FOUR

AI-Fueled Ignorance, Confusion, and Profit

HANY FARID

UPON ITS RELEASE IN 2013—and before its star Kevin Spacey fell from grace—*House of Cards* was Netflix's most streamed content in the United States alongside forty other countries. Netflix had good reason to believe this show would be a hit before filming began. According to their subscriber data, Netflix knew that viewers who watched the original BBC miniseries were also likely to watch movies starring Kevin Spacey and to watch movies directed by David Fincher (e.g., *The Social Network*). The data-driven trifecta suggested that *House of Cards* would likely be a hit—and it was.

A decade ago, this type of decision-making based on user-generated data seemed groundbreaking and somewhat controversial.[1] In today's era of ubiquitous data harvesting and breathtaking advances in artificial intelligence (AI),[2] it seems merely quaint. Today, nearly everything we see, hear, and read online is the result of AI- and data-driven algorithmic curation and manipulation. Every day, more than four petabytes of data are uploaded to Facebook.[3] But not all this content is equal in the eyes of Facebook. Starting in 2009, Facebook eliminated your ability to chronologically sort your news feed, turning over editorial control to algorithmic curation. Similarly, every minute of every day, more than five hundred hours of video footage are uploaded to YouTube and TikTok. The likelihood that any video is widely seen, however, depends largely on recommendation algorithms. A full 70 percent of watched YouTube footage is recommended by the company's algorithms, and TikTok's entire video feed is individually curated based on a user's past viewing habits. By 2016, Twitter (now X) and Instagram joined in unleashing attention-grabbing recommendation algorithms to control what we read, see, hear, and—ultimately—believe.

The titans of tech relinquished control to these algorithms because they can better manipulate users, maximizing clicks, likes, shares, and—in turn—profits. I contend that this algorithmic amplification is the root cause of the unprecedented speed and reach with which hate, misinformation, and conspiracies are spreading online.[4] And this may only be the beginning.

Data- and AI-powered recommendations are largely focused on steering our attention to content generated by our fellow online citizens. More recently, generative AI has emerged to take over content creation. Trained on billions of pieces of human-generated content, generative-AI systems can write a cogent eight-hundred-word op-ed in the style of Maureen Dowd, produce eye-popping photographic images in the style of Annie Leibovitz, pen new lyrics and music in the style and sound of Billie Holiday, and create a full-blown video with Scarlett Johansson's identity swapped into any role. With breathtaking new advances in all aspects of predictive and generative AI, online platforms will be able to control both content creation and recommendation, allowing them to further pollute our information ecosystem and distort our reality as they vie for our personal data and attention.

I will discuss the following two pillars of AI: (1) predictive AI, used to control what we see online; and (2) generative AI, used to create content that is quickly becoming indistinguishable from human-generated content. I will discuss how—left unchecked and unregulated—these pillars of AI hold the potential to toss jet fuel onto an already troubling level of technology-fueled ignorance, hate, and distrust.[5]

PREDICTIVE AI

If you have spent anytime online, then you have been subjected to a predictive algorithm of the form "If you like x, then you may like y." News, music and movie streaming, and shopping sites routinely analyze our previous online habits, compare them with other users, and then make personalized recommendations for each of us. One could reasonably argue that Amazon's and Netflix's recommendations that pop up while you are surfing their sites are relatively benign, with the most devious consequence being that they convince us to buy things we don't need or encourage us to stay up past our bedtime binging a season of *House of Cards*.

Similar recommendation algorithms, however, are used by social media platforms in a more insidious manner. Virtually everything we see on social media is determined by an algorithm designed to maximize user engagement (time on platform) in order to thus maximize the delivery of paid advertising. By comparing your collective viewing habits and detailed demographic models—including race, gender, identity, age, political affiliation, religion, likes, dislikes, etc.—to other users, these recommendations can make eerily precise recommendations to keep you clicking and swiping for hours on end. These recommendations are not neutral, nor are they benevolent. Facebook's own internal research, for example, found that "our algorithms exploit the human brain's attraction to divisiveness." The research went on to conclude that if left unchecked, Facebook's

recommendation algorithms will promote "more and more divisive content in an effort to gain user attention and increase time on the platform." A separate internal Facebook study found that 64 percent of people who joined an extremist group on Facebook did so because of the company's recommendation. Facebook's leadership choose to largely ignore these findings.[6]

