Senate Committee on Commerce, Science, and Transportation
Protecting Consumers from Artificial Intelligence Enabled Fraud and Scams

Hany Farid, Ph.D.

November 2024

Biography

Hany Farid is a Professor at the University of California, Berkeley with a joint appointment in Electrical Engineering & Computer Science and the School of Information. His research focuses on digital forensics, image analysis, and human perception. He received his undergraduate degree in Computer Science and Applied Mathematics from the University of Rochester in 1989, and his Ph.D. in Computer Science from the University of Pennsylvania in 1997. Following a two-year post-doctoral fellowship in Brain and Cognitive Sciences at MIT, he joined the faculty at Dartmouth College in 1999 where he remained until 2019. He is the recipient of an Alfred P. Sloan Fellowship, a John Simon Guggenheim Fellowship, and is a Fellow of the National Academy of Inventors.

Testimony

Overview

Synthetic media – so-called deepfakes – have captured the imagination of some and struck fear in others. Although they vary in their form and creation, deepfakes refer to text, image, audio, or video that has been automatically synthesized by a AI-powered system. While not fundamentally new, today's enhanced ability to easily create, distribute, and amplify manipulated media comes with heightened risks. Reasonable and proportional interventions can and should be adopted that would allow for the creative uses of these powerful new technologies while mitigating the risk they pose to individuals, organizations, and democracies.

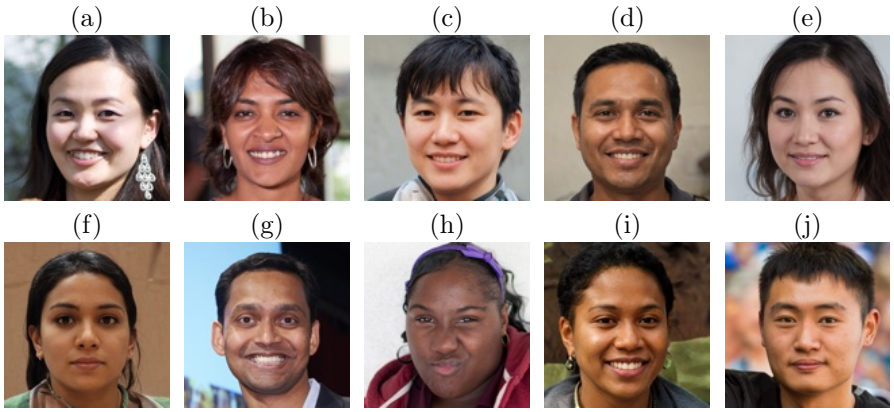

Figure 1: Half of these faces are real and half are AI generated. Can you tell which is which? See footnote at bottom of page for correct answers.

Deepfakes: Creation

Image

A generative adversarial network (GAN) is a common computational technique for synthesizing images of people, cats, planes, or any other category: generative because these systems are tasked with generating an image; adversarial because these systems pit two separate components (a generator and a discriminator) against each other; and network, because the computational machinery underlying the generator and discriminator are deep neural networks (hence the term deepfake).

StyleGAN is one of the earliest and most successful systems for generating realistic human faces, Figure 1. When tasked with generating a face, the generator starts by laying down a random array of pixels and feeding this first guess to the discriminator. If the discriminator, equipped with a large database of real faces, can distinguish the generated image from real faces, the discriminator provides this feedback to the generator. The generator then updates its initial guess and feeds this update to the discriminator in a second round. This process continues with the generator and discriminator adversarially competing until an equilibrium is reached when the generator produces an image that the discriminator cannot distinguish from real faces.[1]

Although highly realistic, GANs generally do not afford much control over the appearance or surroundings of the synthesized face. By comparison, more recent text-to-image (or diffusion-based) synthesis affords more rendering control. Trained on billions of images with an accompanying descriptive caption, each training image is progressively corrupted until only visual noise remains. The model then learns to denoise each image by reversing this corruption. This

[1]The faces in panels Figure 1(a), (b), (g), (h), and (j) are real; the faces in panels (c), (d), (e), (f), and (i) are AI generated.

Figure 2: An AI-generated image created with the prompt "an experienced word carver at work."

model can then be conditioned to generate an image that is semantically consistent with any text prompt like "an experienced word carver at work," Figure 2.

Video

Video deepfakes fall into two broad categories: text-to-video and impersonation.

