WITNESS
SEE IT FILM IT CHANGE IT

CO CR
CO CREATION STUDIO

MIT OPEN DOCUMENTARY LAB

JUST JOKING!

DEEPFAKES, SATIRE, AND THE POLITICS OF SYNTHETIC MEDIA

EXECUTIVE SUMMARY

Around the world, deepfakes are becoming a powerful tool for artists, satirists and activists. But what happens when vulnerable people are not "in on the joke," or when malign intentions are disguised as humor? This report focuses on the fast-growing intersections between deepfakes and satire. Who decides what's funny, what's fair, and who is accountable?

The report analyses more than seventy recent cases of a wide range of deepfakes. Some are examples of potent satire, art, or activism, from mocking authoritarian leaders to synthetically resurrecting victims of injustice to demand action. But others demonstrate how bad actors use comedy as both a sword and a shield, to glorify the powerful and attack marginalized communities, while seeking to escape culpability. Increasingly, satire is used as a defensive excuse — "just joking!" — after a video has circulated and caused harm.

How can we distinguish intentional disinformation or gaslighting from critique? Whose images or voices are fair game to mimic without consent? The report probes the ethical challenges of democratizing synthetic media production, including how to label these videos as they circulate online, and who is responsible for creating and enforcing best practices, protocols, and laws. It considers the ways that media platforms are attempting or failing to make these distinctions in their policies and protocols, as well as the relevant principles and tools in international human rights and U.S. law.

The study concludes with an array of questions that highlight further directions for inquiry and action around the production, circulation, and reception of synthetic media. It argues that an expansive array of voices need to be part of this discussion, including not only technologists, corporations, lawyers, and politicians, but also human rights activists, artists, journalists, and communities from around the globe.

Just Joking! is part of a continuing collaboration between Co-Creation Studio at MIT Open Documentary Lab and the human rights, video and technology network WITNESS. WITNESS's 'Prepare, Don't Panic' initiative (wit.to/Synthetic-Media-Deepfakes) pursues a globally inclusive human rights lead approach to deepfakes, authenticity and media manipulation.

CREDITS

Produced by

Sam Gregory, WITNESS
and
Katerina Cizek, Co-Creation Studio
at MIT Open Documentary Lab

Just Joking! Is a part of an ongoing collaboration, Deepfakery, between WITNESS and Co-Creation Studio at MIT Open Documentary Lab, led by Sam Gregory and Kat Cizek. Deepfakery examines the artistic, political, cultural and policy implications of synthetic media.

Written by
Henry Ajder
Joshua Glick

Lead Research by
Henry Ajder

Additional Research by
Andrea Shinyoung Kim
Srushti Santosh Kamat
Joshua Glick
Jacobo Castellanos
Claudia Romano

With Contributions by
Professor Evelyn Aswad
Attorney Matthew F. Ferraro

Edited by
Joshua Glick
Carl Wilson

Advisors
William Uricchio
Sarah Wolozin

Illustrations by
Alex Wittholz

Designed by
Helios Design Labs

CREDITS

INTRODUCTION

Satire is thousands of years old. Deepfakes are new. Today, they coexist within a vast sea of social media, rife with misinformation, disinformation, and decontextualization. In these turbulent waters, who decides what's funny, what's fair, and who is accountable? How might we consider the progressive possibilities of the comedic arts, but also the weaponization of humor?

Deepfakes are computer-synthesized audio and/or video that make it seem like people have done or said things that they never did or said. The term was first coined on Reddit in 2017, in reference to non-consensual sexual images that attack women.[1] Since then, the majority of malicious deepfakes continue to target women, but deepfakes have also been increasingly tapped for their creative potential by artists, musicians, documentarians, amateur creators, and human rights activists, all with varying motivations and intentions. Satire has emerged as a vibrant realm of deepfakery, with consumer-tool generated memes, sophisticated faceswaps, and more elaborate digital scenes volleying incisive critiques in the service of equity and justice.

While part of a wave of emerging media, deepfakes themselves along with the tools needed to create them are not without historical precedent. They have evolved from a long line of audio-visual media, many of which were first met with debate and even moral panic. As recently as the 1990s, for example, the rise of home computing, digital video cameras, and software such as Photoshop stoked fears as to where digital fabrication might lead.

