Creating, Using, Misusing, and Detecting Deep Fakes

Hany Farid

Abstract. Synthetic media—so-called deep fakes—have captured the imagination of some and struck fear in others. Although they vary in their form and creation, deep fakes refer to text, image, audio, or video that has been automatically synthesized by a machine-learning system. Deep fakes are the latest in a long line of techniques used to manipulate reality, yet their introduction poses new opportunities and risks due to the democratized access to what would have historically been the purview of Hollywood-style studios. This review describes how synthetic media is created, how it is being used and misused, and if (and how) it can be perceptually and forensically distinguished from reality.

1 Introduction

In 1917, two young girls in the English village of Cottingley—the cousins Elsie Wright (1901–1988) and Frances Griffiths (1907–1986)—convinced many around the world of the existence of fairies with the help of Sir Arthur Conan Doyle (ironically, the creator of the famed detective Sherlock Holmes) (Owen 1994). A series of five photographs showing the cousins frolicking with tiny fairies was taken as visual proof of a supernatural phenomena; Doyle vouched for the authenticity of the photos. Six decades after the photos captured the public's imagination, the cousins admitted they were a hoax, created from cardboard cutouts from a children's book; bewilderingly, Frances maintained that one of the photos was genuine.

For as long as we have been recording the sights and sounds around us, we have been manipulating them (Brugioni 1999). The 1917 Cottingley Fairies constituted a low-tech manipulation in which a physically altered scene was photographed. As early as the mid-1860s, photographs of US President Lincoln and General Ulysses S. Grant were manipulated to—it would appear—make them more visually compelling. In the 1930s, Joseph Stalin, Mao Tse-tung, and Adolf Hitler each attempted to erase from history those who fell out of their favor by air brushing them out of photographs. These photo manipulations required a skillfully wielded scalpel and paintbrush to physically alter a negative before re-exposing it to yield a manipulated photo.

In what would be unlikely to raise an eyebrow today, a 1982 *National Geographic* cover caused an uproar after it was revealed that the depicted Great Pyramid of Giza was digitally shifted to fit the magazine's vertical format. Sophisticated digital tools

doi:10.54501/jots.v1i4.56

for recording, editing, and distributing manipulated media became widely available over the next 40 years, leading to various forms of digital chicanery. Today, we are bombarded with manipulated imagery from the humorous to the absurdly unrealistic body images in fashion magazines and Instagram, all of which are being weaponized in the form of non-consensual pornography, political attack ads, scientific fraud, misrepresentations in the media, and disinformation campaigns designed to sow civil unrest and disrupt democratic elections.

From the scalpel and paintbrush in the darkroom to Photoshop on a laptop, the most recent development in digital tampering is in the form of automated AI-powered media synthesis and manipulation (so-called deep fakes). Leveraging modern tools in machine learning, computer graphics, and computer vision, these techniques have lowered the skill and time barriers for manipulating content. While the ability to create audio, image, and video of someone saying or doing something they never did is not fundamentally new, the true power of deep fakes is the widespread access to sophisticated technology that was previously only in the hands of Hollywood-style studios and state-sponsored actors, along with the instantaneous and global distribution channels afforded by social media.

This article reviews how deep-fake audio, images, and videos are created; how they are being used in creative and intriguing ways; how they are being misused in ethically questionable and illegal ways; and if (and how) they they can be perceptually and computationally discriminated from real content. While AI-synthesized text is beyond the scope of this review, recent breakthroughs in natural-language processing have made it possible to automatically synthesize detailed and cogent prose (Zellers et al. 2019; Brown et al. 2020).

2 Creating deep fakes

Before examining the use, misuse, and detection strategies associated with deep fakes, I describe how synthetic media is created using representative examples of recent audio-, image-, and video-synthesis techniques.

2.1 Audio

A prototypical text-to-speech system (Taylor 2009; Shi 2021; Tan et al. 2021) consists of two basic parts. First, the text to be spoken is specified and (typically) 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 machine learning, today's synthetic voices are increasingly more realistic.

Conventional wisdom in audio synthesis has avoided directly synthesizing an audio signal's raw waveform (Figure 1 [top]) because of the extremely high sampling density (16,000 or more samples per second). It was thus surprising when WaveNet produced highly realistic synthesized voices by directly synthesizing the audio waveform (Van Den Oord et al. 2016). This approach was inspired by the creators' previous work in image synthesis (Van den Oord et al. 2016). WaveNet takes as input acoustical features of the desired speech (e.g., the mel spectrum, Figure 1 [bottom]) and synthesizes the corresponding audio waveform. WaveNet's primary computational machinery is an autoregressive neural network, which generates each successive sample of the audio

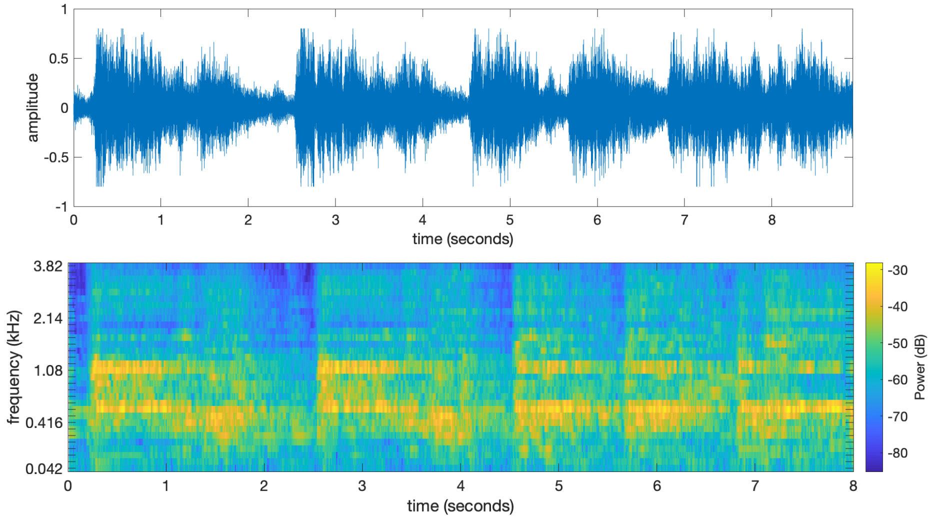

Figure 1: An 8.5-second audio clip of Handel's Messiah (top) and the corresponding frequency-based mel spectrum representation (bottom).

waveform from previously generated samples and iteratively feeds synthesized samples back into the network. This process allows the network to incorporate audio information across thousands of time stamps, yielding highly realistic sounds. The network can also be conditioned on the identity of the speaker, allowing it to generate recordings in different voices.[1]

DeepVoice employs a different network architecture to directly convert text (represented as characters or phonemes) into a spectrogram, which is then converted into an audio waveform (Gibiansky et al. 2017; Ping et al. 2017). The resulting audio is highly realistic, but a bit less so than WaveNet, as measured by user-generated scoring.

While the techniques described above focused on general-purpose voice synthesis, building a system to generate speech in one specific voice appears to be highly effective. In 2019, a small team of Dessa researchers created a system to synthesize speech in the voice of radio personality Joe Rogan (Kadhim and Palermo 2019). Follow the link in the previous reference and listen for yourself. I am not aware of any studies that formally evaluated the realism of the resulting synthesized voice, but except for a slight difference in the cadence of his speech, I find it difficult to distinguish the real Rogan from the synthesized.

Voice synthesis is also quickly entering the commercial world where, for example, a web-based application can be used to synthesize audio in your own voice.[2]

2.2 Images

A generative adversarial network (GAN) (Goodfellow et al. 2014) is arguably the most 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 (the generator and the discriminator) against each other; and *network*, because the computational machinery underlying the generator and discriminator are neural networks. Versions

[1] Audio samples of spoken English and Mandarin Chinese, as well as synthesized music, can be heard at https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio.

[2] https://www.descript.com/lyrebird

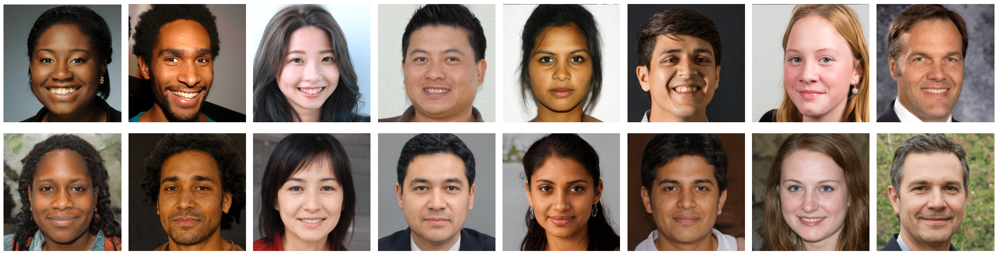

Figure 2: A representative set of real (top) and synthetically generated (bottom) faces (Nightingale and Farid 2022).

1, 2, and 3 of StyleGAN (Karras, Laine, and Aila 2019; Karras et al. 2020; Karras et al. 2021) and its precursor ProGAN (Karras et al. 2017) are some of the most successful techniques for synthesizing realistic faces. Each successive iteration of StyleGAN yielded higher-quality faces with fewer visual artifacts. Although there are many complex and intricate details to these systems, StyleGAN (and GANs in general) follow a fairly straightforward structure.

When tasked with creating a synthesized face, the generator starts by splatting down 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 the 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 competing in an adversarial game until an equilibrium is reached when the generator produces an image that the discriminator cannot distinguish from real faces.

The various versions of StyleGAN are open source, with a fully functioning version of StyleGAN2 available at https://thispersondoesnotexist.com, where each page reload yields a new synthesized face. Shown in Figure 2 (bottom) are representative examples of StyleGAN2-generated faces.

Because this process is randomly seeded, it is not possible to control the properties of a synthesized face (skin tone, age, gender, etc.). A variant of this synthesis pipeline, however, allows for a stylized mixing of facial properties. The five faces shown in Figure 3(c), for example, are generated by mixing the skin/hair tones of the face in panel (a) with the gender/age of the faces in panel (b). This type of controlled generation boosts the power of these synthesis techniques. Arbitrary images and styles can also be stylized. For example, CycleGAN can transform a cityscape photo into the style of the impressionist Monet, replace colorful trees with snow-capped trees to make a fall-time landscape appear to be shot in the winter, or change horses to zebras or vice versa (Zhu et al. 2017). While this GAN is not necessarily focused on synthesizing new faces or scenes, it can convincingly alter the semantic meaning of a photo.

Even more powerfully, this stylizing can be driven by text-based descriptions alone (Razavi, Van den Oord, and Vinyals 2019; Radford et al. 2021; Bau et al. 2021). For example, DALL-E[3]—a 12-billion parameter version of the text-synthesis engine GPT-3 (Brown et al. 2020)—is trained to synthesize images from text descriptions. Text descriptions ranging from “a photo of alamo square, san francisco, from a street” to “an armchair in the shape of an avocado” and “an emoji of a baby penguin wearing a blue hat, red gloves, green shirt, and yellow pants” generate eerily pertinent images.

