months later, in July 2020, a research paper[53] came out that uses content-based predictors in a supervised learning algorithm—not to detect fake news, but to "distinguish influence operations from organic social media activity." They chose only to use predictors that draw from publicly available and human-interpretable aspects of social media posts; this provides transparency and also allows the algorithm to operate on any social media platform. Their predictors include internal characteristics of the post (such as timing, word count, and whether there is a news site URL included) and external ones (such as whether there's a URL in the post with a domain that is popular in the training data for a known troll campaign). The main goal of this paper was actually less about devising a practical state-of-the-art algorithm and more about learning which aspects of social media content are most indicative of influence campaigns in which settings. In particular, they studied Chinese, Russian, and Venezuelan troll activity in the United States on Twitter, Facebook, and Reddit. They found that these predictors worked quite well overall, but which particular ones were most relevant varied quite substantially—across the different social media platforms, across the different influence campaigns, and even across the different months within each campaign. Moreover, to be able to accurately detect posts from a particular influence campaign, the algorithm needed to be trained on data from that same campaign—which certainly limits the real-world efficacy of this approach.

Another paper[54] from 2020 used supervised learning to classify news as fake versus mainstream based on its propagation through Twitter. This one looked in particular at the role played by the different kinds of interactions on Twitter

[53]Meysam Alizadeh et al., "Content-based features predict social media influence operations," *Science Advances* **6** no. 30 (2020): https://advances.sciencemag.org/content/6/30/eabb5824.

[54]Francesco Pierri, Carlo Piccardi, and Stefano Ceri, "A multi-layer approach to disinformation detection in US and Italian news spreading on Twitter," *EPJ* **9** no. 35 (2020): https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-020-00253-8.

(tweet, retweet, mention, reply, and quote), and it also compared performance between the United States and Italy. It was found that the network properties of mentions provides a fairly strong classifier and that the most important network-theoretic predictors in the United States were the same as the most important ones in Italy—but that an algorithm trained in one country did not perform well when applied out of the box in the other country, suggesting that country-specific fine-tuning is necessary. This sensitivity of fake news detection algorithms to language/region is not just a theoretical matter. It was recently found[55] that Spanish-language misinformation in the United States is flourishing on Facebook and appears to be avoiding algorithmic detection much more than English-language misinformation: in a study by Avaaz, seventy percent of misinformation in English was flagged with a warning label, whereas for Spanish it was only thirty percent. Some believe this discrepancy may have been one of the driving factors in President Trump's unexpectedly strong performance in Florida in the 2020 election.

Algorithmic Adjustments

Two of the main ways Facebook has responded to the problem of misinformation are by adding warning labels to questionable content (which is found using a combination of algorithmic detection, professional moderation, and user flagging) and demoting its ranking in the newsfeed algorithm. The Avaaz study on global health misinformation mentioned earlier[56] estimated that by going further and providing all users who have interacted with misinformation on Facebook with corrected factual information would on average cut their belief in that misinformation in half—and that "detoxing" the newsfeed algorithm by downgrading misinformation content itself as well as groups, pages, and users who have a track record of habitually posting misinformation would decrease future views of such content by eighty percent. Facebook was already moving in this direction and implementing some milder versions of these measures (and experimenting with related ones) prior[57] to the Avaaz report, and these efforts have since continued, so Avaaz's all-or-nothing framing of the impact of its proposed modifications is somewhat misleading. Nonetheless, it is still quite helpful to see a quantitative public investigation into the potential effect of these sorts of misinformation mitigation methods since Facebook's internal research is kept private.

---

[55]Kari Paul, "'Facebook has a blind spot': why Spanish-language misinformation is flourishing," *Guardian*, March 3, 2021: https://www.theguardian.com/technology/2021/mar/03/facebook-spanish-language-misinformation-covid-19-election.

[56]See Footnote 14.

[57]Tessa Lyons, "Hard Questions: What's Facebook's Strategy for Stopping False News?" *Facebook newsroom*, May 23, 2018: https://about.fb.com/news/2018/05/hard-questions-false-news/.

The last topic in this lengthy chapter is a twenty-five-year-old law that has been in the news lately for its role in allowing social media companies to avoid responsibility for the content on their platforms.

Section 230

The Communications Decency Act of 1996 was the US federal government's first real legislative effort to regulate indecent and obscene material on the internet. It includes the now-notorious Section 230 stating, among other things, that "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." At the time it was written, the primary intent of this somewhat confusingly worded passage was to prevent internet service providers from being liable for illegal content on the Web. That was before the days of social media, and Section 230 is now interpreted as a liability shield for tech companies like Google and Facebook and Twitter that host user-created content, absolving them of legal responsibility for the content on their platforms. The punchline is that these companies are mostly left to each determine their own policies and moderation methods—and their motivation for doing so is driven primarily by business considerations rather than a direct fear of legal liability.