Text-to-video deepfakes are the natural extension of text-to-image where a model is trained to generate a video to be semantically consistent with a text prompt. A year ago, these systems tasked with creating short video clips from a text prompt like "Will Smith eating spaghetti" yielded videos of which nightmares are made[2].

A typical video consists of 24 to 30 still images per second. Generating many realistic still images, however, is not enough to create a coherent video. These earlier systems struggled to create temporally coherent and physically

[2] `https://www.youtube.com/watch?v=XQr4Xklqzw8`

Figure 3: Selected frames of an avatar deepfake in which from a single photo of a person (bottom left), they are animated based on the movement of another person (bottom right).

plausible videos in which the inter-frame motion was convincing. Just a year later, however, these systems have improved tremendously. While not perfect, the resulting videos are stunning in their realism and temporal consistency, and quickly becoming difficult to distinguish from reality.

Although there are several different incarnations of impersonation deepfakes, two of the most popular are lip-sync and face-swap deepfakes.

Given a source video of a person talking and a new audio track (either AI-generated or impersonated), a lip-sync deepfake generates a new video track in which the person's mouth is automatically modified to be consistent with the new audio track. And, because it is relatively easy to clone a person's voice from as little as 30 seconds of their voice, lip-sync deepfakes are a common tool used to co-opt the identity of celebrities or politicians to push various scams and disinformation campaigns.

A face-swap deepfake is a modified video in which one person's identity, from eyebrows to chin and cheek to cheek, is replaced with another identity. This type of deepfake is most common in the creation of non-consensual intimate imagery. Face-swap deepfakes can also be created in real time, meaning that you will soon not know for sure if the person at the other end of a video call is real or not.

And, the latest incarnation of impersonation deepfakes are puppet-master or avatar deepfakes in which a single image of a person is animated based on the movement and speech of another person, Figurefig:avatar.

The trend of the past few years has been that all forms of image, video, and audio deepfakes continue their ballistic trajectory in terms of realism, ease of use, and accessibility.

Deepfakes: Passing Through the Uncanny Valley

First coined by Japanese roboticist Masahiro Mori in the 1970s, the term uncanny valley describes a phenomenon that occurs when a humanoid robot, or an image or video of a computer-generated human, becomes more human-like. There is a point at which the humanoid depiction becomes eerily similar to humans but is still distinguishable from real humans, causing a significant drop in our emotional comfort and acceptance. This transition is known as the uncanny valley. A humanoid depiction is said to exit the uncanny valley when it becomes so realistic that it is indistinguishable from a real person. Generative AI is well on its way to passing through the uncanny valley.

Half of the faces in Figure 1 are real and half are AI generated. Can you tell which is which? If you are like most others, your performance on this task was at near chance. A recent perceptual study found that when asked to distinguish between a real and AI-generated face , participants performed no better than guessing. In a second study in which participants were provided with training prior to completing the task, their performance improved only slightly. AI-generated faces are highly realistic and extremely difficult to perceptually distinguish from reality.

Performance is only slightly better for videos of people talking. For AI-cloned voices, a recent study found that participants mistook the identity of an AI-generated voice for its real counterpart 80% of the time, and correctly identified a voice as AI-generated only 60% of the time.

While not all forms of AI-generated content have passed through the uncanny valley, what remains will almost certainly follow in the near future. We are quickly entering an era where it is increasingly more difficult for the average person to distinguish between fact and fiction.

Deepfakes: The Good

Hardly a day goes by when I don't use some form of generative AI in my work. From using large language models (LLMs) to write or debug code to using image synthesis to create visuals for a lecture. I cannot recall any other technology that has so dramatically and so quickly altered the way I work (and in some cases, think). My colleagues and students report a similar impact in their work and studies.

Beyond personal uses cases, a particularly empowering example of the use of generative AI was by Representative Wexton of Virginia who used an AI-generated version of her voice to address lawmakers on the House floor: "My battle with progressive supranuclear palsy, or PSP, has robbed me of my ability to use my full voice and move around in the ways that I used." Because today's generative AI can clone a person's voice from as little as a 30 second recording,

Rep. Wexton was able to speak in her own voice as opposed to the tinny and slightly creepy computer-generated voices of just a few years ago.