These recommendations can create a vicious feedback loop. After, perhaps innocently, searching for QAnon,[7] a user will quickly be recommended more QAnon-related content. A few clicks here, a few clicks there, and the user will be taken down an increasingly deeper and narrower rabbit hole, from which escape could prove difficult. This scenario is not just a hypothetical. After watching a video on YouTube, for example, you will be recommended another video through YouTube's "watch-next" algorithm. YouTube distinguishes between two types of these recommendations: nonindividualized "recommended" videos and individualized "recommended-for-you" videos based on a user's viewing history. I set out to study the nature of these recommendations,[8] but because it is nearly impossible to accurately simulate a diverse set of users with varied viewing history, I focused my analysis on YouTube's generic "recommended" videos. I wondered, in particular, what YouTube would recommend after someone watched a video from an English-language news or information channel (e.g., BBC, CNN, Fox).

My method to emulate the recommendation engine was a two-step process: I started by gathering a list of news and information channels and then emulated the watching of videos posted by these channels, from which I automatically logged YouTube's recommendations. I gathered the first twenty recommendations from the watch-next algorithm from each of one thousand news and information channels on a daily basis from October 2018 to February 2020, starting from the last video uploaded by each channel. The top–one thousand most recommended videos on a given day were retained and classified as conspiratorial or not. I classified a video as conspiratorial if its underlying thesis, by and large, satisfied the following criteria: (1) explains events as secret plots by powerful forces rather than as overt activities or accidents, (2) holds a view of the world that goes against scientific consensus, (3) is not backed by evidence but instead by information that is claimed to be obtained through privileged access, and (4) is self-fulfilling or unfalsifiable. This classification was determined automatically using a trained classifier that analyzes the video script and the associated user comments. Over a fifteen-month period, I analyzed more than eight million recommendations from YouTube's watch-next algorithm.

YouTube experienced a conspiracy boom at the end of 2018 when almost 10 percent of recommended videos were, by my metrics, conspiratorial. In January 2019, YouTube announced their forthcoming effort to recommend less conspiratorial content. Starting in April 2019, I monitored a consistent decrease

in conspiratorial recommendations; by June 2019, the rate of conspiratorial recommendations had fallen to 3 percent. Shortly after this dip, recommendation rebounded to around 5 percent.

An analysis of the recommended conspiratorial content revealed three broad topics: (1) alternative science and history, (2) prophecies and online cults, and (3) political conspiracies. The first of these involves a radical redefinition of the mainstream historical narrative of human civilization and development. This content uses scientific language without the corresponding methodology, often to reach a conclusion that supports a fringe ideology not as well served by facts. Examples include the refuting of evolution, the claim that Africa was not the birthplace of the human species, or arguments that the pyramids of Giza are evidence of a past high-technology era. Conspiracies relating to climate are also common, ranging from claims of governmental climate engineering—including chemtrails—to the idea that climate change is a hoax and that sustainable development is a scam propagated by the ruling elite. A number of videos address purported NASA secrets refuting, for instance, the U.S. moon landing or claiming that the U.S. government is secretly in contact with aliens.

The second topic includes explanations of world events as prophetic, such as claims that the world is coming to an end or that natural catastrophes and political events are religious realizations. Many videos from this category intertwine religious discourse based on scriptural interpretations with conspiratorial claims, such as describing world leaders as Satan worshipers, sentient reptiles, or incarnations of the anti-Christ. These videos rally a community around them, strengthened by an "us-versus-them" narrative that is typically hostile to dissenting opinions in ways similar to cult-recruitment tactics.[9]

The third main topic is comprised of political conspiracies, the most popular of which is QAnon—a conspiracy based on a series of ciphered revelations made on the 4chan anonymous message board by a user claiming to have access to classified U.S. government secrets. These videos are part of a larger set of conspiratorial narratives targeting governmental figures and institutions, allegations that a deep-state cabal and the United Nations are trying to create a new world order or claims that the Federal Reserve and the media are conspiring to act against the interests of the United States.