In the ramp-up to the 2020 election, Section 230 caught the skeptical eye of people of all political persuasions: many on the left said it allows Google and the social media companies to serve up harmful extremist content and misinformation without any culpability, while many on the right said it allows these companies to discriminately censor conservatives and stifle free speech. In the spring of 2020, Twitter started labeling false and misleading tweets by President Trump about voter fraud; Trump responded by calling for a total revocation of Section 230. Just prior to that, Joe Biden also called for the revocation of Section 230—but his reason was that it allows companies like Facebook to "propagate falsehoods they know to be false."[58] Even Mark Zuckerberg has expressed support for some revisions to Section 230, though Senator Wyden—one of the co-authors of the original 1996 law—openly questioned Zuckerberg's intentions in this regard:[59] "He made his money, and now he wants to pull up the ladder behind him. The fact that Facebook, of all companies, is calling for changes to 230 makes you say, 'Wait a second'."

[58]Emily Bazelon, "The Problem of Free Speech in an Age of Disinformation," *New York Times*, October 13, 2020: https://www.nytimes.com/2020/10/13/magazine/free-speech.html.

[59]Ben Smith, "It's the End of an Era for the Media, No Matter Who Wins the Election," *New York Times*, November 1, 2020: https://www.nytimes.com/2020/11/01/business/media/ben-smith-election.html.

Just days after the 2020 election, congress invited Mark Zuckerberg and Twitter CEO Jack Dorsey to testify about their platforms, the election, and misinformation.[60] While Zuckerberg welcomed a new cross-platform regulatory framework to ensure that all large tech companies work toward—and are treated equally regarding—content moderation, Dorsey pushed back against this saying that "A centralized global content moderation system does not scale." Dorsey said his focus is on giving users more tools to customize the content they see, but he also called for reforms to Section 230 that would require more oversight of, and transparency in, recommendation/newsfeed algorithms.

Despite calls in 2020 from both then-President Trump and then-candidate Biden for Section 230 to be repealed entirely, it now appears that a milder approach will be pursued—essentially just eliminating protections for certain specific kinds of content in certain specific situations. One bill currently under consideration would hold a tech company responsible if its ranking algorithms amplified the spread of content linked to a real-world act of terrorism.[61] There has also been discussion of whether Section 230 protections should be stripped entirely from online advertising. President Biden's deputy chief of staff co-authored an op-ed[62] stating that "platforms should be held accountable for any content that generates revenue." Whatever legislative decisions ultimately occur in this debate about the fate of Section 230 will almost certainly have a huge impact on the technological approaches the tech giants take next when it comes to content moderation.

Concluding Thoughts

In 2018, Facebook announced plans for a new Oversight Board to help deliberate and adjudicate matters concerning the platform's influence on public discourse. The board, often called "Facebook's Supreme Court," comprises a range of international experts—from top scholars in media studies, law, and public policy to leaders of human rights organizations and think tanks, and even a Nobel Peace Prize winner and a former prime minister of Denmark. As an example of the board's activities, it was tasked with determining whether the indefinite suspension of Trump's Facebook account was justified and whether it should continue.

---

[60] "Zuckerberg and Dorsey Face Harsh Questioning From Lawmakers," *New York Times*, January 6, 2020: https://www.nytimes.com/live/2020/11/17/technology/twitter-facebook-hearings.

[61] David McCabe, "Tech's Legal Shield Appears Likely to Survive as Congress Focuses on Details," *New York Times*, March 9, 2021: https://www.nytimes.com/2021/03/09/technology/section-230-congress.html.

[62] Bruce Reed and James Steyer, "Why Section 230 hurts kids, and what to do about it," *Protocol*, December 8, 2020: https://www.protocol.com/why-section-230-hurts-kids.

However, it was recently pointed out[63] by two scholars at Columbia University's free speech institute that the board is so limited in scope—it focuses almost exclusively on individual instances of content removal—that it is in some sense a façade. They note that specific questions of content moderation are important, but far more consequential are the decisions the company makes about the design of its platform and the algorithms that power it: "[Facebook's] ranking algorithms determine which content appears at the top of users' news feeds. Its decisions about what types of content can be shared, and how, help determine which ideas gain traction. [...] The board has effectively been directed to take the architecture of Facebook's platform as a given."

In other words, the board provides the public impression of external regulation and a dedication to mitigating ill effects on society, but the problem of harmful content spreading on social media and the question of how to moderate it run much deeper than decisions about individual pieces of content. The real discussion must involve investigations into, and possibly a vast rethinking of, the current algorithmic approach to maximizing user engagement. You have seen in this chapter that there are already many insightful investigations into algorithmic amplification of harmful content and even some promising ideas for redesigning the structure of social networks to counteract this. These questions of algorithmic design are central to the way forward, but Facebook has conveniently left all decisions concerning them in its own hands and out of the purview of its Supreme Court.