I have little doubt that generative AI is and will continue to offer positive and exciting use cases, and be an intellectual and creative accelerant, but with a few caveats. All forms of generative AI have been trained on decades of user-generated content, in many cases without permission and in many cases in direct violation of copyright laws. Trying to justify their indiscriminate scraping of online content, OpenAI – one of the leaders in the generative-AI space – admitted it would be "impossible to train today's leading AI models without using copyrighted materials." This is a bit of a hard pill to swallow for a company that in less than 10 years has grown to a valuation of $150 billion.

Deepfakes: The Bad and The Ugly

Non-Consensual Intimate Imagery

Before the less-objectionable term "Generative-AI" took root, AI-generated content was referred to as "deepfakes", a term derived from the moniker of a Reddit user who in 2017 used this nascent technology to create non-consensual intimate imagery, NCII (often referred to by the misnomer "revenge porn," suggesting somehow that the women depicted inflicted a harm deserving of revenge). Seemingly unable to shake off its roots, generative AI continues to be widely used to insert a person's likeness (primarily women and also children) into sexually explicit material which is then publicly shared by its creators as a form of humiliation or extortion.

While it used to take thousands of images of a person to digitally insert them into NCII, today only a single image is needed. This means that the threat of NCII has moved from the likes of Scarlett Johansson, with a large digital footprint, to anyone with a single photo of themselves online.

Shown in Figure 4, for example, is an image that I generated using a free service (that doesn't allow the generation of explicit material) in which I inserted my face into an image of an inmate in an orange jumpsuit.

A recent study surveyed 16,000 respondents in 10 countries and found that 2.2% of respondents reported being a victim of NCII, and 1.8% reported creating NCII. Given that many people may not know they are victims and many may be unwilling to admit creating NCII, this is surely a lower bound. The threats of NCII are not hypothetical nor are they relegated to the dark recesses of the internet. Finding NCII content and creation tools is no further than a Google search away.

Child Sexual Abuse Imagery

The Cyber Tipline at the U.S.-based National Center for Missing and Exploited Children (NCMEC) is a national reporting system for reporting all forms of child sexual exploitation including apparent child sexual abuse material (CSAM). The majority of reports come from electronic service provides including the largest

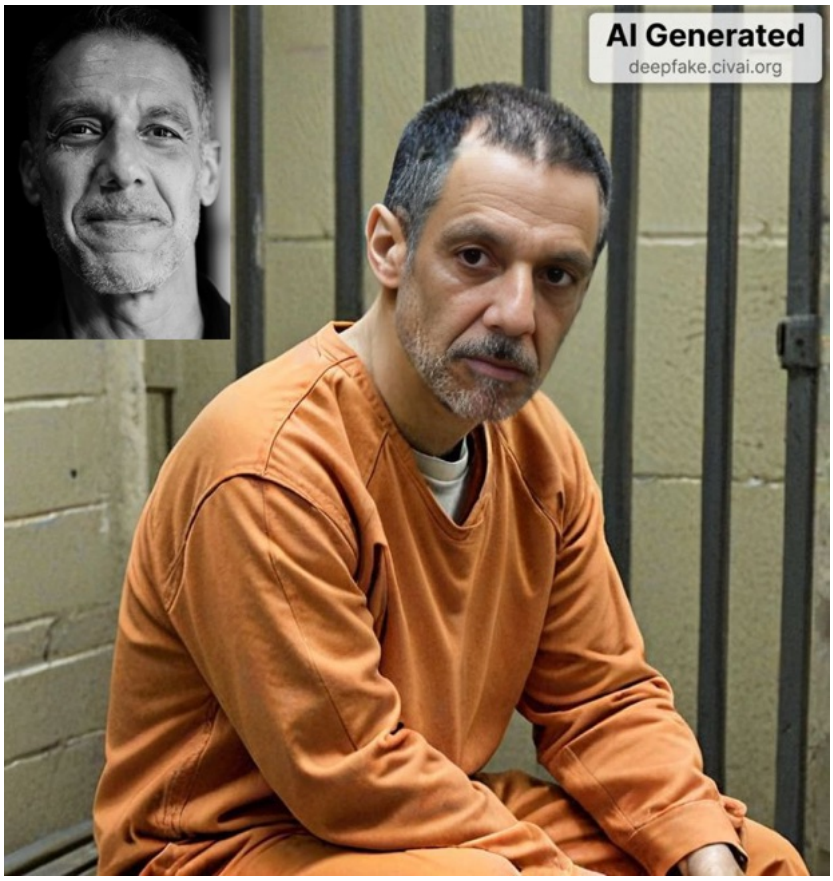

Figure 4: A deepfake in which I inserted my face (source in upper left) into an image of an inmate in an orange jumpsuit.

social media platforms like Facebook and Instagram. In 2010, NCMEC received 132,000 reports. By 2015, the number of reports grew to over 4 million, and then 21 million in 2020 and 36 million in 2023. The average age of a child depicted in this content is 12 and sometimes as young as just a few months.