[3] https://openai.com/blog/dall-e

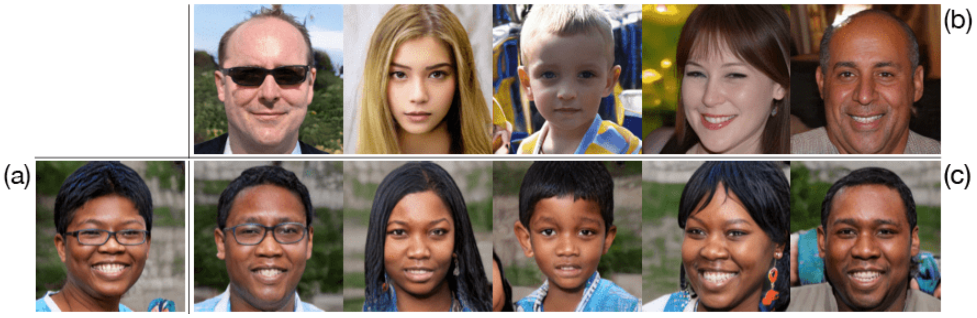

Figure 3: Mixing the skin/hair tones of the face in panel (a) with the gender/age of the faces in panel (b) yields a controlled styling of the synthesized faces in panel (c) (Karras, Laine, and Aila 2019).

2.3 Video

Video-based deep fakes take on one of several different forms: lip sync, face swap, and puppet master. Before discussing these video manipulations, I will briefly review the several decade history that predates deep-fake videos.

Before there were deep neural networks, GANs, massive data sets, and unlimited compute cycles, Chris Bregler and colleagues created what would now be called lip-sync deep fakes (Bregler, Covell, and Slaney 1997). In this seminal video-rewrite work, a video of a person speaking is automatically modified to create a video of them saying things not found in the original footage. After automatically extracting from each original video frame the mouth region and associated spoken phoneme, the application reorders the frames to match the desired text to be spoken. The mouth in the reordered frames is then stitched into the original video, correcting for differences in head position and orientation. The resulting image quality and resolution were generally lower than today's standards, but the results were nevertheless impressive.

Although the disinformation landscape 20 years ago looked radically different than today, and there were clear, positive uses of video rewriting for entertainment-related applications like movie dubbing, this work spawned prescient and early discussions of trust in the age of digital manipulation.

Video: lip sync. A minute-long video of what appears to be former President Obama saying things like "President Trump is a total and complete dip****" 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 deep fakes.[4] The PSA concludes with a Peele-controlled 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***ed up dystopia."

Peele's video was a particularity well executed example of video rewriting's successor, the machine-learning-powered, lip-sync deep fake (Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017), itself the culmination of a rich literature spawned by Bregler et al.'s original work. Using hours of authentic video of President Obama and an arbitrary audio track (either synthesized or impersonated), a lip-sync deep fake generates a synchronized video track of Obama saying anything the creator wants him to. The complete synthesis pipeline consists of four primary steps: (1) a recurrent 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 re-

[4] Deep-fake Obama: https://www.youtube.com/watch?v=cQ54GDm1eL0

gion (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 jaw line is warped to match the shape and position of the chin.

A related text-based video editing technique permits the addition, removal, and alteration of words in a video from, for example "she sells seashells by the sea shore" to "she sells ice cream by the sea shore" (Fried et al. 2019). This technique uses a similar approach to lip-sync deep fakes in which new words are assembled from other parts of the original video and blended back into the video to seamlessly replace another word or phrase.

Video: face swap. TikTok's @deepTomCruise[5] is an impressive example of a face-swap deep fake in which one person's identity, from eyebrows to chin and cheek to cheek, is replaced with another. An early and representative incarnation of this type of deep fake performs a face swap in which, for each video frame of identity *A*, a video frame is synthesized where the original identity is swapped with a new identity *B* (Nirkin, Keller, and Hassner 2019). 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.

Because face-swap deep fakes only replace the facial region, they are most convincing when the imposter's hair, body, and voice are similar to the identity being co-opted as in TikTok's @deepTomCruise, compared to, for example, the face-swap deep fake of Nicolas Cage replacing Julie Andrews as she spins on the mountain top singing "The Hills Are Alive" in the classic *The Sound of Music*.[6]

A number of open-source implementations for creating face-swap deep fakes are freely available, including faceswap-GAN,[7] faceswap,[8] FaceSwap[9], and DeepFace-Live[10].

Video: puppet master. In this type of deep fake, the head movements and facial expressions of one person (the puppet master) are transferred, in real time, to another person (the puppet) (Justus Thies et al. 2016; Nagano et al. 2018; Kim et al. 2018; J. Thies et al. 2018; Siarohin et al. 2019). Unlike lip-sync deep fakes (which only modify the mouth region), or face-swap deep fakes (which only modify the eyebrows to chin and cheek to cheek), a puppet-master deep fake synthesizes the entire head, which is both more difficult and more compelling because it preserves more features of the identity being co-opted.

Face2Face (Justus Thies et al. 2016), an early and representative puppet-master technique, takes as input videos of the puppet master (*A*) and the puppet (*B*) and transfers the facial expressions and head movements from *A* to *B*. This synthesis consists of three basic steps: (1) the facial expressions (e.g., mouth open, eyebrows

[5] https://www.tiktok.com/@deeptomcruise
[6] https://www.youtube.com/watch?v=MHkZEpfUnAA
[7] https://github.com/shaoanlu/faceswap-GAN
[8] https://github.com/deepfakes/faceswap
[9] https://github.com/MarekKowalski/FaceSwap
[10] https://github.com/iperov/DeepFacelive

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.

Open-source implementations for creating puppet-master deep fakes are available, including avatarify.[11]

Puppet-master deep fakes have expanded from head to full-body synthesis (Chan et al. 2019). 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 natural to move to upper-body and then full-body synthesis.

3 Using & misusing deep fakes

Applications of synthetic media do not always fall neatly into one category or another: good/bad, ethical/unethical, legal/illegal (Hancock and Bailenson 2021). In this section I discuss a range of applications and related ethical or legal concerns. With a few exceptions, we have a mostly incomplete understanding of how (or even if) certain forms of synthetic media should be deployed, although Danielle Citron and Robert Chesney have written thoughtfully on this topic (Chesney and Citron 2019), as has Citron in her forthcoming book (Citron 2022).

3.1 Accessibility

Synthesized voices 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, 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." Sonantic, a UK-based firm, cloned Kilmer's voice, allowing him to speak in a voice that is recognizable to him and those around him.[12] Although several hours of audio recording are usually required to recreate a voice, Kilmer's voice was recreated with only 30 minutes of audio (due to movie licensing constraints). With a simple desktop text-to-speech interface, he can now convert his written words into sound.

3.2 Entertainment

Over the past several decades the computer-graphics community has been perfecting software- and hardware-rendering techniques for creating increasingly realistic computer-generated imagery (CGI) (Holmes, Banks, and Farid 2016). These technologies have been most commonly employed in the entertainment industry to generate highly creative content ranging from short animations to feature-length films.

The rate-limiting steps in CGI are the painstaking building of detailed three-dimensional scenes followed by computationally intensive rendering of each video frame. Although training a synthetic-media system can be time consuming and computationally intensive, once trained, the synthesis engine produces content relatively quickly and

[11] https://github.com/alievk/avatarify-python
[12] https://youtu.be/OSMue60Gg6s

free of the type of manual, time-consuming work required in CGI animation. This makes synthetic content highly desirable in the entertainment industry.

Today's synthesis engines are not yet capable of the type of fine control afforded by CGI, but deep fakes have already appeared in feature films. For example, younger versions of performers were synthesized in the recent blockbusters *Rogue One: A Star Wars Story* and *The Irishman*. The startup Synthesia uses similar technology to automatically and more realistically dub movies, 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.[13].

In a more ethically complex application, the documentary *Roadrunner*, about the life and tragic death of Anthony Bourdain, contained a few lines of dialogue in a synthesized version of Bourdian's voice reading an email to a friend ("...and my life is sort of s**t now. You are successful, and I am successful, and I'm wondering: Are you happy?"). The use of a synthesized voice was only revealed after a *New Yorker* reporter asked the filmmaker how he acquired this clip (Rosner 2021). 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."

3.3 Immortality

In early 2018, 17 students were murdered and another 17 injured in a horrific school shooting in Parkland, Florida. Some two and a half years later, as part of a gun-safety voting campaign, Manuel and Patricia Oliver, the parents of slain 17-year-old Joaquin, created a deep-fake video of their son in which he is seen saying, in part: "I've been gone for two years and nothing's changed, bro. People are still getting killed by guns. Everyone knows it, but they don't do anything. I'm tired of waiting for someone to fix it. The election in November is the first one I could have voted in, but I'll never get to choose the kind of world I wanted to live in, so you've got to replace my vote." This deeply emotional and powerful video was met with mixed reactions. Some felt it crossed certain (perhaps not yet clearly delineated) ethical lines, while others felt the parents were within their rights to use their son's likeness to promote a cause that might have prevented his senseless death.

In a related theme, in 2021 the site myheritage.com[14] released Deep Nostalgia, a deep-fake technology that allows users to create short animations from a single photo of a relative. While many thought the short (audio-free) video clips were charming, some expressed concern about what they perceived as the inevitable next step in which voice would be added to create interactive avatars of deceased relatives (or anyone else). For those with an online presence that provides easy access to their likeness, voice, mannerisms, and even ways of thinking, it is perhaps reasonable to think digital avatars could be created to provide an everlasting digital likeness. Will we have to specify upon our death if we want to be kept virtually alive? Will all of this be a good or bad thing (Öhman and Floridi 2017, 2018)?

3.4 Marketing

In March 2022, Stanford Internet Observatory researchers Renée DiResta and Josh Goldstein uncovered more than 1,000 LinkedIn accounts with synthetically generated

[13] https://youtu.be/QiiSAvKJIHo
[14] https://www.myheritage.com/deep-nostalgia

profile pictures (Bond 2022). These accounts also appear to list fictitious job experiences and educational credentials. While fake profiles and padded resumes are nothing new, these accounts seem to have been created as part of a marketing scheme for companies of varying sizes. These fictitious accounts sent messages to LinkedIn users highlighting common associations or backgrounds and then pitched a product or service. The fake profiles spanned more than 70 employers, but their source seems to have been an outside marketing firm hired to bolster sales for their clients. According to LinkedIn's Professional Community Policies, inauthentic profiles, including inauthentic pictures, violate their rules: "Do not use an image of someone else, or any other image that is not your likeness, for your profile photo." The offending accounts were removed shortly after LinkedIn was notified of their existence.

While various actors have long engaged in online and offline marketing schemes, their presence on a professional network, and their flagrant misrepresentations, perhaps signal a troubling trend. As synthetic media becomes more realistic, it seems to be only a matter of time before synthetic audio and video will also be used for a range of marketing schemes, from the above to the below board.

3.5 Politics

Yoon Suk-yeol won the 2022 South Korean presidential election by a paper-thin 0.8% margin. Some have attributed part of his success to an unconventional digital strategy. Analysts have long feared that political candidates would use deep fakes to damage their opponents. It was therefore somewhat surprising to see the first high-profile political deep fakes of a candidate being created by the candidate's own campaign. Starting with hours of studio-recorded video footage of Yoon, the campaign created what it called "AI Yoon." AI Yoon offered voters an interactive experience; the AI version of the candidate was able to answer questions with snappy, meme-ready responses. For instance, when a potential voter asked "President Moon Jae-in and (rival presidential candidate) Lee Jae-myung are drowning. Who do you save?" AI Yoon retorted: "I'd wish them both good luck." Real Yoon may not have been able to be so snarky. Attracting millions of views, AI Yoon appears to have been a hit and seems to have made a middle-aged, establishment candidate more appealing to young voters.

As political candidates flock to social media, AI Yoon may be a harbinger of things to come. It remains to be seen how campaigns and voters will navigate this new medium in which AI-based candidates may be able to do and say things that real candidates cannot, or create highly customized versions of themselves to attract different demographics.