And efforts at Facebook and the other tech giants to improve matters through algorithmic means have not always been undertaken with sufficient gusto. Cathy O'Neil, a prominent data scientist in the first generation of those calling attention to the dangers of society's overreliance on algorithms, wrote[64] in February 2021 that "My own experience with content moderation has left me deeply skeptical of the companies' motives." She said she was invited to work on an AI project at Google concerning toxic comments on YouTube, but she declined after seeing the paltry budget the project was allocated and concluded that "it was either unserious or expected to fail." She had a similar experience with an anti-harassment project at Twitter. And even as these companies make progress on some technical issues with moderation, new challenges rapidly and routinely emerge. For instance, users now use live broadcast features on Facebook and YouTube and other platforms to spread their

[63]Jameel Jaffer and Katy Glenn Bass, "Facebook's 'Supreme Court' Faces Its First Major Test," *New York Times*, February 17, 2021: https://www.nytimes.com/2021/02/17/opinion/facebook-trump-suspension.html.

[64]Cathy O'Neil, "Facebook and Twitter Can't Police What Gets Posted," *Bloomberg*, February 19, 2021: https://www.bloomberg.com/opinion/articles/2021-02-19/facebook-and-twitter-content-moderation-is-failing.

messages, sometimes to enormous audiences, and a live video feed clearly poses numerous vexing technical obstacles when it comes to moderation.

Karen Hao, a technology journalist who has for several years written articles exploring bias in—and unintended consequences of—machine learning algorithms, published an article[65] in March 2021 that she described[66] as "The hardest and most complex story I've ever worked on." The title of the article? "How Facebook Got Addicted to Spreading Misinformation." She said that "Reporting this thoroughly convinced me that self-regulation does not, cannot work" and that the article is "not about corrupt people do corrupt things […] it's about good people genuinely trying to do the right thing. But they're trapped in a rotten system trying their best to push the status quo that won't budge."

Here's the gist of Hao's ambitious article. Facebook built its massive platform by designing algorithms to maximize user engagement at all costs. Numerous internal studies concluded that algorithms designed to maximize engagement also increase polarization and amplify questionable content, but Zuckerberg and others in the company's senior leadership were fixated on engagement as a means to growth. Internal efforts to bring in more ethical considerations to the company's use of AI have mostly focused on algorithmic bias. Mitigating algorithmic bias, especially to prevent federally prohibited discriminatory behavior, is an important and challenging topic, but Facebook allegedly has not prioritized other vital tasks such as rooting out misinformation; some insiders are doubtful it could succeed at that if it tried. Moreover, some of the anti-bias efforts led to misinformation moderation algorithms being shelved when they flagged more conservative content than liberal content—even though it has long been recognized that misinformation disproportionately plagues the far right. The algorithmic, economic, and psychological themes and dynamics at play here are ones you've seen many times in different guises throughout this book, but I encourage you to read Hao's article for yourself to see how all this came together in one company and to see the human side of building, then being unable to control, this misinformation-spreading behemoth.

One additional challenge—seen particularly during the 2020 election—is that content now frequently crosses between different social media platforms, making moderation more of a complex multicompany issue than it was in the past. Curiously, and perhaps tellingly, one of the most effective instances of social media moderation seen in recent times is when the tech giants banded together to essentially pull the plug on the upstart platform Parler. After Facebook and Twitter started clamping down on pro-Trump electoral

[65]Karen Hao, "How Facebook got addicted to spreading misinformation," *MIT Technology Review*, March 11, 2021: https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

[66]Karen Hao, tweet, March 10, 2021: https://twitter.com/_KarenHao/status/1369738426048802817.

disinformation, many disaffected users left and joined Parler, which billed itself as a free speech–oriented alternative that is more friendly to politically conservative discourse. But while Parler did not have policies against the harmful content that was rapidly spreading on its platform, the tech giants powering and distributing its app (Amazon, Apple, and Google) do have such policies and responded promptly and powerfully, leaving Parler struggling to survive. At the risk of putting an overly cynical spin on the situation, I must say that these companies seem much more capable and willing to moderate each other than themselves.

Summary

This chapter opened with a survey showing that people who rely primarily on social media for their news tend to be younger, more exposed to fake news, and less informed about politics and global health than people who get their news from most other forms of media. Next, I discussed the role played by recommendation and ranking algorithms in the spread of fake news, with a particular emphasis on Facebook to give a sense of the problem on that platform and to show some of what's been said and done regarding it. I also discussed how wide-ranging and constantly evolving conspiracy movements like QAnon pose a particular challenge for moderation. I then summarized and contextualized a collection of academic papers that use sophisticated methods to study quantitatively how fake news spreads on social media (mostly Twitter because data is more accessible there than it is for Facebook) and how bots are involved. Next, I looked at some of the machine learning methods that social media companies (mostly Facebook because Twitter has been less forthcoming) have used in their fight against bots and misinformation, and then I looked at some of the methods that academic researchers have proposed for this purpose. Finally, I conveyed some of the recent debates and proclamations concerning Section 230, the law that protects social media companies from liability for the content they host, and then I closed with some thoughts on the reluctance of Facebook and others to step up their game in the fight against fake news. In the next—and final—chapter, I'll tour some of the algorithmic tools available to help in your own personal fight against fake news.