---

1   Samantha Cole's article in VICE was one of the first to raise awareness about deepfakes. Samantha Cole, "AI-assisted Porn is Here and We're All Fucked," Motherboard, VICE, December 11, 2017, https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn. For more on the rise of deepfakes, see the State of Deepfakes reports by Deeptrace (now Sensity) https://regmedia.co.uk/2019/10/08/ deepfake_report.pdf; the Prepare, Don't Panic: Synthetic Media and Deepfakes initiative by WITNESS, https://wit.to/Synthetic-Media-Deepfakes; Tim Hwang, Deepfakes: Primer and Forecast, NATO Strategic Communications Centre of Excellence, 2020, https://stratcomcoe.org/publications/deepfakes-primer-and-forecast/42; Tim Hwang, "Deepfakes: A Grounded Threat Assessment" (Center for Security and Emerging Technology, July 2020), file:///Users/joshuaglick/Downloads/CSET-Deepfakes-Report.pdf; Nina Schick, Deepfakes: The Coming Infocalypse (New York: Twelve, 2020).

In one sense, deepfake satire falls within an expansive tradition of subversive cultural practice that includes modernist Dada installations and collage paintings, theatrical "happenings" and "pranks," guerilla video and graffiti art. More broadly, satire has long held an established place across global popular culture, as well as literary, broadcasting,

and film traditions. Satire draws on narrative and literary tropes such as irony, innuendo, defamiliarization, and hyperbole, often pushing representation beyond the threshold of realism.[2] The aim is generally to hold real-life people and events up for critical evaluation.

Writing in the 18th century, biographer and essayist Samuel Johnson discussed satire in terms of exercising moral judgement, holding "wickedness and folly" up for censure.[3] More recently, media historians Jonathan Gray, Jeffrey P. Jones, and Ethan Thompson have argued that satire exposes or articulates a truth that may be obscured or hidden from public view. In turn, it provides a valuable means through which citizens can "analyze and interrogate power" and entrenched norms.[4] With the rise of online journalism, satire has found a place on media hubs ranging from The Onion (United States) to Al Hudood (Jordan), Punocracy (Nigeria), and Adobo Chronicles (The Philippines). Viewers of satire are encouraged to be alert and engaged. Communication scholar Danna Young writes that satire relies on the audience member's participation in order to achieve its intended effect; they are "in on the joke," with knowledge of the references and necessary context.[5]

Deepfake satire is shaped by at least three intersecting forces:

- First, a growing number of everyday citizens and professional filmmakers have access to a wide breadth of tools for image production and algorithmic design, allowing for heightened realism on a budget.
- Second, the media ecology in which synthetic media circulates has rapidly evolved over the last fifteen years. Deepfakes don't simply appear, rigorously

---

2   Northrop Frye, Anatomy of Criticism (New York: Atheneum, 1970), 223-4. For a cultural history of satire, see Jonathan Greenberg, The Cambridge Introduction to Satire (Cambridge: Cambridge University Press, 2019), 7-26.
3   Samuel Johnson, A Dictionary of The New English Language, ed. E.L. McAdam and George Milne (Mineola: Dover, 2005), 357.
4   Jonathan Gray, Jeffrey P. Jones, and Ethan Thompson, Satire TV: Politics and Comedy in the Post-Network Era (New York: NYU Press, 2009), 8-19.
5   Dannagal Goldthwaite Young, Irony and Outrage: The Polarized Landscape of Rage, Fear, and Laughter in the United States (Oxford: Oxford University Press, 2020), 69-84.

pre-vetted, on a few central media venues, but are uploaded on online platforms geared toward smooth distribution, quick viewing, and from-the-gut user comments. These platforms do make some decisions around what passes as falsified, what degree of deception is allowed, and what might be the relationship between deception and harm. Still, from a human rights perspective, they fail to deal adequately with the emerging influence of deepfakes as well as the nuances of satire and parody.

- Third, in our contemporary moment of fraught electoral politics and rapid technological change, satire has become a powerful means of resistance to the rise of authoritarian leaders and right-wing extremism around the world. Satire can fly "under the radar," smuggling in critiques that might be censored in other media or political contexts. Comedic sensibilities can flourish within an online environment that welcomes remixing and sampling. Additionally, mockery can be a tool of power, used to undercut opponents on all points of a political spectrum.

Online, not everyone is "in on the joke." The flattening of context and the ease and speed with which messages can be separated from messengers frequently lead to misfires, mis-understandings, and distorted meanings. Satire was never a completely "stable" cultural form, but in legacy print media, it often held a more socially legible and legally secure place, appearing in a demarcated column, in an editorial cartoon, or on a designated page of a newspaper. In 2021, satire floats freely in fractured, constantly evolving digital spaces. Additionally, bad actors are weaponizing humor with the aim of doing harm to individuals and communities. Some of their most threatening work takes the form of synthetic media. Deepfakes are well-primed to exploit the gray areas of "JUST JOKING!," providing social critique but also opportunities for malicious media and indi-viduals to gaslight under the guise of humor.