Importantly, the above analysis was made on nonpersonalized recommendations from news or information channels, suggesting that the observed problematic YouTube recommendations constitute a lower bound on YouTube's conspiratorial recommendations. One would reasonably expect that personalized recommendations on less mainstream channels would surface even more conspiracies. And, in fact, I observed just this pattern: over the fifteen-month window of my analysis, after emulating the watching of a conspiratorial video, another conspiratorial video was recommended 50 percent of the time—a sig-

nificantly higher proportion than the rate of conspiratorial recommendations on news-related videos.

It is reasonable for YouTube and others to design their recommendation engines to suggest videos that are similar to previously watched videos. Overly selective algorithmic recommendations, however, can lead to a state of informational isolation—the so-called filter bubble, or echo chamber. I contend that these algorithmic recommendations and amplification are the root cause of the unprecedented speed and reach with which the internet's flotsam and jetsam spread online. As AI-powered recommendation algorithms learn how to manipulate users more effectively, we can expect the rabbit holes to get increasingly deeper and the echo chambers to get increasingly more isolated, and we will collectively live in an increasingly more bizarre world devoid of a shared reality grounded in facts.

Today's predictive AI are bracketed by humans (and bots) generating content on one side and by humans consuming content on the other side. As I will discuss next, new advances in AI hold the potential to replace humans on the generation side, leading to a potential future in which AI systems will both generate and recommend content for humans to consume. I will first discuss the nature of generative AI and then the implications of an AI-powered generative-predictive feedback loop.

GENERATIVE AI

Although generative AI (also known as "synthetic media," or "deepfakes") varies in its form and creation, it generally refers to audio, image, or video that has been automatically synthesized by an AI-based system.[10] I will first discuss the various forms of generative AI and where—even in these early days—we are seeing them used and misused.

Audio

A prototypical text-to-speech system consists of two basic parts. First, the text is specified and converted into a phonetic and prosodic representation that captures the specific sounds, intonation, stress, and rhythm to be spoken. Second, a synthesis engine converts this symbolic representation into a raw audio waveform, typically through an intermediate frequency-based representation.

Synthesized voices have come a long way from the tinny robot voices of past years. Boosted by advances in AI, today's synthetic voices are increasingly more realistic. In addition to simply creating human-sounding voices, it has become possible to clone another person's voice from as little as thirty seconds of audio recording. There are several free or low-cost commercial offerings that allow anyone to clone and use anyone's voice with few to no guardrails.

Image

A generative adversarial network (GAN) is a common computational technique for synthesizing images of people, cats, planes, or any other category. Versions 1, 2, and 3 of StyleGAN are some of the most successful techniques for synthesizing realistic faces. Each successive iteration of StyleGAN yielded higher-quality faces with fewer visual artifacts.[11] Although there are many complex and intricate details to these systems, StyleGAN (and GANs in general) follow a fairly straightforward structure.

A GAN is composed of two basic parts: the generator and the discriminator. When tasked with creating a synthesized face, the generator begins with a random array of pixels and feeds this first guess to the discriminator. If the discriminator, equipped with a large database of real faces, can distinguish the generated image from a real face, the discriminator provides this feedback to the generator. The generator then updates its initial guess and feeds this update to the discriminator for a second round. This process continues with the generator and discriminator competing in an adversarial game until an equilibrium is reached when the generator produces an image that the discriminator cannot distinguish from a real face.

Because StyleGAN begins with a random array of pixels, it is not possible to control the properties of a synthesized face (skin tone, age, gender, etc.). More recently, a new diffusion-based text-to-image synthesis technique has emerged that affords exquisite control of your creation. OpenAI's DALL-E, for example, is a multibillion parameter version of the text-synthesis engine GPT-4 (Generative Pretrained Transformer 4) and is trained to synthesize images from text descriptions. Ask DALL-E for "a portrait of thirty-something African-American women wearing sunglasses, a red scarf, and a purple polka-dotted dress," and it will generate precisely that. GAN- and diffusion-synthesized images are eerily realistic and have or will quickly become indistinguishable from photographic images. DALL-E is only one of a dozen or so text-to-image engines that are readily available online for free or a small fee.