CHAPTER

9

Tools for Truth

Fact-Checking Resources for Journalists and You

> *Falsehood flies, and truth comes limping after it, so that when men come to be undeceived, it is too late; the jest is over, and the tale hath had its effect.*
>
> —Jonathan Swift, 1710

This chapter starts by discussing a collection of publicly available (and mostly free) fact-checking tools that are powered by machine learning; this is to help you know what kinds of tools are available and to provide some insight into how they work. It concludes with a brief look at the role and scope of fact-checking at Google, YouTube, Facebook, and Twitter.

Online Tools

In this section, I describe a handful of online tools that can assist with fact-checking in various ways and that involve machine learning algorithms to varying extents. This is only a sample of what's out there, and some useful products not discussed here helpfully combine multiple approaches into a single user-friendly tool. My main goal here is not to provide a comprehensive list of software packages, for such a list would surely become outdated quite

© Noah Giansiracusa 2021
N. Giansiracusa, *How Algorithms Create and Prevent Fake News*,
https://doi.org/10.1007/978-1-4842-7155-1_9

quickly. Instead, the goal is to illustrate how machine learning is used in the fact-checking process and to help you understand what's actually going on under the hood with automated and semi-automated fact-checking.

Full Fact

The London-based charity organization *Full Fact*[1] provides a range of freely available fact-checking services, including a keyword search for topics that have been fact-checked. While human experts are enlisted for the analyses driving these fact-checks, machine learning is used in several ways to assist these experts with tools that make their job faster and easier. Passages of text—typically coming from articles, political speeches, social media posts, etc.—are first broken down to individual sentences, and then Google's BERT (discussed in Chapter 6) is used to numerically encode the words in each sentence in a contextually meaningful way. This numerical encoding powers an algorithmic detection and classification of "claims" in each sentence: a statement such as "GDP has risen by 5%" is considered a quantitative claim, while "this economic policy leads to a reduction in carbon emissions" is a cause-and-effect claim, and "the economy will grow by 5%" is a predictive claim. This allows human fact-checkers to skim each document and quickly see where the claims that need fact-checking are located and what kinds of claims they are.

The BERT encodings are also used to estimate whether each identified claim matches one in the archive of previously fact-checked claims. Unsurprisingly, most claims don't just appear once, they appear in many different locations and guises, so it saves an enormous amount of time to check each claim in substance once rather than checking every instance and minor variation of it. Machine learning is also used to identify the claims that most urgently require fact-checking each day, based on current events and other factors. This hybrid approach in which machine learning assists human reviewers, rather than replacing them, is very sensible; it is, as you may recall from the previous chapter, similar in spirit to Facebook's approach to content moderation. Full Fact directly states on its web page that "Humans aren't going anywhere anytime soon—and nor would we want them to be."

Logically

A startup based in the UK and India called *Logically*[2] offers a suite of tools to combat fake news. Like Full Fact, this company uses machine learning to assist human fact-checkers rather than to replace them—or, as Logically's website

[1] https://fullfact.org/.
[2] https://www.logically.ai/.

poetically puts it, to "supplement human intelligence, not supplant it." In broad strokes, the fact-checking service works as follows. First, a user submits a link to an article or post, and Logically uses machine learning to identify the key claims in it, similar to what we saw with Full Fact except that here the claims are not classified—instead, the user is prompted to select one of the identified claims to focus on. Next, machine learning is used to search for evidence and previous fact-checks by Logically related to the selected claim. If a close fact-check match is found, then no human intervention is needed; otherwise, the claim is sent to the human fact-checking team, and a full report is returned (and added to the database of completed fact-checks) once it is ready.

The company also provides a service to "identify the accuracy and credibility of any piece of text content" by using machine learning algorithms that combine the three different types of predictors discussed in the previous chapter: content-based features that extract meaning directly from text, network-based features quantifying the spread of content across social media, and metadata-based features that consider things like who posted the content and where it originated from. But no details are provided about the predictors beyond this vague list, nor about the algorithm itself. In August 2020, Logically launched a Chrome web browser extension with some of the company's services, and in February 2021 a partnership with TikTok was announced to help detect misinformation on that platform.

Squash

The Reporter's Lab[3] housed at Duke University trialed an experimental automated fact-checking service nicknamed *Squash*.[4] It is similar to Full Fact and Logically in that the main step is to match claims to a database of human-conducted fact-checks—but the emphasis here is on real-time spoken statements, so a speech-to-text algorithm is applied first. The result is that users can watch live political speeches and events with pop-up fact-check bubbles appearing automatically, although the product is considered to be still in the research and development stage.