3.6 Disinformation

While some analysts worry that deep fakes would amplify existing disinformation campaigns, plunging users into even deeper wells of lies and conspiracies (Garfinkle 2020), so far such campaigns have been fueled by cheap fakes, mis-attribution fakes, and algorithmic amplification (Farid 2021). For example, in mid-May of 2020, in the midst of the global pandemic, 20% of the US public believed Bill Gates was planning to use COVID-19 to implement a mandatory vaccine program with tracking microchips (Nightingale and Farid 2020). This conspiracy spread through simple social media posts; deep fakes were not involved. The far-reaching, far-right QAnon conspiracy claims, among other things, that a cabal of Satan-worshipping cannibalistic pedophiles and child sex traffickers plotted against Donald Trump during his term as president. A poll administered in 2020 found that 37% of Americans were unsure whether QAnon was true or false, and 17% believed it was true (Newall 2020). Again, this conspiracy was created and spread through social media posts, not deep fakes. A pair of videos depicting Speaker Nancy Pelosi looking and sounding inebriated during public appearances were not—as was often mistakenly reported—deep fakes. The low-tech videos were created by simply slowing the original footage by 25% to simulate slurred speech (Spencer 2020). Sophisticated technology was not needed to convince millions that Pelosi was "blowed [sic] out of her mind," as the accompanying social media post claimed.

Deep fakes are not the common thread connecting Bill Gates's COVID-19 microchips, QAnon's Satan-worshipping cannibals, and the litany of conspiracies and outright lies polluting the internet. The common thread is the more mundane recommendation algorithms that aggressively promote the internet's flotsam and jetsam to drive user engagement and, in turn, ad revenue (Farid 2021). Where manipulated media is playing a role, it is typically in the form of cheap fakes consisting of simple Photoshop-style manipulations or slowing a video to slur speech, and misattribution fakes consisting of mislabeling a photo or video as purportedly being from a different time or place. As the quality and sophistication of synthetic media continues to improve, however, it is reasonable to predict that deep fakes may be used to fuel disinformation campaigns.

A study carried out in India, for example, found that fake videos bolstered the credibility and virality of misinformation more than audio or text alone (Sundar, Molina, and Cho 2021). The study found that socio-demographic factors were relevant (categorized as urban vs. rural, an acknowledged crude proxy for a range of user differences), with urban users somewhat less likely to find fake stories and videos credible or share them. These results fit the conventional wisdom that "seeing is believing."

A recent study, however, challenges this conventional wisdom (Groh, Sankaranarayanan, and Picard 2022). After seeing or reading a short political statement by either Joe Biden or Donald Trump, participants were asked to determine if the statement could be attributed to either individual. Participants more accurately assessed the truthfulness of political speeches from video than from written transcripts. These results cannot be explained by unrealistic fake videos because users were not able to reliably detect video manipulations, and were less capable of assessing truthfulness in closed-captioned silent videos. This contradictory finding may be the result of familiarity with the speakers and what they have said or are likely to say, as was the case in an earlier study (Sundar, Molina, and Cho 2021) which found that familiarity with statements reduced perceived credibility.

Regardless of how or even if they will play a direct role in fueling disinformation, deep fakes have already had an indirect impact on fueling disinformation in the form of the liar's dividend (Chesney and Citron 2019). The liar's dividend posits that when we enter a world where any audio, image, or video recording can be manipulated, then nothing has to be real, providing the liar with the double-fisted weapon of both spreading falsehoods and using the specter of digital manipulation to cast doubt on the veracity of any inconvenient truths. For example, candidate Trump initially apologized in 2016 for his earlier misogynistic remarks caught on the Access Hollywood tape. Just a year later, President Trump deployed the liar's dividend, claiming the recording was fake.

The power of the liar's dividend also extends to those receiving information that may run counter to their world view. In one study, for example, informing participants about deep fakes did not enhance their ability to detect them, but did empower them to dismiss real videos as fake (Ternovski, Kalla, and Aronow 2022).

3.7 War

In the early days of Russia's February 2022 invasion of Ukraine, President Zelenskyy warned the world that Russia's digital disinformation machinery would create a deep fake of him admitting defeat and surrendering. A few weeks later a deep fake of him appeared with just this message. This video was quickly debunked thanks to the rather crude audio and video and to Zelenskyy's pre-bunking. This type of deep fake, however, is likely just the beginning of a new digital front that we might expect in future conflicts.

On the other side, a video released and promoted by the Ukrainian government of the Eiffel Tower destroyed by missiles, and Parisian neighborhoods consumed in smoke, sirens, and screams amplified by overhead jets, ends with a disclosure that the video is not real and the dire warning "Just think if this were to happen in another European capital. Close the sky over Ukraine. If we fall, you fall." While nobody can blame the Ukrainians for releasing a dramatic and powerful video with the goal of rallying global support, circulating a manipulated video makes it harder to obtain reliable information, particularly in the fog of war. It also plays into the hands of those attempting to deny visual evidence of Russian aggression and war crimes: if this video could be faked, the average person may wonder, what else is fake?

3.8 Science

Cases of scientific misconduct involving manipulated images began to rise in early 2000 (Farid 2006; Bik, Casadevall, and Fang 2016; Acuna, Brookes, and Kording 2018). Before the advent of GANs, unscrupulous researchers could employ Photoshop and similar photo-editing software to edit biomedical imagery to subtly or dramatically change the interpretation of their scientific results. Twenty years later, the same GAN machinery that can be used to synthesize faces (see above) can be adapted to synthesize different types of biomedical imagery (L. Wang et al. 2022). These new synthesis techniques further lower the barriers to committing image-based scientific fraud, and may make it even more difficult to computationally or perceptually detect these manipulations.

3.9 Fraud

In early 2020, a United Arab Emirates' bank was swindled out of $35 million. A bank teller was convinced to transfer the funds after receiving a phone call from the purported director of a company who 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 describing the acquisition, 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.

This was not the first time AI-synthesized content had been used to steal large sums of money. In 2019, a UK-based company suffered a similar fate when an imposter used an AI-synthesized voice to steal $243,000 in a similar type of scam. These two incidents are almost certainly the canaries in the coal mine. As synthesized audio and video continue to improve in quality and accessibility, it is reasonable to predict that they will be used to commit future fraud. Of particular concern will be attacks that combine sophisticated synthesized content with individualized attacks powered by personal data leaked in all-too-common data breaches, as has occurred in recent phishing scams.

3.10 Non-consensual sexual imagery

The term "deep fake" comes from the moniker of a Reddit user who in 2017 used the then nascent AI-synthesis technology to create non-consensual sexual imagery (often referred to by the misnomer "revenge porn," suggesting somehow that the women depicted inflicted a harm deserving of revenge). Targeting primarily female celebrities, politicians, public figures, and sometimes citizens who attract unwanted attention, this abusive application immediately tarnished the reputation of the underlying synthesis technologies.

By 2019, the $50 DeepNude app allowed a user to digitally remove clothing from any image of a woman, rendering a detailed nude photo. The creator removed the app shortly after its release due to wide-spread criticism. However, several copycats quickly emerged. There are now multiple apps and websites to create non-consensual sexual imagery with tag lines like "the superpower you always wanted," "see any girl clothless [sic] with the click of a button," "undress anyone online," and, of course, "...we also do not take any responsibility for images created with this software."

There have been a number of legal responses to this weaponization of technology against (primarily) women (Citron and Franks 2014; Citron 2022). 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. As of 2022, California, Hawaii, and New York have similar restrictions. In 2021, Australia amended its laws to include synthesized or manipulated content; violations can incur both criminal charges and monetary fines.

Yet it remains unclear whether legislation can rein in these abuses. Hollywood actress Scarlett Johansson—a frequent target of non-consensual sexual imagery—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."

4 Detecting: Perceptual

From investigative journalists to amateur sleuths, law enforcement, and the average citizen browsing their social media feed, we often have to rely on our own senses and reasoning to assess whether an image or video is real or fake. In this section, I review recent studies describing how well human observers can distinguish between real and fake audio, image, and video.

4.1 Audio

Using audio samples from the 2019 ASVspoof Challenge (X. Wang et al. 2020), a recent study examined individuals' ability to detect audio deep fakes (Müller, Markert, and Böttinger 2021). The study participants (n = 200) correctly classified 80% of the audio samples (7,355 real and 4,919 synthesized). The study author's AI-based detector significantly outperformed the human participants, but where the AI system struggled on specific audio samples, humans excelled and vice versa.

Voice conversion (VC) encompasses computational techniques for changing a recorded voice to sound like another person's voice, without changing the semantic meaning of what was originally spoken. VC predates deep fakes by 65 years, although recent efforts have incorporated deep learning (Sisman et al. 2020). The Voice Conversion

Challenge (VCC) [15] is a semiannual event inviting researchers to develop and submit VC techniques, which are evaluated for naturalness (rated from 1 = completely unnatural to 5 = completely natural) and speaker identity (rated on a scale of “same, absolutely sure,” “same, not sure,” “different, not sure,” or “different, absolutely sure”). In the first challenge in 2016, 200 native listeners rated the VC results from 17 submissions. The best-performing system received an average of 3.0 on the five-point naturalness scale, and 70% of the samples were judged on identity to be “same.” In the second challenge in 2018, 260 native listeners rated 23 submissions. The best-performing system received an average 4.1 naturalness score, and 80% of the samples were judged on identity to be “same.” In the most recent challenge in 2020, several systems achieved nearly perfect identity ratings, and naturalness scores continued to hover around 4.0. Over the years voice conversion has continually improved in quality and naturalness, with this trend expected to continue.

Google’s Tactoron 2 and a modified version of DeepMind’s WaveNet technologies combine to convert text into a spectogram (a frequency-based representation of sound) and then convert the spectogram into speech. The authors of WaveNet (Van Den Oord et al. 2016) asked participants to rate the naturalness of their synthesized speech on a five-point scale (1 = Bad; 5 = Excellent). Compared to an average rating of 4.55 on real US English human speech, participants rated WaveNet 4.21. Real Chinese Mandarin speech was rated 4.21 vs. WaveNet’s 4.08. Although difficult to compare directly, these results are slightly better than the best-performing VCC models.

These recent studies on the perceptual naturalness of synthesized speech suggest that it will soon be possible to synthesize speech that is indistinguishable from real speech. Synthesized English-language speech will likely arrive at this milestone first, with other languages following closely behind.

While naturalness and identity are important qualities to consider in synthetic speech, other characteristics may also be important. For example, an interactive, automated conversational agent deployed to counsel those in need should sound trustworthy and honest. Persuasiveness may be desirable (although ethically questionable) for advertising and marketing (see (Dubiel et al. 2020) for an examination of the persuasiveness of synthetic voices.) Different speech patterns (e.g., rate, pitch, volume, fluency) influence their perceived persuasiveness, intelligence, and attractiveness, among other characteristics. As synthetic voices become more realistic, it will become more important to study these other perceptual qualities.

4.2 Images

Although the GANs described in Section 2 can synthesize any type of image (people, cats, planes, etc.), I focus on images of people, which are arguably some of the most intriguing. Three recent perceptual studies (Nightingale and Farid 2022) examined the ability of trained and untrained observers to distinguish between real and synthesized faces. Shown in Figure 2 are representative examples from the full dataset of 400 real and 400 synthetic faces. The synthetic faces were generated using the state-of-the-art StyleGAN2 (Karras et al. 2020), ensuring diversity across gender, race, and apparent age. For each synthesized face, the corresponding real face was matched in terms of gender, age, race, and overall appearance.