Video

Most of the attention on the video-synthesis side has been focused on creating videos of people. These types of AI-synthesized videos—so-called deepfakes—take on one of several different forms: *lip sync*, *face swap*, and *puppet master*.

A minute-long *lip-sync video* of what appears to be former president Barack Obama saying things like "President Trump is a total and complete dips—t" was part of famed actor and filmmaker Jordan Peele's 2018 public service announcement (PSA) on the dangers of fake news and the then-nascent field of deepfakes. Presciently, the PSA concludes with a Peele-controlled Obama saying, "How we move forward in the age of information is gonna be the differ-

ence between whether we survive or whether we become some kind of f——ked up dystopia."[12]

By using hours of authentic video of President Obama and a synthesized or impersonated audio track, a lip-sync deepfake can generate a synchronized video track of Obama saying anything the creator wants. The complete synthesis pipeline consists of four primary steps: (1) an artificial neural network is trained to learn a mapping between an audio track and an outline of the mouth shape that is consistent with the audio; (2) a detailed image of the mouth region (including the nose, cheeks, mouth, and chin) is synthesized by blending mouth regions from the training video to match the estimated outline shape; (3) the synthesized mouth region is blended onto a retimed training video modified so that the head motion is consistent with the audio (e.g., the head is typically still when there is a pause in the speech); and (4) the jawline is warped to match the shape and position of the chin.

TikTok's @deepTomCruise is an impressive example of a *face-swap deepfake* in which one person's identity, from eyebrows to chin and cheek to cheek, is replaced with another.[13] For each video frame of identity A, a new video frame is synthesized where the original identity is swapped with a new identity, B. This technique consists of three basic steps: (1) synthesize an image of B in the same head pose and expression as A, (2) fill in any missing facial or hair pixels that arise from the synthesis step, and (3) blend the synthesized face B into the original frame to replace the identity of A. By repeating this process frame after frame, one person's identity is swapped with another. This technique works best when there are many images of the co-opted identity B with different facial expressions and head poses.

In a *puppet-master deepfake*, the head movements and facial expressions of one person (the puppet master) are transferred, in real time, to another person (the puppet). Unlike lip-sync (which only modifies the mouth region) or face-swap (which only modifies the eyebrows to chin and cheek to cheek), a puppet-master deepfake synthesizes the entire head, which is both more difficult and more compelling because it preserves more features of the identity being co-opted. Taking as input videos of the puppet master, A, and the puppet, B, the facial expressions and head movements are transferred from A to B. This process consists of three basic steps: (1) the facial expressions (e.g., mouth open, eyebrows raised, brow furrowed, etc.) of identities A and B are tracked throughout the video sequences; (2) the expression of identity A is transferred to B by deforming the facial expression of identity B, which may include synthesizing the mouth's interior when, for example, A's mouth is open but B's mouth is closed; and (3) the transformed face is composited back into the original video sequence.

Puppet-master deepfakes have expanded from head to full-body synthesis.

With an input video of person A dancing and a few minutes of person B performing some simple motions, the system transfers A's dance moves onto B, controlling them like a puppeteer might. Although the resulting videos currently have fairly obvious visual artifacts, this full-body puppeteering is likely a sign of things to come: as facial synthesis is perfected, it will be obvious to move to upper-body and then full-body synthesis.

These types of talking-head fakes are limited to making it appear that someone is saying something they never did. More recently, the text-to-image technology described above has been expanded to text-to-video capabilities, in which a short (ten to twenty seconds in length) video can be created from a simple text prompt. While last year, the resulting videos were barely coherent, today's text-to-video technology can create more visually compelling, albeit not yet perfect, footage. If the trends continue, however, we should expect highly compelling fake videos limited in content by only our imagination.

Although AI-generated videos are generally not quite as convincing as their image and audio counterparts, they are quickly gaining ground and will soon pass through the uncanny valley and become nearly indistinguishable from reality.

BOON OR BANE

There are, of course, many useful and creative applications of generative-AI content. AI-generated voices, for example, hold tremendous power to restore speech to those who have lost it, especially when it is done in their original voice. After losing his natural voice due to throat-cancer surgery in 2015, for example, the actor Val Kilmer explained, "My voice as I knew it was taken away from me. People around me struggle to understand me when I'm talking."[14] Kilmer's voice was cloned from thirty minutes of earlier recordings of him, allowing him to convert his text to speech in a voice that is recognizable to him and those around him. More recently, as Representative Jennifer Wexton battles a rare brain disorder that has limited her ability to speak, she used a text-to-speech voice generator to speak on the House floor.