FakerFact

Mike Tamir, head of data science at Uber's self-driving car division, produced a free tool called *FakerFact*[5] for analyzing passages of text that users either paste in or provide a link to. While sometimes billed as an online fact-checker,

[3] https://reporterslab.org/.
[4] Jonathan Rauch, "Fact-Checking the President in Real Time," *Atlantic*, June 2019: https://www.theatlantic.com/magazine/archive/2019/06/fact-checking-donald-trump-ai/588028/.
[5] https://www.fakerfact.org/.

the website for this tool says explicitly that it "will never tell you if an article is True or Not" and that its job is instead to "enable readers to detect when an article is focused on credible information sharing vs. when the focus is on manipulating or influencing of the reader by means other than the facts." This lofty description sounds quite useful, but really the tool is just a style classifier—it uses supervised deep learning, trained on millions of documents, to label each article as journalism, wiki, satire, sensational, opinion, or agenda-driven. It doesn't try to identify facts in articles, let alone verify their accuracy; instead, it looks for patterns in word usage that correlate with these various styles. One nice feature of the tool is that instead of just giving a single predicted style label, it outputs multiple style labels each with an assigned confidence score. Another nice feature is that it often highlights particular sentences in the passage of text that were most responsible for the algorithm's choice of label(s).

However, when teaching a data science class one semester, I assigned my students to experiment with FakerFact, and we found the results quite unreliable—and often laughable. One student noted that the US Constitution was labeled opinion and satire. When I pasted in the first chapter of this book, FakerFact said "this one sounds silly" and also deemed it opinion and satire. Just now, I fed FakerFact the first link on the *New York Times*, which was an article[6] about a COVID vaccine trial, and it was labeled sensational. Curiously, some of the passages highlighted as influential for this decision actually made sense: "Miami-Dade County, which includes Miami Beach, has recently endured one of the nation's worst outbreaks, and more than 32,000 Floridians have died from the virus, an unthinkable cost that the state's leaders rarely acknowledge." But others don't seem the least bit sensational: "Two-thirds of participants were given the vaccine, with doses spaced four weeks apart, and the rest received a saline placebo." Let's hope Uber's self-driving cars are a little more accurate than this text analysis tool.

Waterloo's Stance Detection

A research team[7] based at Canada's Waterloo University broke down the fully automated fact-checking process into four steps:

1. Retrieve documents relevant to the claim in question.

[6]"Federal Health Officials Say AstraZeneca Vaccine Trial May Have Relied on 'Outdated Information'," *New York Times*, March 22, 2021: https://www.nytimes.com/live/2021/03/22/world/covid-vaccine-coronavirus-cases.

[7]Chris Dulhanty et al., "Taking a Stance on Fake News: Towards Automatic Disinformation Assessment via Deep Bidirectional Transformer Language Models for Stance Detection," presented at NeurIPS 2019: https://arxiv.org/pdf/1911.11951.pdf.

2. Determine the "stance" of each document, meaning whether it supports, rejects, or is ambivalent/unrelated to the claim.
3. Assign a credibility score to each document based on its source.
4. Assign a truthfulness score to the claim by combining the document stances weighted by the document credibility scores (and highlight relevant facts/context from the documents).

The first step can be a tricky one, because it is more open-ended than the matching step conducted by Full Fact, Logically, and Squash: rather than searching through a specific database of fact-checks for a match to the claim, one needs to scour the Web for any documents that might have information related to the claim. Google's use of BERT to analyze the text in user search queries is bringing us closer to this task, but many challenges still remain. The third step is essentially what Google and Facebook already do—through Google's use of PageRank and human evaluators and through Facebook's NEQ scores, as you saw in Chapters 6 and 8. The fourth step is straightforward once the first three are complete, so the Waterloo team decided to focus on the second step—stance detection—to see how well that could be automated.

They used a data set of fifty thousand articles to train a deep learning classification algorithm that looks at the body of each article and the headline of the article and estimates whether the body agrees with the headline, disagrees with it, discusses it without taking a stance, or is unrelated to the headline. Their algorithm scored a very respectable ninety percent accuracy, which was considerably higher than previous attempts by other researchers. The main insight in their work was to start with Facebook's massive pre-trained deep learning algorithm RoBERTa and then do additional focused training to fine-tune the algorithm for the specific task at hand. This general process of fine-tuning a massive pre-trained deep learning algorithm is called "transfer learning," and it has been an extremely successful method in AI, so it is not at all surprising that this is the right way to go when it comes to stance detection—it just wasn't possible before BERT and RoBERTa came out.

I don't believe a prototype of this Waterloo stance detection method is publicly available yet, and to really be effective we need progress on the document retrieval step as well. Nonetheless, it is promising work that may well find itself in user-friendly software in the near future.

SciFact

The Allen Institute for Artificial Intelligence (which you briefly encountered in Chapter 2 for its Grover system for detecting GPT-2 type generated text) developed a free tool called *SciFact*[8] to help with fact-checking medical claims related to COVID-19. The user types an assertion, or chooses from a list of suggestions, such as "Higher viral loads of SARS-CoV-2 are associated with more severe symptoms," and a list of medical research publications is returned, each with an estimated score of how strongly it supports or rejects the assertion—similar to the Waterloo team's stance detection—and a few potentially relevant excerpts from each publication are provided. When I tried this higher viral load assertion, seven articles were returned, four supporting and three refuting—and while not all of them seemed relevant or accurately labeled, the results were enough to show that there is not a strong scientific consensus on this issue.

