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.

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 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.

[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.

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 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*).

[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*.)

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 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

[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.

[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 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.

[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.