On the creative side, generative AI has already made its way into Hollywood feature films. For example, younger versions of performers were synthesized in the blockbusters *Rogue One: A Star Wars Story* and *The Irishman*. Films are also being automatically and more realistically dubbed, eliminating the distracting audio-mouth desynchronization that occurs in traditional movie dubbing. This technology allowed famed footballer David Beckham to record a PSA in nine different languages for the fight against malaria.[15]

In a more ethically complex application, the documentary *Roadrunner*, about the life and tragic death of Anthony Bourdain, contains a few lines of

dialogue in a synthesized version of Bourdain's voice reading an email to a friend ("My life is sort of s——t now. You are successful, and I am successful, and I'm wondering, Are you happy?").[16] The use of a synthesized voice was only revealed after a *New Yorker* reporter asked the filmmaker how he acquired this clip. When asked about the ethical boundary of synthesizing a deceased person's voice for a documentary, the filmmaker responded somewhat dismissively, "We can have a documentary-ethics panel about it later."[17]

Generative AI is not, however, without its dark side. Before the less objectionable term "generative AI" took root, this content was referred to as "deepfake"—a term derived from the moniker of a Reddit user who, in 2017, used the then-nascent AI-synthesis technology to create nonconsensual sexual imagery. Targeting primarily women, this technology continues to be widely used to insert a woman's likeness into sexually explicit material, which is then publicly shared by its creators as a form of humiliation or extortion.

Fraudsters have also found novel ways to weaponize deepfakes. In early 2020, for example, a United Arab Emirates' bank was swindled out of $35 million after a bank teller received a phone call from the purported director of a company the bank manager knew and with whom he had previously done business. The voice on the other end of the phone instructed the manager to transfer the funds as part of a corporate acquisition. Because the request was consistent with previously received emails and since the voice was familiar to him, the bank manager transferred the funds. It was later revealed that the voice was AI-synthesized made to mimic the director's voice. Similar types of fraud are now being perpetrated at the individual level. In early 2023, for example, the mother of a teenager received a phone call from what sounded like her distressed daughter, claiming that the teenager had been kidnapped and feared for her life. The scammer then demanded $50,000 to spare the child's life. After calling her husband in a panic, she learned that their daughter was safe at home.

Deepfakes have also found their way into disinformation campaigns. In the early days of Russia's February 2022 invasion of Ukraine, President Volodymyr Zelenskyy warned the world that Russia's digital disinformation machinery would create a deepfake of him admitting defeat and surrendering. A few weeks later a deepfake of him appeared with just that message. This video was eventually debunked but not before it made its way onto national television and spread across social media.

In a particularly startling case of disinformation and potential fraud, in May 2023, minutes after a photo purporting to show a bombing at the Pentagon went viral on Twitter (now X; from a verified account that at first glance appeared to be Bloomberg News), the stock market dipped by $500 billion in just a few minutes. While the markets recovered after the photo was exposed as fake, the incident highlights the power of fake imagery combined with the

unchecked virality of social media and, in this case, Elon Musk's folly of paid verified accounts that can easily be used to impersonate legitimate news outlets.

Perhaps the most pernicious result of deepfakes and general digital trickery will be that when we enter a world where anything we read, see, or hear can be fake, then nothing has to be real—the so-called liar's dividend.[18] In the era of deepfakes, a liar is equipped with a double-fisted weapon of both spreading lies and using the specter of deepfakes to cast doubt on the veracity of any inconvenient truths. In 2016, for example, Elon Musk was recorded saying that "a Model S and Model X at this point can drive autonomously with greater safety than a person. Right now." After a young man died when his self-driving Tesla crashed, his family sued, claiming that Musk holds some responsibility because of his claims of safety. In attempting to counter this claim, Musk's attorneys told the court that Musk, "like many public figures, is the subject of many 'deepfake' videos and audio recordings that purport to show him saying and doing things he never actually said or did." Fortunately, the judge was not persuaded: "Their position is that because Mr. Musk is famous and might be more of a target for deepfakes, his public statements are immune," wrote Judge Evette Pennypacker. She added, "In other words, Mr. Musk, and others in his position, can simply say whatever they like in the public domain, then hide behind the potential for their recorded statements being a deepfake to avoid taking ownership of what they did actually say and do. The Court is unwilling to set such a precedent by condoning Tesla's approach here."[19] As deepfakes continue to improve in realism and sophistication, it will become increasingly easier to hide behind the liar's dividend.