In the first study, 315 paid, online participants were shown—one at a time—128 faces, half of which were real, and asked to classify each as either real or synthetic. The average accuracy on this task was 48.2%, close to chance performance of 50%. While

[15] http://vc-challenge.org

participants were equally likely to say a real face was synthetic as vice versa, there was a significant gender × race interaction: synthetic white faces were harder to distinguish than other races, and synthetic white male faces were harder to distinguish than female white faces. This racial and gender difference is hypothesized to be the result of male white faces being over-represented in the StyleGAN2 training data set, leading to their more realistic appearance.

In a second study, 219 new participants were initially provided with a brief training on examples of specific rendering artifacts that can be used to identify synthetic faces. Throughout the experiment, participants were also provided with trial-by-trial feedback informing them if their response was correct. This training and feedback led to a slight improvement in average accuracy from 48.2% to 59.0%.

A third study examined if participants were able to glean any anomalous properties of synthetic faces as an indirect measure of realism. A new set of 223 participants rated the trustworthiness of faces on a scale of 1 (very untrustworthy) to 7 (very trustworthy). The average rating for synthetic faces (4.82) was higher than for real faces (4.48). Although only 7.7% more trustworthy, this difference is statistically significant. Black faces were rated as slightly more trustworthy than South Asian faces, but otherwise there was no effect across race, but women (4.94) were rated as significantly more trustworthy than men (4.36). The authors hypothesized that synthetically generated faces are more trustworthy than real faces because synthesized faces tend to look more like average faces—an artifact of the GAN generation process—which themselves are deemed more trustworthy (Sofer et al. 2015).

The finding that synthetically generated faces are nearly indistinguishable from real faces is consistent with those of other studies (Hulzebosch, Ibrahimi, and Worring 2020; Lago et al. 2021). This result, however, is not without its caveats. The 400 synthetic faces used in (Nightingale and Farid 2022), for example, were manually selected from a larger set of images to only include those with a mostly uniform background and without any obvious rendering artifacts. This culling makes the perceptual task harder. In practice, this means that some synthesized faces from, for example, https://thispersondoesnotexist.com, will be easily detected. Because the synthesis process is so fast and easy, however, it is reasonable to assume that a patient fraudster can apply the same culling process to avoid any obvious visual artifacts.

While synthesized faces are highly realistic, there are some clues that can help distinguish them from real faces: (1) StyleGAN-synthesized faces have a common structure consisting of a mostly front-facing person from the neck up and with a mostly uniform or nondescript background (Figure 4); (2) facial asymmetries are a tell-tale sign of a synthetic face—this is often seen in mismatched earrings (Figure 4(a)), or glass frames; (3) when viewing a face, we tend to focus most of our attention in a "Y"-pattern shifting our gaze between the eyes and mouth, often missing some glaring artifacts in the background which may contain physically implausible structures (Figure 4(b)); and (4) StyleGAN-synthesized faces enforce a facial alignment in the training and synthesis steps resulting in the spacing between the eyes being the same and the eyes being horizontally aligned in the image. Shown in Figure 4(c), for example, is a superposition of all eight faces shown in the lower portion of the figure, demonstrating that the eyes are nearly perfectly aligned across the synthesized faces.

4.3 Video

While synthesized faces are nearly indistinguishable from real faces, it is more technically challenging to synthesize videos in which a person's identity is co-opted to make them say something they never did. More artifacts are also likely to be left

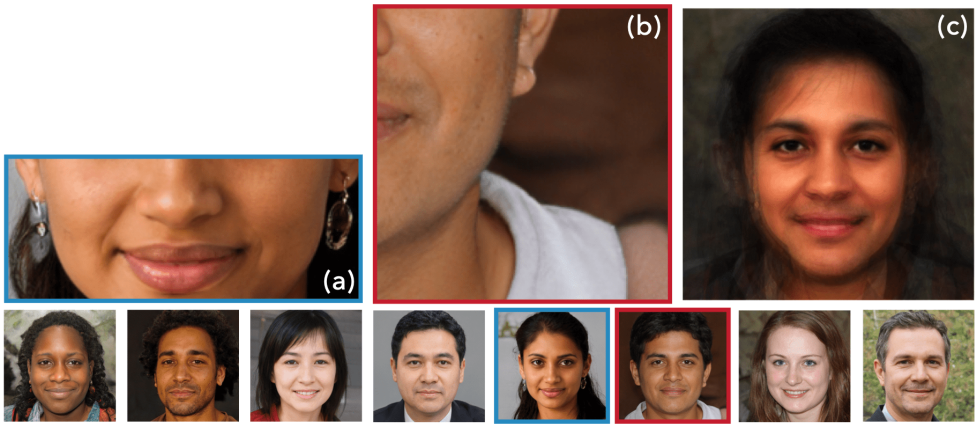

Figure 4: Eight synthesized faces (bottom) and three common synthesis artifacts: (a) asymmetric earrings; (b) physically implausible background structures; and (c) a superposition of all eight faces reveals a nearly perfectly alignment of the eyes.

behind in the individual video frames as well as the temporal relationship between video frames, and the accompanying audio.

A recent set of perceptual studies (Groh et al. 2022) examined untrained observers' ability to identify 50 real and synthesized 10-second videos from the Deepfake Detection Challenge (DFDC) (Ferrer 2020) of unknown actors making uncontroversial statements in nondescript locations. In one experiment, participants viewed a single video and were asked to categorize it as real or synthesized, and specify how confident (on a scale of 50% to 100%) they were in their assessment. Each response (normalized from 0 to 1) was scored proportional to correctness: a correct response with 85% certainty, for example, yields a response score of 0.85, whereas an incorrect response at this level of certainty is scored 1.00 – 0.85 = 0.15. Participants achieved an average score of ≈0.66, compared to a chance performance of 0.50. Assigning to a video with a score greater than 0.50 as "correct," participants correctly identified ≈66% of the deep-fake videos (in another experimental condition, participants achieved greater accuracy when a real and corresponding fake video were presented side by side). By comparison, the leading DFDC predictor correctly identified 80% of the deep-fake videos, although this accuracy is significantly higher than the some 65% accuracy achieved by the top-performing model on the entire DFDC holdout dataset of 4,000 videos.

Pooled responses from all participants yields an improved crowd score of ≈0.80 and the number of correctly identified videos increases to 80%, nearly the same as the top-performing predictor. This crowd wisdom suggests that identification errors are at least somewhat independent, with some participants noticing artifacts that others do not.

In a second condition, after specifying their confidence in their assessment, participants were shown the prediction by the top-performing DFDC model and given the opportunity to update their response. Participants updated their response in 24% of the trials, half of which led to a crossing of the 50% confidence threshold. In this collaborative condition, participants correctly identified 73% of the deep-fake videos, up from 66% in the first condition.

Although perceptual detection of deep-fake videos is—unsurprisingly—harder than still images, participants have some ability to perform this task. There is also wisdom in the crowd: pooling responses from multiple, independent moderators improves accuracy, as does giving moderators more information from computational predictors (discussed in the next section).

5 Detecting: Computational

Human observers can distinguish between real and fake content with varying degrees of efficacy. If past trends continue, it is reasonable to predict that synthetically generated audio, image, and video will eventually become indistinguishable from reality. It thus seems reasonable to look for relief from the computational machinery used to create this content. In the remainder of the section I discuss representative examples of recent audio-, image-, and video-synthesis detection techniques.

5.1 Audio

The computational detection of synthesized human speech is relatively nascent compared to detection techniques for images and videos (see below). However, speaker recognition (what somebody is saying) and speaker identification (who is saying it) have been well studied, and many of the underlying computational approaches used to identify speakers have been shown to be capable of distinguishing real from synthetic speech.

Audio: artifact based. The most common approach to detecting synthetic speech is to quantify artifacts introduced by the audio-synthesis pipeline. This approach consists of three basic parts: (1) the raw audio signal is transformed into a representation that reveals discriminating features; (2) specific features are measured from the transformed signal; and (3) any of a number of pattern-recognition techniques is trained to distinguish the extracted features between real and synthetic speech.

The popular mel-frequency cepstral coefficients (MFCCs), for example, begin by transforming an audio signal into the frequency (Fourier) domain, followed by a series of linear and non-linear filtering to yield a transformed sound in which the representation of different frequencies more closely matches human perception. Shown in Figure 1 is a sample audio (top panel) and the corresponding mel-spectrum representation (bottom panel), from which the cepstral coefficients are extracted. The MFCCs have proven effective at a variety of tasks from speaker recognition to automatic music retrieval. Pairing MFCCs extracted from real and synthesized speech along with a variety of pattern recognition techniques from logistic regression to support vector machines, or neural networks, yields a predictor of audio authenticity.

The plethora of audio detection techniques vary in their specific choices of initial transform, feature extraction, and pattern recognition machinery. These three steps can also be incorporated into one end-to-end, learning-based system typically using deep neural networks (see, for example, (Chen et al. 2020)). Below we review the state-of-the-art detection accuracy of these types of predictors as reported in a recent semiannual challenge.

Audio: statistical. At the time of writing, the top ten most populous countries (in millions) are: China (1,439), India (1,380), United States (331), Indonesia (273), Pakistan (220), Brazil (212), Nigeria (206), Bangladesh (164), Russia (145), and Mexico (128). Five of these populations begin with the number 1 (China, India, Bangladesh, Russia, and Mexico), four begin with 2, and one begins with 3. Compared to a uniform distribution in which each digit would appear with an 11.1% frequency, this lopsided distribution of first digits generalizes to all the world populations in which 1 appears with a ≈30% frequency, 2 with a ≈18% frequency, followed by a rapidly decaying frequency for subsequent digits, until 9 appears with only ≈5% frequency. This distribution

($P(d) = \log_{10}(1 + 1/d)$) of first digits d—termed Benford's Law (Fewster 2009)—has been observed in a large number of data sets from world populations to river lengths, house prices, and tax returns.

This exponentially decaying distribution of first digits has also been observed in human speech (Hsu and Berisha 2022). An audio signal is subjected to a standard frequency decomposition from which the first digits of human speech are found to nearly perfectly follow Benford's law. Synthetically-generated speech deviates from this distribution. On an admittedly small dataset of natural and synthetic speech, deviations from Benford's first-digit distribution achieves a 91% classification accuracy. Although it is unlikely to generalize to all forms of synthetic audio, this technique can be a useful tool in any computational detection toolkit.

Audio: evaluation. Starting in 2015, and running every 2 years, the Automatic Speaker Verification Spoofing and Countermeasures challenge (ASVSpoof, https://www.asvspoof.org) invites researchers to develop and submit detection techniques to detect fake, spoofed, or synthesized speech, with the goal of developing automatic speaker verification (ASV) systems. In the most recent 2021 challenge (Yamagishi et al. 2021), the top-performing deep-fake audio classifier achieved an equal error rate of 15.6% (15.6% of synthetic audio clips were misclassified as real, and 15.6% of real audio clips were misclassified as synthetic). This relatively low performance highlights the continued challenge of detecting synthetically generated human speech.

5.2 Image

The computational detection of synthesized images can be partitioned into two basic categories: (1) learning based, in which features that distinguish real from synthetic content are explicitly learned by any of a range of different machine-learning techniques and (2) artifact based, in which a range of low-level (pixel-based) to high-level (semantic-based) features are explicitly designed to distinguish between real and synthetic content. This section provides an overview of a representative example of these computational techniques and discusses their advantages and disadvantages.