Authoritarian leaders and the movements that support them use humor as both a sword and a shield, to attack their enemies while also defending their actions as "merely a joke." Figures in power use synthetic media (and the plausible deniability they readi-ly enable) to dismiss forms of official evidence, citizen journalism, and widely known truths, while evading conventional protocols of responsibility and accountability.

In late 2020, WITNESS and the Co-Creation Studio at the MIT Open Documentary Lab produced Deepfakery, a series of critical conversations examining deepfakes within the broader framework of synthetic media and its relationship to art, cinema, journalism,

and disinformation. We came together with the understanding that no one sector or form of expertise can tackle these issues by itself. These discussions featured experts and practitioners from around the world, emphasizing the ways that synthetic media might be used for malicious as well as progressive purposes. As we look towards the next stage of the Deepfakery program, we continue to expand our investigation with an eye towards the possibilities and pitfalls of AI-enabled comedy.

This report draws on a wide range of over seventy recent synthetic-media case studies from around the world. We highlight areas that need further focus and discussion—issues we feel are best addressed by way of interdisciplinary, global, and collaborative approaches that bring together technologists, journalists, artists, legal scholars, and human rights activists.

We begin by introducing some fresh terminology into the discourse of synthetic media, as well as clarifying some buzz terms. Then, in Part I, we focus on the ethics and aesthetics of satirical deepfakes, and note key themes that require further engagement. Part II reflects on the far right's weaponizing of humor and its relationship to the contemporary media landscape. Part III features the voices of legal scholar Evelyn Aswad and attorney Matthew Ferraro, whose writing sheds light on the way deepfake art should be viewed within an international human rights framework, and how it can be protected by law. Part IV takes a critical look at how social media platforms, through a combination of bottom-up and top-down prods, pleas, and demands, are attempting to moderate synthetic media.

We conclude with an array of questions that distill some of the most important topics and debates, not to prompt quick answers, but to prepare for future inquiry. These questions are best addressed through informed discussion, centering the diverse global voices of the people most at risk of harm from synthetic media's abuses, and also those who stand to benefit most from the potential to challenge oppressive power via new forms of art.

*WITNESS helps people use video and technology to protect and defend human rights. Over the past three years, our **Prepare, Don't Panic** initiative has led a critical effort to provide a framework for understanding deepfakes and other forms of synthetic media, and ensuring that a global, human rights-led perspective guides both prioritization of threats and action on solutions.*

The Co-Creation Studio is part of the MIT Open Documentary Lab (ODL), which convenes storytellers, technologists, and scholars to advance the art of documentary with emergent technologies and collective methodologies.

DEFINITIONS

Synthetic Media

Media that is enabled or modified by Artificial Intelligence (AI) has come to be termed "synthetic media." There are many kinds of synthetic media, and the field is constantly expanding to include image and text generation, music composition, and voice production. Just as with legacy technologies, the use or application of synthetic media can be for the civic good or societal harm. There is increasing interest in using synthetic media for art, urban planning, environmental science, and health care.

Deepfake

"Deepfakes" use deep-learning algorithms to create realistic simulations of a person's face, voice, or body. As a form of synthetic media, deepfakes can make it seem like someone did or said something they never did or said. The term was first used in reference to the "faceswaps" where female celebrities' faces were swapped without consent with those of adult film actresses in sexually explicit videos. However, as the term quickly caught on in mainstream entertainment culture, journalism, and fan communities, it began to refer to a broader AI-enabled practice of making media, rather than just faceswaps or a specific malicious application.

A 2018 deepfake features the lip movements of former President Barack Obama synthetically synchronised with director-comedian Jordan Peele's voice.

Shallowfakes and Cheapfakes

"Shallowfakes" are audiovisual media deliberately re-presented in a new context. This can involve re-labelling media to reframe when or where something happened, or captioning a video to alter its perceived meaning. For example, a video of a burning building in India was re-packaged and circulated in South Africa to give the impression that it depicted xenophobic attacks in Johannesburg. Shallowfakes can also include "lookalike" footage, making it seem that a particular person is present in a scene, when in fact it's only a person who resembles the individual in question.

"Cheapfake" refers to a broader category that may involve similar techniques of recontextualizing and re-labelling, but also speeding up or slowing down footage, and other forms of audio-visual manipulation such as removing parts of an image or inserting new elements.[6] As their names imply, cheapfakes and shallowfakes are created through widely accessible, relatively inexpensive software.