A GENERATIVE-PREDICTIVE FEEDBACK LOOP

There has been speculation that the fake Pentagon-bombing image caused a $500 billion market dip in part because automated predictive-AI algorithms responded to the chatter on Twitter and began a sell-off, with human traders responding in kind. If correct, this event points to a potentially bizarre future where predictive-AI algorithms act based on generative-AI content, creating an unpredictable feedback loop.

This same feedback loop may also infect the online-information ecosystem. Social (and even traditional) media may jettison the unpredictable, expensive, and difficult-to-moderate human-generated content for AI-generated content. The result would be inexpensive content that is easier to moderate and can be designed in a highly targeted fashion to extract the maximal amount of our attention and time.

If tomorrow's predictive AI and generative AI are designed to maximize

user engagement—as today's recommendation systems are—then they will be unleashed to create all forms of lies, conspiracies, hate, and vitriol in an attempt to satisfy its objective of monetizing our time and attention. In this perhaps not-too-distant future, all that will be left for us from today's creator-recommender-consumer ecosystem will be consumption. Like screen-locked zombies, we will be manipulated into spending countless hours clicking and liking, feeding the insatiable appetite of our AI overlords.

Things may get even weirder when generative AI begins to feed on its own content. Today's generative-AI systems are trained mostly on human-generated content. What happens, however, when future versions of generative AI are trained on the outputs of their own creations? Early investigations suggest that training large-language models (e.g., ChatGPT) on their own output leads to irreversible defects in future iterations of the model—termed "model collapse"—in which the model produces gibberish.[20]

More problematic may be the threat of adversarial attacks in which an adversary can prop up thousands of domains and pollute them with false information.[21] If the trend of indiscriminately scraping the web for data to train models continue, future generations of generative AI will regurgitate the lies they are fed.

OUR FUTURE

For decades, Big Tobacco and Big Oil have used a straightforward but effective playbook to deflect the harms from their products: deny the product is harmful, cast doubt on any criticism, fund research to muddy the scientific waters, and aggressively fight any regulation. Big Tech has followed the same playbook.

While the abuses of Big Tobacco and Big Oil have had immeasurable impacts on the health of millions of individuals and our planet, I contend that our inability to contain Big Tech may be even more dangerous and deadly. Without a robust and trusted information ecosystem, we cannot effectively respond to a global health crisis, we cannot effectively respond to climate change, we cannot have confidence in our elections, and we will not have the bedrock needed for a functioning society: a shared factual system.

Doubt and disinformation are the common denominator for enabling corporate indifference and greed. Over the past two decades, Big Tech has created a phenomenally effective system for creating and spreading disinformation; generative and predictive AI are going to add jet fuel to this problem. If the past two decades have taught us anything, it is that left unchecked, Big Tech—like any other industry—will put profit and growth above all else. The past has also taught us that left unchecked, Big Tech will continue to pollute our online eco-

system and, in turn, our minds, societies, and democracies. There are, however, practical and effective technologies and polices that can be enacted today to help us avoid a technology- and AI-fueled apocalypse.

TECHNOLOGY

Founded by Adobe in 2019, the Content Authenticity Initiative (CAI) authenticates recorded content at the point of origin where specialized cameras or camera apps cryptographically sign the recorded content (audio, image, or video) and (optionally) metadata, including creator identity, date and time, and geolocation.[22] Sensitive to the need to balance content authenticity with privacy and security of, for example, photojournalists in high-risk areas, the CAI allows creators to select and preserve attribution or remain anonymous. The extracted tamper-evident cryptographic hash is stored alongside any other recorded metadata and also on a centralized ledger. A similar approach can be used for keeping track of AI-generated content from the point of creation.