Image: learning based. A typical image-forensic classifier, designed to distinguish a real from a synthesized image, takes as its input a single image and outputs a real-valued number corresponding to the likelihood that the input image is synthetically generated. Wang *et al.*'s forensic classifier (S.-Y. Wang et al. 2020), for example, is based on a standard neural-network architecture (ResNet-50), pre-trained on a standard image database (ImageNet), and then refined to classify an image as real or synthesized. The training dataset contains 720,000 training and 4,000 validation images, half of which are real images; the other half are synthesized images created by ProGAN (Karras et al. 2017). To ensure the classifier is resilient to simple image modifications, the training images are augmented by standard manipulations including spatial blurring and JPEG compression. With an average precision[16] greater than 90%, the trained classifier can accurately classify ProGAN synthesized images as well as those from other previously unseen synthesizers. Using a similar computational approach, the forensic classifier of Frank *et al.* (Frank et al. 2020) yields a similarly accurate and generalizable classifier.

It is not uncommon for these types of learning-based approaches to correctly classify images similar to those used in training. What was more impressive was the network's

[16] Precision is defined as the ratio TP / (TP+FP), where true positive (TP) is the rate at which a synthesized image is correctly classified, and false positive (FP) is the rate at which a real image is incorrectly classified as synthesized.

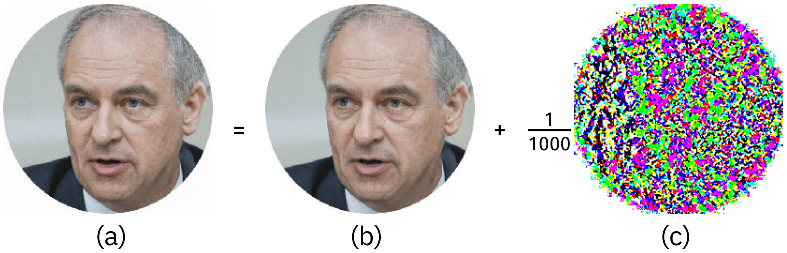

Figure 5: According to his Twitter account, Andrew Walz was running for a congressional seat in Rhode Island. In reality, he was the creation of a 17-year-old high school student. Walz's profile picture (b) was synthesized using StyleGAN2. A state-of-the-art synthetic-media detector (S.-Y. Wang et al. 2020) would have flagged this picture as 87% fake. Adding a perceptually indistinguishable perturbation (c) to this image, however, causes the detector to classify the resulting picture (a) as 99% real (Carlini and Farid 2020).

uncharacteristic ability to generalize to synthetic images that were not part of the training—likely because the classifiers learned an image artifact that was introduced when images were resized within the image-synthesis pipelines. Unfortunately, this means that the intentional or unintentional removal of this single artifact will render the classifier useless (Dong, Kumar, and Liu 2022).

It is well established that machine-learning classifiers are vulnerable to adversarial attacks (Carlini and Wagner 2017): for a given classifier and input image X, it is possible to construct an imperceptible additive perturbation Y so that the image $X + Y$ is misclassified (Figure 5). Most strategies construct the adversarial perturbation Y with respect to the input X through a gradient-descent optimization in which the adversarial, additive pattern is explicitly learned (Carlini and Wagner 2017).

Shortly after their release, the two forensic-image classifiers described above were found not to be adversarially robust: an arbitrary image correctly classified as synthetic can be imperceptibly modified (Figure 5) to be misclassified as real (Carlini and Farid 2020). These adversarial attacks are carried out assuming the attacker has (white box) or does not have (black box) full access to the image classifier's parameters. White-box attacks reduce the area under the ROC curve (AUC) from 0.95 to ≈0.1, compared to an AUC of 0.5 for a chance-performing classifier. Black-box attacks reduce the AUC to ≈0.2. In both cases, the adversarial attacks yield a classifier that is most likely to incorrectly classify a real image as synthetic, and a synthetic image as real.

The advantage of learning-based techniques is that they can uncover subtle and non-obvious discriminatory artifacts. However, these techniques often require a significant amount of training data, can struggle to generalize to new synthesis techniques, and are vulnerable to intentional and unintentional attacks.

Image: artifact based. It is often said that "the eyes are a window to the soul," but it would be more accurate to say "the eyes are a reflection of the individual's world." Because the eye's cornea is a partially reflective surface, it reflects back to the camera a glimpse into one's surroundings. Corneal reflections have been used to reconstruct a view of a person's physical surroundings (Nishino and Nayar 2004) and to detect physical inconsistencies in synthesized faces. Because facial synthesis is not necessarily aware of the axial symmetry of the face, the corneal reflection and pupil shape of synthetically generated faces are often—implausibly—different between the left and right eyes (Guo et al. 2021; Hu, Li, and Lyu 2021). Such inconsistencies constitute a simple, yet robust, indicator of a synthetically-generated face.

Ocular asymmetries caused by the lack of explicit semantic knowledge in image-synthesis pipelines has also been exploited in the analysis of facial feature configurations (Yang et al. 2019). Particularly in earlier incarnations of synthetically generated faces, the location of facial features (eyes, tip of nose, corners of mouth, chin, etc.) on synthetically generated faces was found to be different than on real faces.

While these types of semantic artifacts are more resilient to adversarial attacks than learning-based approaches, facial synthesis will continue to evolve and incorporate this type of semantic information. These detection techniques will thus become less effective over time and new techniques will need to be developed that exploit different artifacts. This is the inevitable outcome of the cat-and-mouse game between synthesis and detection.

5.3 Video

The computational detection of synthesized videos can be partitioned into three basic categories: (1) learning based, in which features that distinguish real from synthetic content are explicitly learned by a range of different machine-learning techniques; (2) artifact based, in which a range of low-level (pixel-based) to high-level (semantic-based) features are explicitly designed to distinguish between real and synthetic content; and (3) identity based, in which biometric-style features are used to determine whether the person depicted in a video is who it purports to be. This section provides an overview of a representative sample of these computational techniques, and discusses their advantages and disadvantages.

Video: learning based. This category of computational forensic techniques leverages modern machine learning to extract artifacts that arise from the synthesis process that (typically) are not visually salient. In these supervised-learning approaches, a system (typically, a neural network) is trained to distinguish real from synthesized content. Depending on the initial representation, network architecture, and objective function being optimized, these networks can learn different features to distinguish the real from the synthesized.

One such approach (Zhou et al. 2017) trains two separate neural networks to distinguish real from synthesized faces on individual video frames, and to distinguish arbitrary image patches taken from the same or different image. The first network is designed to learn artifacts that emerge directly from the synthesis process, while the second is designed to learn low-level properties shared by parts of an image recorded by the same camera (e.g., sensor noise and compression artifacts). These two predictors are then combined into a single classifier. This technique is somewhat effective at detecting synthetic faces, able to detect synthetic faces 70% of the time, but with a 5% false-alarm rate (incorrectly classifying a real face as synthetic).

Different network architectures lead to different features used to distinguish the real from the synthesized. For example, a network designed to focus on mesoscopic (intermediate-scale) features (Afchar et al. 2018) rather than low-level (e.g., sensor noise) or high-level (e.g., facial features) characteristics yields an improved detection accuracy of 90% with a 5% false-alarm rate.

A face-swap or lip-sync deep fake starts by synthesizing, one frame at a time, a face or mouth to digitally replace another face or mouth. The synthesized part of the face is then blended into the original video frame, which requires some post-processing image blending to remove any visible artifacts at the facial boundary. Because this blending introduces a slight artifact along the facial boundary, a neural network can be trained to detect artifacts at the boundary of a stitched face (L. Li et al. 2019). Although not applicable to puppet-master deep fakes, this approach yields improved detection

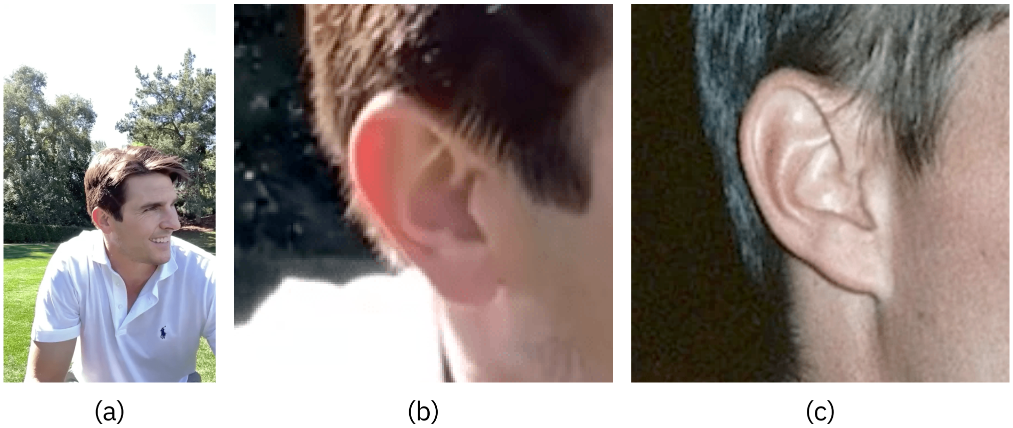

Figure 6: Ear biometrics for deep-fake detection: (a) one frame of face-swap, deep-fake video of Tom Cruise; (b) a magnified view of deep-fake Cruise's ear; and (c) the real Cruise's ear, which has a distinctively different aural pattern (Agarwal and Farid 2021).

accuracy to ≈95% with a 5% false-alarm rate.

The advantage of these and related learning-based approaches is they are able to learn detailed and subtle artifacts resulting from video synthesis. The disadvantage is they typically require large amounts of labelled training data that may not always be easy to acquire. These techniques often struggle to generalize to new synthetic content that is not explicitly part of the training data set, and can—as was the case with learning-based synthetic-image detection—be vulnerable to adversarial attacks, including simple laundering attacks in which the synthetic media is trans-coded and resized (Barni, Stamm, and Tondi 2018).

Video: artifact based. This category of computational forensic techniques is based on the observation that even if a deep-fake video is visually compelling, it may contain spatial and temporal physiological artifacts that are not always perceptually obvious.

One of the first artifact-based techniques was based on the clever observation that earlier incarnations of face-swap deep fakes contained unnaturally few eye blinks (Li, Chang, and Lyu 2018). Creating a face-swap deep fake typically requires many photos of the identity in a range of head poses and expressions. Yet since most photos of a person show them with their eyes open, synthesis techniques struggled to accurately capture eye blinks. The average eye blink rate in natural interview-style videos is 34 blinks/min, while the rate in deep-fake videos is only 3.4 blinks/min; blink rate can thus be used to detect unnaturally few eye blinks in videos as short as 30 seconds. As is often the case in the cat-and-mouse game between synthesis and detection, face-swap deep fakes were quick to adapt: today's higher-quality, deep-fake videos do not contain this artifact.

Because a face-swap deep fake only replaces the face from the eyebrows to the chin and cheek to cheek, the ears in a synthesized video belong to the imposter rather than the co-opted identity. Aural features, which have previously been used as a biometric, can therefore be used to determine if they match the purported identity (Agarwal and Farid 2021). For example, shown in Figure 6(a) is a single frame of deep-fake Tom Cruise. Shown in panels (b) and (c) are views of deep-fake Cruise's ear and the real Cruise's ear, in which clear differences are apparent from the detached vs. attached earlobe, and differences in the overall shape and structures.