Their algorithm uses BERT to process text, and it was fine-tuned as follows. First, a modestly sized collection of medical publications was assembled and the citation sentences (i.e., sentences that include a citation to another research publication) were extracted. Next, human experts manually rewrote these sentences as medical assertions; they were allowed to use the text surrounding each citation sentence, but not the paper being cited. The human experts also created negated versions of these assertions, so that the algorithm would have examples of assertions refuted by the literature, not just ones supported by it. They then went through by hand and decided whether each assertion is indeed supported by the article it cites, or refuted by it, or whether there is insufficient information to make this decision—and in the cases where it was labeled as supported or refuted, the experts highlighted the passages in the article's abstract that provided the strongest basis for this label. Roughly speaking, this process is how they trained their BERT-based algorithm to do all three tasks involved: retrieve articles relevant to a given assertion, estimate the stance of each article in relation to the assertion, and extract relevant passages from each article's abstract.

The researchers' framework in principle applies to all kinds of medical and scientific assertions—not just COVID-related ones—but since training their algorithm involves careful manual work with the data, they decided to launch this tool initially in a limited setting and scope. The reader is cautioned, and the researchers readily admit, that this tool is largely intended to show what is possible and what challenges remain, rather than to be blindly trusted. They tested a few dozen COVID-19 assertions and found the algorithm returned relevant papers and correctly identified their stance about two-thirds of the time. While this tool should not replace finding and reading papers manually, it might still help with a quick first-pass assessment of medical claims.

[8] https://scifact.apps.allenai.org/.

Diffbot

One way to record factual information is with a *knowledge graph*, a network structure that encodes interrelated descriptions of entities. Think of a collection of facts that are essentially in a subject-verb-object format, and they fit together to provide elaborations and contextualization for each other. A very simple example is nesting of information—Boston is located in Massachusetts, Massachusetts is located in New England, so there is an implied factual connection that Boston is located in New England—but many other more varied and flexible relational configurations are possible. Knowledge graphs are a convenient way of storing information in a computer that is easily searchable, sharable, and expandable. Google has built a massive knowledge graph that powers the information panels that show up on many searches. Google said that its knowledge graph draws from hundreds of sources, with Wikipedia a "commonly-cited source," and that as of May 2020 it contained half a trillion facts on five billion entities[9]—but it has been extremely reluctant to reveal any of the technical details underlying the construction of this knowledge graph.

Another massive knowledge graph is being assembled by a startup called *Diffbot*[10] that has been scraping the entire public Web for facts. (Diffbot, Google, and Microsoft are supposedly the only three companies known to crawl the entire public Web for any purpose.) It has been adding over a hundred million entities per month. Diffbot offers a handful of services based on its knowledge graph, and the CEO said[11] that he eventually wants to use it to power a "universal factoid question answering system." It is curious that he used the term "factoid" here, which in general usage can refer to either a snippet of factual information or a statement that is repeated so often that it becomes accepted as common knowledge whether or not it is actually true. I suspect Diffbot is actually capturing the latter, because it draws from all of the Web rather than just using certain vetted sources as Google does. And this worries me. We all know not to trust everything we read on the Web, so why should we trust an algorithm that gathered all of its knowledge by reading the Web? I'm optimistic that the developers at Diffbot have attempted to include only accurate information in their knowledge graph, but I'm far less optimistic that they've been successful in this regard. This project strikes me as valuable

[9]Danny Sullivan, "A reintroduction to our Knowledge Graph and knowledge panels," *Google blog*, May 20, 2020: https://blog.google/products/search/about-knowledge-graph-and-knowledge-panels/.

[10]https://www.diffbot.com/.

[11]Will Douglas Heaven, "This know-it-all AI learns by reading the entire web nonstop," *MIT Technology Review*, September 4, 2020: https://www.technologyreview.com/2020/09/04/1008156/knowledge-graph-ai-reads-web-machine-learning-natural-language-processing/.

and worthwhile but also suffused with the kind of hubris and hype that has consistently damaged the public image of artificial intelligence when the inevitable shortcomings and biases arise.

Twitter Bot Detection

Bot Sentinel[12] is essentially a public-facing version of the bot detection systems used internally by Twitter to automatically detect bot accounts. It uses machine learning techniques (along the lines discussed in the previous chapter) to detect bots on Twitter and lets users freely explore a database of these detected accounts and track their activity. *Botometer*[13] is a related free tool that lets you type in a specific Twitter username and then applies machine learning classification methods (again, like the ones discussed last chapter) to estimate whether that account is a bot or a human; it also provides bot-versus-human estimates on all the followers of the specified account. *BotSlayer*[14] is a free browser extension that helps users detect and track hashtags and other information spreading across Twitter in a coordinated manner suggestive of a bot campaign.