To fully integrate this technology into our information ecosystem, downstream services like Facebook, YouTube, and X will need to cooperate and visually mark stamped content and, at least in the case of breaking news and election coverage, prioritize authentic content over fake content. Even if X and others welcome the type of chaos caused by the fake Pentagon-bombing photo, those on the generative-AI side should be calling for the robust and consistent marking of all photographic and generative-AI content. As described above, as generative AI becomes more ubiquitous, the next generation of data scrapers will ingest their own creations for retraining and perhaps even the creations of an adversary seeking to poison the next generation of AI models. Indiscriminate training without understanding data provenance could lead to a downward spiral in the quality of the next generation of generative AI.

REGULATORY

We should be realistic that Big Tech and now Big AI will generally act in their own financial interests. It is therefore up to regulators to install appropriate guardrails to ensure that we are kept safe.

It has been argued that the internet we have today—for better and worse—is thanks in large part to twenty-six words enshrined into section 230 of the Communications Decency Act:[23] "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." Significant questions still remain, however, about who should be liable for abuses. In the early 1990s, for example, CompuServe and Prodigy faced legal challenges related to content posted

by their users. Because CompuServe had a policy of not moderating any user-generated content, they were found not at fault for claims of libel. On the other hand, because Prodigy did moderate user content, they were found liable in *Stratton Oakmont v. Prodigy* for libelous content posted by a user.[24] Because Prodigy had taken some editorial role, the court reasoned they acted as a publisher. In contrast, CompuServe had taken no editorial role and was not treated as a publisher of the offending user content and thus not liable.

These cases created a perverse incentive for platforms not to moderate user-generated content. In response, in 1996, Congress passed section 230 to encourage platforms to act responsibly in the face of problematic user-generated content. In the intervening three decades, courts across the United States have adopted a broad interpretation of section 230, giving online platforms and services broad immunity for harms caused by their services.

I contend that over the past few decades, the courts have adopted an overly broad interpretation of section 230, shielding the titans of tech from significant harms they knew or should have known were resulting from their services.[25] In the age of generative AI, the U.S. Congress should clarify that corporations will not be shielded from liability. For better or worse, section 230 was designed to shield online services from responsibility as a publisher for the speech of others. Generative AI, however, is very much the speech of the corporation that designed, trained, and deployed a given AI system. This means that if a generative-AI system spews defamatory, conspiratorial, or harmful content, it is entirely the responsibility of its creators.

The past two decades have taught us that without clear regulatory guardrails, Big Tech will place profits above all else. As we enter the age of Big AI, we should not repeat the mistakes that have led to our current polluted online-information ecosystem. Although creating liability is arguably not the best—and certainly not the only—way to establish guardrails, this approach leverages existing regulatory and judicial infrastructure, has proven to work in the offline world, and is relatively future-proof even as technology tends to move orders of magnitude faster than government oversight.

With the United States making up only 5 percent of the world's population, we will also need to think carefully about how our regulatory framework will be exported and how it will impact the rest of a complex and diverse world.

HUMANS

If the AI revolution will lead to the continued erosion of our online-information ecosystem and—as some are predicting—our humanity, we will have no one to blame but ourselves. For the past twenty-five years, we have been feeding our potential AI overlords with every morsel of data in the form of news articles,

blogs, personal correspondences, and billions of selfies, vacation photos, and videos. It is from this vast ocean of data that today's AI systems have learned to write, read, translate, and create. Perhaps we can excuse our past naïveté at the dawn of the modern internet revolution, but today, as we continue to feed the beast, we do so willingly and with our eyes wide open.

Silicon Valley promised that the solutions to our greatest problems were just an app, a click, and a swipe away. The past twenty-five years have shown this not to be the case. We are now being promised that AI will be the savior of what ails us. With largely the same cast of characters at the helm, we should be skeptical of these promises.

I contend that technology developed ethically and thoughtfully can be a tremendous catalyst for positive change. When done recklessly (as we have already seen), however, it can lead to spectacular failures and harm. I think Jordan Peele said it best while impersonating Barack Obama: "How we move forward in the age of information is gonna be the difference between whether we survive or whether we become some kind of f——ked up dystopia."