Starting in 2023, NCEMC has received a small but steadily increasing number of reports that appears to be AI generated or AI manipulated. Given the escalating volume of CSAM reports over the past two decades it was, sadly, predictable that this nascent technology would quickly be weaponized in this horrific way.

18 U.S. Code § 2252A uses a standard prohibiting any visual depiction of CSAM that is "virtually indistinguishable" from a minor engaging in sexual conduct. That is, creation or possession of CSAM can extend beyond material depicting an actual child to material that is computer generated. Beyond the legal standing of AI-generated CSAM, the large-scale creation of abusive content

holds the potential to both normalize the sexual abuse of children for offenders and overwhelm an already strained CyberTipline.

While some generative-AI systems placed reasonable guardrails to prevent the creation of CSAM, others did not. Stability AI's first version of their image generator — Stable Diffusion — was open-sourced with no guardrails. In response to concerns of potential abuse, the company's founder, Emad Mostaque, said "Ultimately, it's peoples' responsibility as to whether they are ethical, moral and legal in how they operate this technology." Depending on your viewpoint, this is spectacularly naive, cynical, or simply indifferent.

Fraud

First it was Instagram ads of Tom Hanks promoting dental plans. Then it was TV personality Gayle King hawking a sketchy weight-loss plan. Next, Elon Musk was shilling for the latest crypto scam, and Taylor Swift was announcing a giveaway of Le Creuset cookware. More recently it has been Brad Pitt and Cristiano Ronaldo promoting phony medicines to treat serious diseases like cancer. All, of course, were deepfake scams.

AI-powered scams are not just impacting individuals, they are also impacting small- to large-scale organizations. Earlier this year, a finance worker in Hong Kong was tricked into paying out $25 million to fraudsters using deepfake technology to pose as the company's chief financial officer in a video conference call.

This was not the first such example. In 2019, a United Kingdom based company suffered the same fate when an imposter used an AI-synthesized voice to steal $243,000 in a similar type of scam. And, in early 2020, a United Arab Emirates' bank was swindled out of $35 million when the bank teller was convinced to transfer the funds after receiving a phone call from the purported director of a company whom the bank manager knew and with whom he had previously done business. It was later revealed that the voice was that of an AI-synthesized voice made to sound like the director. These incidents are almost certainly the canaries in the coal mine.

Similar types of fraud are also being carried out at the individual level. In early 2023, 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 demanded $50,000 to spare the child's life. After calling her husband in a panic, she learned that the daughter was safe at home.

Generative AI is a powerful new weapon in the arsenal of cyber criminals. As synthesized audio and video continue to improve in quality and accessibility, it is reasonable to predict that these technologies will continue to be used to commit a range of small- to large-scale frauds.

Disinformation

By mid-May of 2020, in the midst of the global pandemic, 28% of Americans believed Bill Gates planned to use COVID-19 to implement a mandatory vaccine program with tracking microchips. Belief in this conspiracy is not unique to Americans. In global surveys across Central and South America, the Middle East, Northern Africa, the United States, and Western Europe, 20% of the public believes this bizarre claim.

As of this year, 22% of Americans do not believe in climate change with only 54% believing that climate change is human-driven. Understanding of climate change is highly partisan with 93% of Democrats and only 62% of Republicans believing in climate change.

The far-reaching, far-right QAnon conspiracy alleges a cabal of Satan- worshipping cannibalistic pedophiles is running a global child sex-trafficking ring that was plotting against Donald Trump. A recent poll finds 37% of Americans are unsure whether QAnon is true or false, and 17% believe it to be true.

Our global health, our planet's health, and our democratic institutions are all under attack due to rampant disinformation, conspiracies, and lies. It seems likely that deepfakes will be an accelerant to disinformation campaigns that until today managed to take significant hold without accompanying visual "evidence."