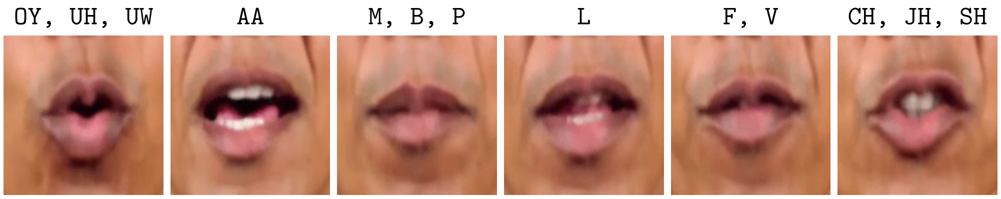

Figure 7: Six example visemes and their corresponding phonemes. The M, B, P phonemes, for example, correspond to the the first syllable of “mother,” “brother,” or “parent,” in which the lips are pressed tightly together, leading to the viseme shown (Agarwal, Farid, Fried, et al. 2020).

Although both the aural and eye-blink techniques are applicable to face-swap deep fakes, they are unsuitable for examining lip-sync deep fakes since only the mouth region and audio track are manipulated. However, since the ear and ear canal are related to movement of the lower jaw, which moves in response to face and head movements, onset of vocalization, swallowing, and laughing, dynamic aural features can be used to determine if the dynamics of aural motion are decoupled from the mouth and jaw motion, as would be likely in a lip-sync deep fake (Agarwal and Farid 2021). Because static and dynamic aural features are relatively complex, it is likely to be harder to circumvent these techniques, as was the case with eye blinking. Fully automatic localization and tracking of the human ear remains challenging, so this technique is best performed with manual assistance.

A third physiologically based artifact is based on the shape of the mouth (viseme) needed to utter specific sounds (phonemes). For example, your lips have to close to say a word that begins with the M, B, P phoneme—*mother, brother, parent* (Figure 7). Similarly, your upper teeth close slightly on your inwardly curled lower lip to enunciate *victor, favor*—the V, F phoneme. Because lip-sync and text-to-voice deep fakes do not explicitly model the appearance of the mouth at the phoneme/viseme level (see Section **Creating**), these phoneme–viseme pairings are occasionally violated in these types of deep fakes (Agarwal, Farid, Fried, et al. 2020). While it is relatively straightforward to extract spoken phonemes using standard voice-to-text conversion, precise and automatic measurement of the mouth shape (viseme) remains challenging. Thus this technique also works best with manual assistance.

Video: identity based. This category of computational forensic techniques is based on the assumption that as a person speaks, they have distinct facial expressions and movements that can be used to distinguish them from others, impersonators, and deep fakes (Agarwal et al. 2019; Agarwal, Farid, El-Gaaly, et al. 2020; Agarwal et al. 2021; Cozzolino et al. 2021). A person may, for example, raise her eyebrows when she is emphasizing a certain word, tilt her head upward when she is smiling, or tilt her head downward when she furrows her brow.

In a face-swap or puppet-master deep fake, an impersonator is driving the facial expressions and head movements. It is thus difficult for the impersonator to capture all of the idiosyncratic mannerisms of the co-opted identity. Similarly, although the head movements and facial expressions in a lip-sync deep fake are authentic, this manipulation de-synchronizes the mouth from the rest of the head. If, for example, a person tends to tilt her head upwards when she smiles, this mannerism may be lost in a lip-sync deep fake.

Underlying this category of techniques is a system that learns distinct facial expressions and movements from authentic videos of an individual, and then compares these learned mannerisms to the video in question. The advantage of this approach is that the mannerisms are measured over a 5–10 second window consisting of 150–300 video frames in a standard video format. Because deep-fake videos are synthesized one (or maybe a few) frames at a time, the synthesis engine at time T is (for now at least) unaware of what is happening at time $T + 150$ and beyond. As a result, current synthesis engines will struggle to easily circumvent this forensic approach. Compared to learning-based approaches, the measurement of facial expressions and movement is robust even for relatively low video resolution and quality. And, by attacking the fundamental flaw of a deep fake—the person depicted is not who it purports to be—these techniques are applicable to all forms of deep-fake videos. The disadvantage of these approaches is that a model must be constructed for each individual. This may be practical when it comes to protecting a few world leaders from deep fakes—for which hours of video can typically be found online—but is otherwise impractical. The other disadvantage is that the learned mannerisms are somewhat context dependent: when a world leader is giving a public address, for example, she may be more formal than when she is giving an unscripted interview, so specific mannerisms may not generalize across different contexts.

Video: evaluation. Thanks to a number of industry- and academic-led efforts, several datasets are available to evaluate deep-fake video detection, including:

1. FaceForensics++ (Rossler et al. 2019) consisting of 1,000 YouTube videos of 1,000 different people, mostly news anchors and video bloggers, from which four types of deep-fake videos are created: faceswap[17], FaceSwap[18], Face2Face (Justus Thies et al. 2016), and Neural Textures (Thies, Zollhöfer, and Nießner 2019).

2. DeepFake Detection dataset (Dufour and Gully 2019) consisting of 363 real and 3,068 face-swap deep fakes of 28 paid and consenting actors. Each individual was made to perform tasks including walking, hugging, talking, etc. in different expressions such as happy, angry, neutral, or disgust.

3. DeepFake Detection Challenge (DFDC) (Dolhansky et al. 2019) consisting of 1,131 real and 4,113 face-swap deep fakes of 66 paid and consenting actors.

4. Celeb-DF (Y. Li et al. 2019) dataset consisting of 5,639 face-swap deep fakes generated from 590 YouTube videos of 61 celebrities speaking in different settings including interviews, TV-shows, and award ceremonies.

In the 2019–2020 Deepfake Detection Challenge (Ferrer 2020), 2,116 teams competed for $1 million in prizes. Teams were provided 23,654 real videos and 104,500 deep-fake videos created from the provided real videos. The top-performing learning-based detector achieved a detection accuracy of only 65% on a set of 4,000 holdout videos, half of which were real and half of which were deep fakes (i.e., chance performance is 50%). Compared to reported detection accuracies greater than 90% in many published studies, these results reveal that fully automatic detection of deep fakes remains a challenging problem in practice.

Video: active. During the COVID-19 pandemic, video calls frequently replaced in-person meetings and phone calls. Although not yet perfected, deep fakes can be synthesized in real time and piped through a virtual camera (e.g., https://github.com/alievk/avatarify-python and https://github.com/iperov/DeepFacelive). Despite not yet being perfected, this technology is good enough that the mayor of Berlin, Franziska Giffey, spent 15 minutes on a video conference call with a deep-fake version of the mayor of Kyiv, Vitali Klitschko (Oltermann 2022). It will therefore become increasingly difficult to distinguish a real person from a synthesized person on a video call. The

[17] https://github.com/deepfakes/faceswap
[18] https://github.com/MarekKowalski/FaceSwap

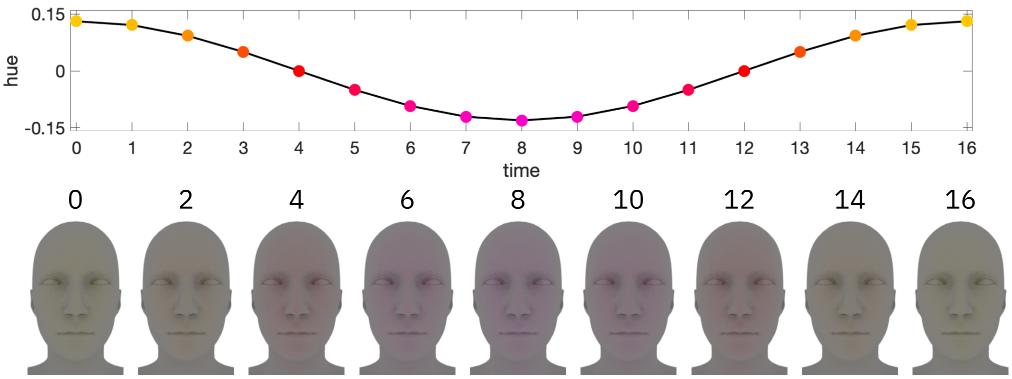

Figure 8: Shown in the top panel is a visualization of the dynamic change in the hue of a uniform-colored area light source (simulating a computer screen). Shown below are nine renderings of a 3D model illuminated with a different light-source hue at nine distinct moments in time (Gerstner and Farid 2022).

creation of real-time deep fakes poses unique threats because of the general sense of trust surrounding a live video call, and the challenge of detecting deep fakes in real time, during a call.

While some of the computational techniques described above might help detect real-time deep-fake videos, a class of promising approaches (Shang and Wu 2020; Gerstner and Farid 2022) exploits the unique physical constraints of a live video call: call participants are in front of a light source (the computer display) that can be actively adjusted in real time to induce specific lighting patterns on a user's face. Shown in Figure 8, for example, is a simulation of the change in facial appearance of a person sitting in front of a display that is slowly changing color (hue) over time. The consistency of a call participant's appearance under this lighting can then be used to verify their liveness and physical presence in front of the camera. This approach is effective because either the synthesis engine simply fails to transfer the active illumination or there is a temporal delay in transferring the active illumination.

This illumination pattern can be induced by a call participant sharing her screen and displaying a temporally varying pattern, or directly integrated into the video-call client. In either case, no specialized imaging or lighting hardware is required.

6 Detecting: Provenance

The perceptual and computational techniques described in the two previous sections are based on a somewhat unreasonable model: billions of daily users upload hundreds of hours of video to YouTube and TikTok every minute, and make hundreds of millions of daily tweets and Facebook posts. Billions of daily users are then largely left to their own devices to separate fact from fiction, while trying to make sense of a global pandemic, national elections, human rights atrocities, and war. It is no wonder we are awash in a sea of misinformation, conspiracies, and lies.

An alternate model would be to shift the responsibility of verification away from the content recipient and onto the content creator and publisher—particularly for multimedia content (audio, image, and video). In this flipped model, the content creator and publisher are responsible for authenticating content from the point of creation and edits, through editorial, publishing, and eventual delivery (the so-called glass-to-glass pipeline of content creation to delivery).

Three separate, but now unified, organizations have been established to formalize this new model of content creation and delivery. The Content Authenticity Initiative (CAI)[19] considers the initial content creation, Project Origin[20] considers content publication, and a third umbrella group, the Coalition for Content Provenance and Authenticity (C2PA)[21] has unified these efforts through the creation of an end-to-end technical specification.

The CAI was founded in 2019 by Adobe, *The New York Times*, and Twitter, and now includes over 700 members across the technology sector, publishing, NGOs, device manufacturers, and academia. Under the CAI, content authentication begins at the point of recording where a specialized camera app cryptographically hashes (Preneel 2010) the recorded content (audio, image, or video) and (optionally) metadata including creator identity, date/time, and geo-location. Sensitive to the need to balance content authenticity with privacy and security of, for example, photojournalists in high-risk areas, the CAI protocol allows creators to select and preserve attribution or remain anonymous. The extracted, tamper-evident cryptographic hash is stored alongside any other recorded metadata and stored on a centralized trust list (Linn 2000). Although the CAI does not use a blockchain (an immutable, decentralized ledger) to store the cryptographic hashes, the protocol can be adapted to use a blockchain as in, for example, https://www.starlinglab.org/78days.

Founded in 2020 by the BBC, CBC/Radio Canada, Microsoft, and *The New York Times*, Project Origin picks up where the CAI leaves off. Prior to publication, an Origin-compliant publisher attaches to a piece of media, metadata consisting of the date, source, and description. A cryptographically signed hash of the media and metadata is then stored on a trust list. As a consumer views the media, the browser or media player compares the extracted signature against the trust list to verify the content provenance. Whereas the CAI is focused on verifying the integrity of the underlying media, Origin is focused on when, where, and by whom a piece of media was published.