Images of a burning building in India circulating online suggesting xenophobic attacks in Johannesburg

Satire

Satire is a genre of creative expression that draws on comedic devices (hyperbole, irony, etc.) to cast judgment on a person, group of people, a set of norms, or a larger idea. At

---

6   Britt Paris and Joan Donovan, Data & Society, Deepfakes and Cheap Fakes: The Manipulation of Audio-Visual Evidence (2019), https://datasociety.net/library/deepfakes-and-cheap-fakes/

their best, such works of art aim to reveal larger social truths. Humor is crucial to the technique and popular appeal of satire, whose history spans folk and oral traditions, visual art, music, poetic and essayistic forms, and theatrical productions. Satire has continued to live on via large and small screens in the 20th and 21st centuries.

Satirical newspapers from around the world. From top to bottom, left to right: USA's The Onion; Nigeria's Punocracy; Jordan's Al Hudood, and Mexico's El Deforma

Parody

Parody is a comic imitation of a person, work, or style, often involving exaggeration or playful stylization. It can serve a critical and edifying function, poking fun at a target and skillfully showing how it operates. Parody frequently takes the form of light entertainment, such as musician Weird Al Yankovic's reimagining of pop songs. It can also offer critique, such as the Radi-Aid campaign irreverently mimicking Western charity advertisements and music videos.

Malicious Media Manipulation

Malicious media manipulation refers to the deliberate alteration or fabrication of a media object (image, photo, video, audio) for harmful purposes. This often functions as a form of disinformation (false information deliberately disseminated to trick viewers). Misinformation is a more capacious category of false information that may or may not have been created or shared with intent to harm.

Gaslighting

Gaslighting is a form of psychological manipulation that aims to erode an individual or group's perception of reality. Originally referring to emotional abuse in romantic relationships, the term stems from Patrick Hamilton's play Gas Light (1938). The plot centers on a husband who makes his wife doubt her own sanity by causing their home's gas lights to flicker, then claiming she's simply imagining it when she notices. In politics, gaslighting describes disinformation campaigns that not only spread false narratives, but destabilize opponents by adamantly denying known facts and their own experiences. To gaslight somebody or an entire community is not simply to lie to them, but to directly or insidiously undermine their hold on truth, sowing seeds of confusion and distrust.

PART I: DEEPFAKES FOR THE CIVIC GOOD

Deepfakes can augment legacy practices of civic art, while also enabling new forms of political satire, social advocacy, or pop fandom. Still, these works prompt ethical questions concerning how synthetic media is created and distributed. Who it's "okay" to deepfake and under what circumstances are hotly contested.

Profiles in Political Folly

Whether in the form of a newspaper cartoon, sketch-comedy routine, or theatrical puppet show, political satire has long targeted authoritarian leaders and dictators. In Berlin in the 1930s, John Heartfield's anti-fascist typography and photomontage took aim at Adolf Hitler, at once critiquing the flagrant hypocrisy and brutality of the Third Reich and mocking its affinity for pageantry.[7] A few years later, Charlie Chaplin embodied the Hitleresque dictator Adenoid Hynkel in The Great Dictator (1940), exposing the fascist mobilization of fantasy and the vanity and delusion of authoritarian leaders. More recently, satirists have gravitated to deepfakes, using synthetic versions of politicians' faces in irreverent resistance to oppressive regimes and the extremist movements behind them.

The majority of these "faceswap" videos are created by individual artists or amateurs who go by pseudonymous handles on social media. Most are quickly and even crudely fashioned, and do not intend to fool viewers per se. They make their subjects look ridiculous, ignorant, or evil by way of a grotesque performance. These videos reveal what may be masked in the individual's day-to-day professional life or hidden from public view. In some cases, lip-synched impersonations or synthetically replicated speech give viewers the impression that these politicians are denigrating their own platforms or party ideologies.

---

7   Sabine Kriebel, "Manufacturing Discontent: John Heartfield's Mass Medium," New German Critique, No. 107 (Summer 2009): 53-88; Jodi Sherman, "Humor, Resistance, and the Abject: Robert Benigni's Life is Beautiful and Charlie Chaplin's The Great Dictator," Film & History: An Interdisciplinary Journal of Film and Television Studies 32:2 (2002): 72-81.

Artist Bruno Sartori was an early innovator in this realm of deepfakery, making a name for himself through his mocking of Brazil's far-right president, Jair Bolsonaro. One video depicts Bolsonaro's paranoid demonization of his go-to scapegoat, the former president Lula (Luiz Inácio Lula da Silva), by faceswapping both of them into Mariah

Carey's Obsessed music video. In Lava Uma Mao, Sartori calls attention to Bolsonaro's reprehensible response to the coronavirus pandemic, portraying the embattled president singing about the urgency of proper hand washing.