Liar's Dividend

While the harms from deepfakes are real and already with us, perhaps the most pernicious result of deepfakes and general digital trickery is that when we enter a world where anything we see or hear can be fake, then nothing has to be real. 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 – the so-called liar's dividend.

In 2016, for example, Musk was recorded saying "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 deep fakes, 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 deep fake 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."

As deepfakes continue to improve in realism and sophistication it will become increasingly easier to wield the liar's dividend.

Deepfakes: Mitigations

Generative AI continues its ballistic trajectory in terms of its ability to create content that is – or soon will be – nearly indistinguishable from reality. While there are many exciting and creative applications, this technology is also being weaponized against individuals, societies, and democracies.

If we have learned anything from the past two decades of the technology revolution and the disastrous outcomes in terms of invasions of privacy and toxic social media, it is that things will not end well if we ignore, or downplay as the cost of innovation, the malicious uses of generative AI.

I contend that reasonable and proportional interventions from creation through distribution, and across academia, government, and the private sector are both necessary and in the long-term interests of everyone. I will enumerate a range of interventions that are both practical and when deployed properly can keep us safe and allow for innovation to flourish.

Academe

In criticizing the reckless use of scientific advancements without considering the ethical implications, Jeff Goldblum's character, Dr. Ian Malcolm in the 1993 blockbuster movie *Jurassic Park*, said: "Your scientists were so preoccupied with whether they could, they didn't stop to think if they should."

I am, of course, not equating advances in AI with the fictional resurrection of dinosaurs some 66 million years after extinction. The spirit of Goldblum's sentiment, however, is one all scientists should absorb.

Many of today's generative-AI systems used to create NCII and CSAM are derived from academic research. For example, *pix2pix* developed by University of California, Berkeley researchers uses a GAN to transform the appearance or features of an image (e.g., transforming a day-time scene into a night-time scene). Shortly after its release, this open-source software was used to create *DeepNude*, a software that transforms an image of a clothed woman into an image of her unclothed. The creators of *pix2pix* could and should have foreseen this weaponization of their technology and developed and deployed their software with more care.

This was not the first such abuse nor will it be the last. From inception to creation and deployment, researchers need to give more thought on how to develop technologies safely and, in some cases, if the technology should be created in the first place.

1. During the peer-review process, reviewers should assess if any ethical or safety concerns should be considered and/or addressed by the authors prior to publication.

2. While open-source deployments are of great benefit to the larger research community, this benefit should be counter-balanced by the potential risks (as we saw in the above example).

3. At both the undergraduate and graduate levels, mandatory curricular additions are needed to expose math and engineering students to more ethics, history, philosophy, political science, and a broader swath of the liberal arts than most typically see. Our future innovators need the proper scaffolding to think about the broader issues of how technology is intersecting with society and the world beyond Silicon Valley.

Creation

When text-to-image image generators first splashed onto the scene, Google initially declined to release its technology while OpenAI took a more open, and yet still cautious, approach, initially releasing its technology to only a few thousand users. They also placed guardrails on allowable text prompts, including no nudity, hate, violence or identifiable persons. Over time, OpenAI has expanded access, lowered some guardrails and added more features. Stability AI took yet a different approach, opting for a full release of their Stable Diffusion with no guardrails. And most recently Elon Musk's image generator, Grok, followed a similar course leading to all sorts of ridiculous content from Kamala Harris romantically embracing Donald Trump to Mickey Mouse wielding an AR-15, to the more offensive and dangerous.

Regardless of what you think of Google's or OpenAI's approach, Stability AI and Grok made their decisions largely irrelevant: when it comes to this type of shared technology, society is at the mercy of the lowest common denominator. Nevertheless, generative-AI systems should follow several simple rules to mitigate the harm that comes from their services, and the remaining bad actors will have to be dealt with through legislation and litigation (see below).

1. The Coalition for Content Provenance and Authentication (`https://c2pa.org`) is a multi-stake holder, open-source initiative aimed at establishing trust in digital audio, image, and video. The focus of the C2PA is creating standards to ensure the authenticity and provenance of digital content. This standard includes the addition of metadata and embedding an imperceptible watermark into content, and extracting a distinct digital signature from content that can identify content even if the attached content credentials are stripped out. Any AI-generated service should implement this standard to make it easier to identify content as AI-generated.