In 2021, the C2PA unified these two efforts by creating an end-to-end technical specification that encompasses the entire publishing pipeline. Released in January 2022, the open, technical standard (v.1.0) gives creators, publishers, and consumers a way to understand the authenticity and provenance of multi-media content. As the C2PA continues to focus on creating open standards for provenance and authenticity, the CAI and Project Origin focus on advocacy and adoption among their respective constituencies.

A provenance-based approach need not only focus on authentic content. Synthesis pipelines can also incorporate the same standard to make downstream identification of synthetic media easier. Because the CAI-based provenance information is embedded in the content's metadata, this critical information can easily be stripped from the media. An alternate, and more robust, solution would be to embed a unique identifying digital watermark directly into the underlying content (Yu et al. 2022). This watermark can be baked into the synthesis pipeline so that content can be traced back to a specific synthesis engine.

The advantage of these provenance-based approaches is they can contend with internet-scale content creation and consumption. The disadvantage is they will require broad adoption by creators, publishers, and social media giants. It seems likely, however, that some version of a provenance-based solution, alongside the detection schemes described in the previous sections, will be needed help re-establish trust in

[19] https://contentauthenticity.org
[20] https://www.originproject.info
[21] https://c2pa.org

online content.

7 Closing Thoughts

Given the rate at which synthetic media has been advancing, this article will likely need to be updated in a few years. The underlying synthesis techniques, use and misuse cases, and ethical and legal landscape will most likely continue to quickly evolve as we enter a world in which reality will become increasingly more difficult to ascertain. While there are creative applications to deep-fake technologies, we cannot ignore the threats posed by the widespread access to deep-fake synthesis technology and content. I argue that today there are three primary threats, two of which are being realized, and the third is likely to be more fully realized in the near future. First, non-consensual sexual imagery is being weaponized against (primarily) women as many apps and web services operationalize deep-fake technology with the explicit purpose of creating this material. As this technology improves, anyone with even a minimal online footprint could be a victim. Second, the existence of deep fakes enables the liar's dividend to be wielded as a powerful weapon to dismiss inconvenient facts and realities. Third, although still nascent, deep fakes are likely to play a larger role in small- to large-scale fraud and disinformation campaigns in the future.

It is more difficult to predict how the courts, and society, will contend with the myriad of complex legal and ethical questions arising from deep-fake and related technologies. As we wrestle with these issues, academic researchers—the primary driving force of deep-fake technology—should consider not just *how* to do something, but *if* something should be done—and if so, how to put proper safeguards in place before unleashing their technologies on the world.

Recent history can provide some guidance. After decades of research in facial recognition, we are now contending with troubling consequences from demonstrated racial and gender bias (Buolamwini and Gebru 2018), as well as privacy violations in applications like Clearview AI (Roussi 2020). What started out as a mostly academic research effort that gave little thought to ethics or consequences has led to the broad deployment of biased and privacy-invading technologies without proper safeguards. The time to consider potential harms is before, not after, technology is developed and deployed.

While deep fakes are the flavor of the year (or decade), they are not the first—nor will they be the last—form of digital manipulation. They are also not currently the most dangerous example of the weaponization of the internet, given the long litany of online harms including child exploitation, extremism and terrorism, the sale of illegal and deadly drugs, small- to large-scale fraud, invasions of privacy, and the spread of deadly and dangerous disinformation campaigns designed to sow civil unrest and disrupt elections. Deep fakes should therefore be considered within a broader range of online harms that are threatening individuals, societies, and democracies.

References

Acuna, Daniel E, Paul S Brookes, and Konrad P Kording. 2018. "Bioscience-scale automated detection of figure element reuse." *BioRxiv,* 269415.

Afchar, Darius, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. "MesoNet: A compact facial video forgery detection network." In *IEEE International Workshop on Information Forensics and Security.* https://doi.org/10.1109/wifs.2018.8630761.

Agarwal, Shruti, and Hany Farid. 2021. "Detecting deep-fake videos from aural and oral dynamics." In *CVPR Workshop on Media Forensics,* 981–89. https://doi.org/10.1109/cvprw53098.2021.00109.

Agarwal, Shruti, Hany Farid, Ohad Fried, and Maneesh Agrawala. 2020. "Detecting deep-fake videos from phoneme-viseme mismatches." In *CVPR Workshop on Media Forensics,* 660–61. https://doi.org/10.1109/cvprw50498.2020.00338.

Agarwal, Shruti, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. 2020. "Detecting deep-fake videos from appearance and behavior." In *International Workshop on Information Forensics and Security,* 1–6. https://doi.org/10.1109/wifs49906.2020.9360904.

Agarwal, Shruti, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 2019. "Protecting world leaders against deep fakes." In *CVPR Workshop on Media Forensics,* vol. 1.

Agarwal, Shruti, Liwen Hu, Evonne Ng, Trevor Darrell, Hao Li, and Anna Rohrbach. 2021. *Watch those words: Video falsification detection using word-conditioned facial motion.* arXiv:2112.10936.

AlBadawy, Ehab A, Siwei Lyu, and Hany Farid. 2019. "Detecting AI-synthesized speech using bispectral analysis." In *CVPR Workshop on Media Forensics,* 104–9.

Barni, Mauro, Matthew C Stamm, and Benedetta Tondi. 2018. "Adversarial multimedia forensics: Overview and challenges ahead." In *European Signal Processing Conference,* 962–66. IEEE, September. https://doi.org/10.23919/eusipco.2018.8553305.

Bau, David, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. 2021. *Paint by word.* arXiv:2103.10951.

Bik, Elisabeth M, Arturo Casadevall, and Ferric C Fang. 2016. "The prevalence of inappropriate image duplication in biomedical research publications." *MBio* 7, no. 3 (July): e00809–16. https://doi.org/10.1128/mbio.00809-16.

Bond, Shannon. 2022. *That smiling LinkedIn profile face might be a computer-generated fake.* https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles.

Bregler, Christoph, Michele Covell, and Malcolm Slaney. 1997. "Video rewrite: Driving visual speech with audio." In *24th Annual Conference on Computer Graphics and Interactive Techniques,* 353–60. https://doi.org/10.1145/258734.258880.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. "Language models are few-shot learners." *Advances in Neural Information Processing Systems* 33:1877–901.

Brugioni, Dino A. 1999. *Photo fakery: the history and techniques of photographic deception and manipulation.* Potomac Books Incorporated.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability and transparency,* 77–91.

Carlini, Nicholas, and Hany Farid. 2020. "Evading deepfake-image detectors with white- and black-box attacks." In *CVPR Workshop on Media Forensics,* 658–59. https://doi.org/10.1109/cvprw50498.2020.00337.

Carlini, Nicholas, and David Wagner. 2017. "Adversarial examples are not easily detected: Bypassing ten detection methods." In *10th ACM Workshop on Artificial Intelligence and Security,* 3–14. https://doi.org/10.1145/3128572.3140444.

Chan, Caroline, Shiry Ginosar, Tinghui Zhou, and Alexei Efros. 2019. "Everybody Dance Now." In *2019 IEEE/CVF International Conference on Computer Vision (ICCV).* IEEE, October. https://doi.org/10.1109/iccv.2019.00603.

Chen, Tianxiang, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury. 2020. "Generalization of audio deepfake detection." In *Odyssey 2020: The Speaker and Language Recognition Workshop,* 132–37. ISCA, November. https://doi.org/10.21437/odyssey.2020-19.

Chesney, Bobby, and Danielle Citron. 2019. "Deep fakes: A looming challenge for privacy, democracy, and national security." *California Law Review* 107:1753. https://doi.org/10.2139/ssrn.3213954.

Ciftci, Umur Aybars, and Ilke Demir. 2019. *FakeCatcher: Detection of synthetic portrait videos using biological signals.* arXiv:1901.02212.

Citron, Danielle. 2022. *The fight for privacy.* W.W. Norton.

Citron, Danielle Keats, and Mary Anne Franks. 2014. "Criminalizing revenge porn." *Wake Forest L. Rev.* 49:345.

Cozzolino, Davide, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. 2021. "ID-reveal: Identity-aware deepfake video detection." In *International Conference on Computer Vision and Pattern Recognition,* 15108–17. https://doi.org/10.1109/iccv48922.2021.01483.

Dolhansky, Brian, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. 2019. *The deepfake detection challenge (DFDC) preview dataset.* arXiv:1910.08854.

Dong, Chengdong, Ajay Kumar, and Eryun Liu. 2022. "Think Twice Before Detecting GAN-Generated Fake Images From Their Spectral Domain Imprints." In *International Conference on Computer Vision and Pattern Recognition,* 7865–74.

Dubiel, Mateusz, Martin Halvey, Pilar Oplustil Gallegos, and Simon King. 2020. "Persuasive synthetic speech: Voice perception and user behaviour." In *Conference on Conversational User Interfaces,* 1–9. https://doi.org/10.1145/3405755.3406120.

Dufour, Nicholas, and Andrew Gully. 2019. *Contributing data to deepfake detection research.* https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

Farid, Hany. 2006. "Exposing digital forgeries in scientific images." In *Proceedings of the 8th workshop on Multimedia and Security,* 29–36. https://doi.org/10.1145/1161366.1161374.

———. 2021. "On algorithmic amplification." *Inference* 6 (1). https://doi.org/10.37282/991819.21.10.

Ferrer, Cristian Canton. 2020. *Deepfake detection challenge results: An open initiative to advance AI.* https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai.

Fewster, Rachel M. 2009. “A simple explanation of Benford’s Law.” *The American Statistician* 63 (1): 26–32. https://doi.org/10.1198/tast.2009.0005.

Frank, Joel, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. *Leveraging frequency analysis for deep fake image recognition.* arXiv:2003.08685.

Frank, Joel, and Lea Schönherr. 2021. *WaveFake: A data set to facilitate audio deepfake detection.* arXiv:2111.02813.

Fried, Ohad, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. 2019. “Text-based editing of talking-head video.” *ACM Transactions on Graphics* 38, no. 4 (August): 1–14. https://doi.org/10.1145/3306346.3323028.

Garfinkle, Adam. 2020. “Disinformed.” *Inference* 5 (3). https://doi.org/10.37282/991819.20.35.

Gerstner, Candice, and Hany Farid. 2022. “Detecting real-time deep-fake videos using active illumination.” In *CVPR workshop on Media Forensics.* https://doi.org/10.1109/cvprw56347.2022.00015.

Gibiansky, Andrew, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. “Deep voice 2: Multi-speaker neural text-to-speech.” *Advances in Neural Information Processing Systems* 30.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Wade-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. “Generative adversarial nets,” 2672–80. https://doi.org/10.1145/3422622.

Groh, Matthew, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. “Deepfake detection by human crowds, machines, and machine-informed crowds.” *Proceedings of the National Academy of Sciences* 119 (1). https://doi.org/10.1073/pnas.2110013119.

Groh, Matthew, Aruna Sankaranarayanan, and Rosalind Picard. 2022. *Human detection of political deepfakes across transcripts, audio, and video.* arXiv:2202.12883.

Guo, Hui, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. 2021. *Eyes tell all: Irregular pupil shapes reveal GAN-generated faces.* arXiv:2109.00162.

Hancock, Jeffrey T, and Jeremy N Bailenson. 2021. “The social impact of deepfakes.” *Cyberpsychology, Behavior, and Social Networking* 24 (3): 149–52. https://doi.org/10.1089/cyber.2021.29208.jth.

Holmes, Olivia, Martin S Banks, and Hany Farid. 2016. “Assessing and improving the identification of computer-generated portraits.” *ACM Transactions on Applied Perception* 13, no. 2 (March): 1–12. https://doi.org/10.1145/2871714.