Google Reverse Image Search

The image search tab of Google allows you to do keyword searches that result in images on the Web rather than links to websites. You can also drag and drop an image onto the search bar in this tab, and Google will search the Web for images that are visually similar to your image; this is officially called a "reverse image search," though I think a better name might be "image-based search" to contrast it with the usual keyword-based search. This can help you fight against fake news in several ways. If you think an image in an article you're reading has been doctored, try a search with that image and you may find the original undoctored version. Even if the image hasn't been doctored, the article you're reading might be using the image misleadingly out of context (recall the deceptive caption examples from Chapter 3); searching for this image will show you where else it has appeared on the Web, which can help you track down the original context.

So how does this kind of image search work? You may remember from Chapter 3 that an autoencoder is a deep learning architecture that teaches itself how to compress data, and when applied specifically to image data, it finds numerical ways of encoding meaningful visual structure in the images. An overly simplified example would be that pictures of faces might be reduced to

[12] https://botsentinel.com/.
[13] https://botometer.osome.iu.edu/.
[14] https://osome.iuni.iu.edu/tools/botslayer/.

one number indicating the subject's age, another the hair color, another the hairstyle, etc.—but in reality the numbers autoencoders use don't have such simple human-interpretable meanings since they are generated by the algorithm itself. The most common approach to the reverse image search is the following. First, an autoencoder is used to represent every image on the Web as a numerical vector. Picture this as a road map to these images, where the numerical vectors are a higher-dimensional analogue of latitude/longitude coordinates. The beauty of autoencoders is that visually similar images will be placed near each other on this map. Then, whenever a user searches with an image, that image is converted to a numerical vector by the same autoencoder—meaning its coordinates on our map are determined—and the images returned by this search are those whose coordinates are nearest to the coordinates of this input image.

There are also tools to do reverse searches for videos. Rather than attempting to directly encode entire videos numerically, usually this is performed simply by sampling several still images from the video and then searching for other videos that contain similar still frames. In other words, the autoencoder is applied at the level of images rather than videos.

Additional Tools

Hoaxy[15] is a free keyword search tool that builds interactive network visualizations for the diffusion across Twitter of claims that have been fact-checked by one of the main fact-checking sites. *The Factual*[16] is a free mobile app and browser extension that uses machine learning to estimate the quality of news articles; it does this by combining a few different estimated quantities pertaining to the article, such as a reputation score for the journalist(s), an NEQ-type score for the publisher, and a measure of how opinionated the article's language is.

Fact-Checking on the Big Platforms

In this section, I look at the ways that Google, YouTube, Facebook, and Twitter have embedded fact-checking into their platforms.

Google

The Google search itself is certainly a common source of information that can help with fact-checking, but here I want to highlight a few additional features that more directly focus on fact-checking. I already mentioned Google's

[15] https://hoaxy.osome.iu.edu/.
[16] https://www.thefactual.com/.

knowledge graph that powers the information panels appearing on certain searches—while the information there is not always accurate, Google allows users to provide feedback and corrections, and in September 2020 Google said[17] that it "deepened our partnerships with government agencies, health organizations and Wikipedia" to increase the accuracy of the knowledge graph. I also mentioned how Google's reverse image search can be a helpful tool for uncovering the provenance of images. Let me turn now to some other helpful tools and fact-checking topics related to Google.

If you begin a Google search with the keywords "fact check," then below many of the search results Google will attempt to extract a one-sentence claim, the people or organizations making the claim, and the beginning of a fact-check of the claim provided by a fact-checking organization if this information can be found. For instance, just now I searched for "fact check AstraZeneca vaccine banned," and the top result is a link to a fact-check from Full Fact—and just below the link Google says "Claim: Seventeen countries have banned the AstraZeneca vaccine outside of the UK" and "Claimed by: Facebook users" and "Fact check by Full Fact: This is not the case. At the time of…." Google automatically includes this kind of fact-check information on certain keyword searches, and in June 2020 it started including it below some image thumbnails for certain image searches as well.[18] Google also has a tool[19] that lets users directly do keyword searches for fact-checks and browse recent fact-checks.

YouTube

In March 2019, YouTube began adding information panels written by third-party fact-checking organizations to searches for topics prone to misinformation—but these fact-checks really were about the keyword search, not any of the particular videos that it returned. This feature launched initially in India and Brazil and then reached the United States in April 2020, in large part to help deal with the flood of COVID-related misinformation.[20] Some

[17]Pandu Nayak, "Our latest investments in information quality in Search and News," *Google blog*, September 10, 2020: https://blog.google/products/search/our-latest-investments-information-quality-search-and-news.

[18]Harris Cohen, "Bringing fact check information to Google Images," *Google blog*, June 22, 2020: https://www.blog.google/products/search/bringing-fact-check-information-google-images/.

[19]https://toolbox.google.com/factcheck/explorer.