2. Because text-to-image and text-to-video systems are capable of producing content limited only by the imagination of the creator, some reasonable semantic guardrails should be implemented on both the input and output. On the input side, a large language model (LLM) can flag prompts that includes requests for NCII, CSAM, or other violative or illegal content. On the output side a multimodal LLM can similarly flag violative content that managed to slip through the input guardrails.

3. Although content credentials and semantic guardrails are important steps in mitigating harms, they are not infallible. Generative-AI services should

adopt a know your customer (KYC) approach common in all financial institutions. This will both put creators on notice that their content creation is not anonymous, and allow platforms to aid investigations into illegal uses of their services.

Distribution

There are three main phases in the life cycle of online content: creation, distribution, and consumption. I have addressed creation in the previous section and will address consumption next. On the distribution side, social media needs to take more responsibility for everything from the unlawful to the lawful-but-awful content that is both shared on their platforms and amplified by their own recommendation algorithms.

While it is easy to single out social media platforms for their failure to rein in the worst abuses on their platforms from CSAM, to NCII, fraud, violence, and dangerous disinformation campaigns, these platforms are not uniquely culpable. Social media operates within a larger online ecosystem powered by advertisers, financial services, and hosting/network services.

Each of these – often hidden – institutions must also take responsibility for how their services are enabling a plethora of online harms.

1. In addition to improving on their content moderation policies and enforcement, social media can create a global shared database of identified NCII as they have previously done for CSAM and terror-related content. Once NCII is identified, such a shared database would prevent NCII from being re-uploaded thus reducing the continued harm to victims.

2. Pressure to effect change on platforms rarely comes from users because we are not the customer, we are the product. The real customers are advertisers who should wield their power to effect change by insisting, for example, that their products and services not be advertised along side CSAM, NCII, and violent content. This isn't just the right thing to do, it is the smart thing to do for brand protection.

3. The largest financial services (Visa, MasterCard, PayPal, etc.) should not be in business with services that primarily host or produce NCII or other illegal and harmful content. There are at least two examples of where financial services were able to effect change when they withheld service from PornHub (for hosting CSAM and NCII) and Backpage (for enabling sex trafficking).

4. While more fraught, computing infrastructure services from GitHub to Amazon/Google/Microsoft cloud, to network services like Cloudflare can also act as better stewards. For example, at a hate-filled neo-Nazi march in Charlottesville, Virginia in 2017, violence erupted between marchers and counter protesters leading to the horrific murder of a counter-protester. In the aftermath, companies like Cloudflare came under heavy criticism for

providing services to neo-Nazi groups like Daily Stormer, and for giving them personal information on people who complain about their content. Despite initially refusing to act, Cloudflare eventually terminated the account of Daily Stormer. While these groups will eventually find another home, that doesn't mean that we should not continually make the internet – where they can amplify their hate and violence – an increasingly unwelcome place.

Consumption

When discussing deepfakes, the most common question I'm asked is "how can the average consumer distinguish the real from the fake?" My answer is always the same: "nothing." After which I explain that artifacts in today's deepfakes – seven fingers, incoherent text, mismatched earrings, etc. – will be gone tomorrow, and my instructions will have provided the consumer with a false sense of security. The space of generative AI is moving too fast and the forensic examination of an image is too complex (see next section) to empower the average consumer to be an armchair forensic detective.

There are, however, things that consumers can do to protect themselves from the being defrauded or fooled by deepfakes.

1. Protecting against fraudulent phone calls from scammers claiming to be a family member can be as simple as having an agreed upon family code word that would have to be produced when an unexpected or emergency call is received.

2. Protecting against disinformation is, of course, more challenging as more and more people get the majority of their news from increasingly louder echo chambers. Here, I propose the development of a national K-12 effort to educate students on how to strike a balance between skepticism and vigilance, how to spot signs of disinformation, how to fact check, and how to generally be better digital citizens than the previous generation.

3. Protecting against being a victim of NCII effectively requires being completely invisible online, which in today's world is nearly impossible. If you are a victim of NCII, several organizations may be able to provide assistance or advice, including the Cyber Civil Rights Initiative (`https://cybercivilrights.org`).