Hsu, Leo, and Visar Berisha. 2022. *Does human speech follow Benford’s law?* arXiv:2203.13352.

Hu, Shu, Yuezun Li, and Siwei Lyu. 2021. “Exposing GAN-generated faces using inconsistent corneal specular highlights.” In *International Conference on Acoustics, Speech and Signal Processing,* 2500–2504. https://doi.org/10.1109/icassp39728.2021.9414582.

Hulzebosch, Nils, Sarah Ibrahimi, and Marcel Worring. 2020. “Detecting CNN-generated facial images in real-world scenarios.” In *CVPR workshop on Media Forensics,* 642–43. https://doi.org/10.1109/cvprw50498.2020.00329.

Kadhim, Hashiam, and Joseph Palermo. 2019. *RealTalk: We replicated a real person’s voice with AI.* https://medium.com/dessa-news/real-talk-speech-synthesis-5dd0897eef7f.

Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. *Progressive growing of GANs for improved quality, stability, and variation.* arXiv:1710.10196.

Karras, Tero, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. “Alias-free generative adversarial networks.” In *Neural Information Processing Systems.*

Karras, Tero, Samuli Laine, and Timo Aila. 2019. “A style-based generator architecture for generative adversarial networks.” In *International Conference on Computer Vision and Pattern Recognition,* 4401–10. IEEE, June. https://doi.org/10.1109/cvpr.2019.00453.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. “Analyzing and improving the image quality of StyleGAN.” In *International Conference on Computer Vision and Pattern Recognition,* 8110–19. https://doi.org/10.1109/cvpr42600.2020.00813.

Kim, H., P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. 2018. “Deep video portraits.” *ACM Transactions on Graphics* 37, no. 4 (August): 1–14. https://doi.org/10.1145/3197517.3201283.

Lago, Federica, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. 2021. *More real than real: A study on human visual perception of synthetic faces.* arXiv:2106.07226.

Li, Lingzhi, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2019. *Face X-ray for more general face forgery detection.* arXiv:1912.13458.

Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. 2018. “In ictu oculi: Exposing AI created fake videos by detecting eye blinking.” In *International Workshop on Information Forensics and Security,* 1–7. IEEE, December. https://doi.org/10.1109/wifs.2018.8630787.

Li, Yuezun, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2019. *Celeb-DF: A new dataset for deepfake forensics.* arXiv:1909.12962.

Linn, John. 2000. “Trust models and management in public-key infrastructures.” *RSA Laboratories* 12.

Lorenzo-Trueba, Jaime, Fuming Fang, Xin Wang, Isao Echizen, Junichi Yamagishi, and Tomi Kinnunen. 2018. “Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama’s voice using GAN, WaveNet and low-quality found data.” In *The Speaker and Language Recognition Workshop (Odyssey).* Les Sables d’Olonne, France. https://doi.org/10.21437/odyssey.2018-34.

Müller, Nicolas M, Karla Markert, and Konstantin Böttinger. 2021. *Human perception of audio deepfakes.* arXiv:2107.09667.

Nagano, Koki, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. 2018. “PaGAN: Real-time avatars using dynamic textures.” *ACM Transaction on Graphics,* https://doi.org/10.1145/3272127.3275075.

Newall, Mallory. 2020. *More Than 1 in 3 Americans believe a 'deep state' is working to undermine Trump.* Ipsos.

Nightingale, Sophie, and Hany Farid. 2020. *Examining the global spread of COVID-19 misinformation.* arXiv:2006.08830.

———. 2022. “AI-synthesized faces are indistinguishable from real faces and more trustworthy.” *Proceedings of the National Academy of Sciences* 119 (8). https://doi.org/10.1073/pnas.2120481119.

Nirkin, Yuval, Yosi Keller, and Tal Hassner. 2019. “FSGAN: Subject agnostic face swapping and reenactment.” In *International Conference on Computer Vision and Pattern Recognition,* 7184–93. https://doi.org/10.1109/iccv.2019.00728.

Nishino, Ko, and Shree K Nayar. 2004. “The world in an eye.” In *International Conference on Computer Vision and Pattern Recognition,* vol. 1.

Öhman, Carl, and Luciano Floridi. 2017. “The political economy of death in the age of information: A critical approach to the digital afterlife industry.” *Minds and Machines* 27 (4): 639–62. https://doi.org/10.1007/s11023-017-9445-2.

———. 2018. “An ethical framework for the digital afterlife industry.” *Nature Human Behaviour* 2 (5): 318–20. https://doi.org/10.1038/s41562-018-0335-2.

Oltermann, Philip. 2022. *European politicians duped into deepfake video calls with mayor of Kyiv.* https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko.

Owen, Alex. 1994. “‘Borderland forms’: Arthur Conan Doyle, Albion's daughters, and the politics of the Cottingley fairies.” In *History Workshop,* 48–85. 38.

Ping, Wei, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. *Deep voice 3: 2000-speaker neural text-to-speech.* arXiv:1710.07654.

Preneel, Bart. 2010. “The first 30 years of cryptographic hash functions and the NIST SHA-3 competition.” In *Cryptographers' track at the RSA conference,* 1–14. Springer.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. “Learning transferable visual models from natural language supervision.” In *International Conference on Machine Learning,* 8748–63. PMLR.

Razavi, Ali, Aaron Van den Oord, and Oriol Vinyals. 2019. “Generating diverse high-fidelity images with VQ-VAE-2.” *Advances in Neural Information Processing Systems* 32.

Rosner, Helen. 2021. *The ethics of a deepfake Anthony Bourdain voice.* https://www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice.

Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. “FaceForensics++: Learning to Detect Manipulated Facial Images.” In *2019 IEEE/CVF International Conference on Computer Vision (ICCV).* IEEE, October. https://doi.org/10.1109/iccv.2019.00009.

Roussi, Antoaneta. 2020. “Resisting the rise of facial recognition.” *Nature* 587 (7834): 350–54. https://doi.org/10.1038/d41586-020-03188-2.

Shang, Jiacheng, and Jie Wu. 2020. "Protecting real-time video chat against fake facial videos generated by face reenactment." In *International Conference on Distributed Computing Systems,* 689–99. https://doi.org/10.1109/icdcs47774.2020.00082.

Shi, Zhaofeng. 2021. *A Survey on Audio Synthesis and Audio-Visual Multimodal Processing.* arXiv:2108.00443.

Siarohin, Aliaksandr, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. "First order motion model for image animation." *Advances in Neural Information Processing Systems* 32.

Sisman, Berrak, Junichi Yamagishi, Simon King, and Haizhou Li. 2020. "An overview of voice conversion and its challenges: From statistical modeling to deep learning." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29:132–57. https://doi.org/10.1109/taslp.2020.3038524.

Sofer, Carmel, Ron Dotsch, Daniel HJ Wigboldus, and Alexander Todorov. 2015. "What is typical is good: The influence of face typicality on perceived trustworthiness." *Psychological Science* 26 (1): 39–47. https://doi.org/10.1177/0956797614554955.

Spencer, Saranac Hale. 2020. *Viral video manipulates Pelosi's words.* Factcheck.

Sundar, S Shyam, Maria D Molina, and Eugene Cho. 2021. "Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps?" *Journal of Computer-Mediated Communication* 26 (6): 301–19. https://doi.org/10.1093/jcmc/zmab010.

Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. "Synthesizing Obama: Learning lip sync from audio." *ACM Transactions on Graphics,* https://doi.org/10.1145/3072959.3073640.

Tan, Xu, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. *A survey on neural speech synthesis.* arXiv:2106.15561.

Taylor, Paul. 2009. *Text-to-Speech Synthesis.* Cambridge University Press, February. https://doi.org/10.1017/cbo9780511816338.

Ternovski, John, Joshua Kalla, and Peter Aronow. 2022. "The negative consequences of informing voters about deepfakes: Evidence from two survey experiments." *Journal of Online Trust and Safety* 1 (2). https://doi.org/10.54501/jots.v1i2.28.

Thies, J., M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. 2018. "HeadOn: Real-time reenactment of human portrait videos." *ACM Transactions on Graphics.*

Thies, Justus, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. "Face2face: Real-time face capture and reenactment of RGB videos." In *International Conference on Computer Vision and Pattern Recognition,* 2387–95. IEEE, June. https://doi.org/10.1109/cvpr.2016.262.

Thies, Justus, Michael Zollhöfer, and Matthias Nießner. 2019. "Deferred neural rendering: Image synthesis using neural textures: image synthesis using neural textures." *ACM Transactions on Graphics* 38, no. 4 (August): 1–12. https://doi.org/10.1145/3306346.3323035.

Van den Oord, Aaron, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. "Conditional image generation with pixelCNN decoders." *Advances in Neural Information Processing Systems* 29.

Van Den Oord, Aäron, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. 2016. "WaveNet: A generative model for raw audio." *SSW* 125:2.

Wang, Liansheng, Lianyu Zhou, Wenxian Yang, and Rongshan Yu. 2022. "Deepfakes: A new threat to image fabrication in scientific publications?" *Patterns* 3 (5): 100509. https://doi.org/10.1016/j.patter.2022.100509.

Wang, Sheng-Yu, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. "CNN-generated images are surprisingly easy to spot... for now." In *International Conference on Computer Vision and Pattern Recognition,* 8695–704. https://doi.org/10.1109/cvpr42600.2020.00872.

Wang, Xin, Jaime Lorenzo-Trueba, Shinji Takaki, Lauri Juvela, and Junichi Yamagishi. 2018. "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis." In *International Conference on Acoustics, Speech and Signal Processing,* 4804–8. IEEE, April. https://doi.org/10.1109/icassp.2018.8461452.

Wang, Xin, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, et al. 2020. "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech." *Computer Speech &amp; Language* 64 (November): 101114. https://doi.org/10.1016/j.csl.2020.101114.

Yamagishi, Junichi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. 2021. *ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection.* arXiv:2109.00537.

Yang, Xin, Yuezun Li, and Siwei Lyu. 2019. "Exposing deep fakes using inconsistent head poses." In *IEEE International Conference on Acoustics, Speech and Signal Processing,* 8261–65. IEEE, May. https://doi.org/10.1109/icassp.2019.8683164.

Yang, Xin, Yuezun Li, Honggang Qi, and Siwei Lyu. 2019. "Exposing GAN-synthesized faces using landmark locations." In *ACM Workshop on Information Hiding and Multimedia Security,* 113–18. ACM, July. https://doi.org/10.1145/3335203.3335724.

Yu, Ning, Vladislav Skripniuk, Dingfan Chen, Larry S. Davis, and Mario Fritz. 2022. "Responsible disclosure of generative models using scalable fingerprinting." In *International Conference on Learning Representations.*

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. "Defending against neural fake news." *Advances in Neural Information Processing Systems* 32.

Zhou, Peng, Xintong Han, Vlad I. Morariu, and Larry S. Davis. 2017. "Two-Stream Neural Networks for Tampered Face Detection." In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* IEEE, July. https://doi.org/10.1109/cvprw.2017.229.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In *International Conference on Computer Vision and Pattern Recognition,* 2223–32. IEEE, October. https://doi.org/10.1109/iccv.2017.244.

Author

Hany Farid is a Professor in Electrical Engineering & Computer Sciences and the School of Information at the University of California, Berkeley.

Data Availability Statement

Not applicable

Funding Statement

Not applicable

Conflict of Interest

The author is a paid advisor to Truepic, a company that contributes to the C2PA effort discussed in Section 6. In this capacity, the author is a member of the C2PA's Steering Committee.

Ethical Standards

Not applicable

Keywords

synthetic media, deep fakes, disinformation, digital forensics