[20]"Expanding fact checks on YouTube to the United States," *YouTube blog*, April 28, 2020: https://blog.youtube/news-and-events/expanding-fact-checks-on-youtube-to-united-states/.

fact-checking organizations such as PolitiFact offer fact-checks of individual YouTube videos, though you have to access these from PolitiFact's website[21] rather than directly through YouTube.

Facebook

In addition to the efforts described in the previous chapter to remove or down-rank questionable content, Facebook also partners with a variety of fact-checking organizations to provide warnings, additional context, and fact-checks to some types of misinformative posts. In the nine months leading up to the 2020 election, Facebook placed warning labels on nearly two hundred million pieces of content that had been debunked by third-party fact-checkers. Facebook's content flagging system described in the previous chapter is used to identify posts needing fact-checks; then—much like what we saw above with some of the public fact-checking tools—machine learning is used to group posts according to the claims involved in order to reduce the number of fact-checks humans need to perform.

Since information, and misinformation, is often shared on Facebook in the form of visual memes, the company has gone to particular lengths to develop machine learning methods that can tell when two images have similar content. Rather than feeding the entire image into an autoencoder as would be done with a reverse image search, Facebook's approach is to first identify key objects in the photo and only use the subregions containing them in the autoencoding process. A company blog post[22] explains that "this allows us to find reproductions of the claim that use pieces from an image we've already flagged, even if the overall pictures are very different from each other." An extension of RoBERTa is also used to combine word and sentence embeddings with image embeddings, similar to the company's "holistic" approach to content detection mentioned in the previous chapter.

One of the fact-checking organizations that partnered with Facebook in the aftermath of the 2016 election was Snopes, but after two years of collaboration, Snopes decided to withdraw from the partnership in February 2019. Snopes' official statement[23] on the matter included the following remark: "At this time we are evaluating the ramifications and costs of providing third-party fact-checking services, and we want to determine with certainty that our efforts to aid any particular platform are a net positive for our online community,

[21] https://www.politifact.com/personalities/youtube-videos/.

[22] "Here's how we're using AI to help detect misinformation," *Facebook blog*, November 19, 2020: https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/.

[23] Vinny Green and David Mikkelson, "A Message to Our Community Regarding the Facebook Fact-Checking Partnership," *Snopes*, February 1, 2019: https://www.snopes.com/2019/02/01/snopes-fb-partnership-ends/.

publication, and staff." In an interview[24] published the same day as this announcement, Snopes' vice president of operations clarified that the main issue behind this decision was that Snopes, an organization employing only sixteen people, was overwhelmed with the flood of fact-checks Facebook required and questioned whether that labor-intensive approach made sense for both parties. He hinted that perhaps Snopes' limited resources were best put elsewhere and that perhaps Facebook needed to find a more efficient way to limit the spread of fake news on its platform. He also complained about Facebook's proprietary, platform-specific approach to fact-checking: "The work that fact-checkers are doing doesn't need to be just for Facebook—we can build things for fact-checkers that benefit the whole web, and that can also help Facebook." Another Snopes employee expressed that the organization should return to its focus on original reporting rather than diluting its efforts with the deluge of fact-checking requests for content on Facebook.

Twitter

In June 2020, just a few days after it started flagging some tweets from President Trump as potentially misleading or glorifying violence, Twitter discussed in a series of tweets[25] the platform's focus on "providing context, not fact-checking" when it comes to public discourse. For the most part, Twitter has chosen a minimalist approach and only links to fact-checking sites in limited instances. In January 2021, Twitter launched[26] a small experimental "community-driven" pilot program called *Birdwatch* in which a select group of users can add notes to anyone's tweets to provide them with additional context. Initially, only one thousand users are able to write these notes, and they are only visible on a Birdwatch website, but Twitter said it plans to expand the program and to eventually make the notes visible directly on the tweets they pertain to "when there is consensus from a broad and diverse set of contributors."

---

[24]Daniel Funke, "Snopes pulls out of its fact-checking partnership with Facebook," *Poynter*, February 1, 2019: https://www.poynter.org/fact-checking/2019/snopes-pulls-out-of-its-fact-checking-partnership-with-facebook/.

[25]Sherisse Pham, "Twitter says it labels tweets to provide `context, not fact-checking'," *CNN*, June 3, 2020: https://www.cnn.com/2020/06/03/tech/twitter-enforcement-policy/index.html.

[26]Keith Coleman, "Introducing Birdwatch, a community-based approach to misinformation," *Twitter blog*, January 25, 2021: https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html.

Summary

This chapter opened with a list of publicly available tools that use machine learning to assist with fact-checking tasks, and it closed with a brief discussion of the fact-checking tools and activities at Google, YouTube, Facebook, and Twitter. Now you can go forth and do your own part in the fight against fake news!

I

Index

© Noah Giansiracusa 2021
N. Giansiracusa, *How Algorithms Create and Prevent Fake News*,
https://doi.org/10.1007/978-1-4842-7155-1

E

F

G

Y, Z