Legislation

Existing legislation should be sufficient to combat child sexual abuse material (CSAM) and fraud, weather AI-powered or not. Here interventions to protect the public and prosecute perpetrators are primarily limited by law enforcement resources and the inaction of the largest social media platforms. With tens of millions of CSAM reports each year to NCMEC, for example, law enforcement is simply overwhelmed. With billions of uploads each day to social media, these

platforms are incapable – and too often unwilling – to combat illegal activity on their services.

Most agree that bans or restrictions should be placed on the creation and distribution of non-consensual intimate imagery (NCII), but the law has not fully caught up with the latest technology that now makes it too easy to create and distribute this type of content.

In recent years, however, there has been a patchwork of national and international legislation enacted. In 2019, the US state of Virginia expanded its 2014 "revenge porn" laws to include synthesized or manipulated content, making it illegal to share nude photos or videos of anyone – real or fake – without their permission. California, Hawaii, New York, and Texas have similar restrictions, but as of yet there is no federal legislation. In 2021, Australia amended its laws to include synthesized or manipulated content; violations can incur both criminal charges and monetary fines.

Because the internet is borderless, nations should now band together to move from a patchwork of legislation to a consistent set of rules and regulations to combat NCII. It remains unclear, however, whether legislation can fully rein in these abuses. Hollywood actress Scarlett Johansson – a frequent target of NCII – told the *Washington Post*, "I think it's a useless pursuit, legally, mostly because the internet is a vast wormhole of darkness that eats itself."

With some exceptions including speech designed to interfere with elections or the peaceful transfer of power, mitigating the harms from various forms of political speech is complex, and legally fraught. It is not, after all, illegal for a politician to lie or for anyone to believe those lies.

Nevertheless, several states have recently passed legislation designed to protect the integrity of elections from misleading deepfakes. In 2024, in the lead up to a contentious national election in which deepfakes have already played a roll, both the states of Minnesota and California passed legislation to impose varying civil and criminal penalties to those creating, distributing, and in some cases, hosting, AI-powered election misinformation. These laws are not without controversy and they will soon be challenged on First Amendment grounds.

Nevertheless, practical and proportional responses to existing and emerging threats are within reach.

1. Despite Scarlett Johansson's perfectly reasonable assessment of the state of the internet, a combination of updating of existing legislation and crafting new legislation to combat emerging threats is necessary, if not sufficient. To date, only a handful of nations and a handful of U.S. states have moved to mitigate the harms from deepfakes. While I applaud individual U.S. states for their efforts, internet regulation cannot be effective with a patchwork of state laws. A national and coordinated international effort is required. In this regard, the European Union's *Digital Safety Act*, the United Kingdom's *Online Safety Act*, and Australia's *Online Safety Act* provide a road map for the U.S. While regulation at a global scale will not be easy, some common ground can surely be found among the U.S. and it allies, thus serving as a template for other nations to customize and

adopt.

2. In the absence of sweeping legislation, liability can be a powerful motivating factor for the technology sector to make sure their products and services are not harmful. But, penetrating the powerful liability shield of Section 230 of the Communications Decency Act, has proven challenging. Written in 1996, Section 230 provides broad immunity to online platforms (including social media) from being held liable for user-generated content, and it allows platforms to moderate content in "good faith" without being treated as the publisher or speaker of the content. The U.S. Congress has repeatedly tried (and failed) to modernize this outdated law that could not of and does not work in today's modern technology landscape. The U.S. Congress needs to revisit the issue by modernizing Section 230 to create some liability to motivate a mindset of safety by design, not safety as (at best) an afterthought.

3. On the specific issues of CSAM and NCII, more resources and training should be provided to law enforcement to provide resources for victims and support for investigations and, where appropriate, prosecutions.

Summary

There is much to be excited about in this latest wave of the technology revolution. But, if the past few technology waves have taught us anything, it is that left unchecked, technology will begin to work against us and not for or with us. We need not make the mistakes of the past. We are nearing a fork in the road for the type of future we want and what role technology will play.

Famed actor and filmmaker Jordan Peele's 2018 public service announcement on the dangers of fake news and the then-nascent field of deepfakes[3] offers words of advice and caution. The PSA concludes with a Peele-controlled President Obama saying "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****d up dystopia." I couldn't agree more.

[3]Deepfake Obama: `https://www.youtube.com/watch?v=cQ54GDm1eL0`