In an interview for the Deepfakery web series, Sartori explained that he was not only attacking Bolsonaro, but addressing the leadership vacuum in Brazil and the stakes of public policy: "I am using the image of the president of Brazil, and unfortunately the person who occupies that chair is Bolsonaro. So, the image I am using is a public image, the image of the President, to create a critique of the attitudes he takes, not portraying their personal life, but their public life."

Bruno Sartori's Lava Uma Mao comically portrays President Bolsonaro singing about the importance of washing hands due to the global pandemic.

Other videos call out oppressive leaders in different ways. Turkey's President Recep Tayyip Erdogan focuses on his administration's poor record on women's rights. French Faker's Marine Le Pen parle Arabe? mocks the xenophobic policies of France's right-wing National Rally politician Marine Le Pen by showing her wearing a hijab

and speaking Arabic, thus turning her into the thing she disparages. FacetoFake faceswapped Santiago Abascal, the leader of the Spanish far-right party, with the mischievous "problem child" character from the film, Este Chico es un Demonio—El Rey del Colegio?, in a YouTube video by the same name.

TheFakening's whimsical XI Jinping as Winnie the Pooh Dancing to Bat Out of Hell By Meatloaf shows the Chinese President cosplaying as the portly fictional bear, a comparison that has become a symbol of resistance against Xi's Communist government. Ever since Xi-as-Pooh took off as a meme in 2013, Chinese censors have been struggling to erase all such depictions from public view, recently resulting in the blocking of the Disney film Christopher Robin from Chinese distribution.

Recognition of deepfakes' popular appeal has also fast-tracked their entry into mainstream entertainment. In October 2020, South Park creators Trey Parker and Matt Stone, along with comedian and voice artist Peter Serafinowicz, launched Sassy Justice, a web series featuring deepfaked U.S. politicians and celebrities as the central characters. To make the series, the showrunners formed Deep Voodoo, a studio comprised of graphic designers, digital artists, and engineers. Poking fun at both local station newscasters and figures such as Al Gore, Donald Trump, Ivanka Trump, and Jared Kushner, Sassy Justice used AI as part of its narrative fabric. In a similar fashion, the Italian program Striscia la notizia used deepfakes in send-ups aimed at politicians and celebrities such as former Italian Prime Minister Matteo Renz

Left: Sassy Justice creates a deepfake of former President Donald Trump. Right: Striscia la notizia creates a deepfake of former Prime Minister Matteo Renzi.

Framing the Conversation: Fair Game for Public Figures?

For many democratic societies with a tradition of free speech, an individual's "public" or "private" status is important when considering whether their consent is necessary before they become the target of a cultural work. Somebody whose words and actions are of legitimate public interest and concern is generally deemed to merit less control over their likeness than an everyday private citizen.

The artist and coder Daniel Howe argues in his contribution to the WITNESS/Open Doc Lab Deepfakery series that "it's useful to characterize some of these techniques as ethically justified in specific contexts, when there are power differentials at work. ... Those who are subject to deepfake manipulation are figures who themselves seek to manipulate masses of people with digital technologies." Howe's position echoes political scientist James Scott's "weapons of the weak" theory, which claims that art can provide a voice and a means of resistance to those lacking other forms of political agency.[8]

The discourse surrounding public figures could help offer broad protections to those synthetic media makers aiming to satirize figures who enjoy the societal spotlight. But the boundaries are not always clear between a public and a private person, especially

given the highly visible way that people live their lives online, and the increasing intersections between the spheres of politics and popular culture.

Questions of ownership can also be murky around the commodification of likeness. While a skilled actor doing an impersonation might clearly be in rightful possession of their filmed performances as intellectual property, the synthetic voice of a politician constitutes a noticeable distinction and might come with some restrictions for use. The profit motive is also an important consideration. Should it be necessary to obtain Leonardo DiCaprio's consent to make him into a realistic-looking avatar in a video game, reality TV show, or documentary? As an actor, he makes a living from the use of his image, and should have the agency to control and share in any money made from it. However, perhaps an AI DiCaprio should be seen as continuous instead with how the actor might be depicted or impersonated in traditional forms of animation, sketch comedy, or literary satire.

---

8   James Scott, Weapons of the Weak: Everyday Forms of Peasant Resistance (New Haven: Yale University Press, 1985).

Advocacy

Deepfakes are also becoming a tool for progressive advocacy, with approaches ranging from biting satire to more somber PSA-style messages. This kind of media takes many forms, and resonates with modernist art movements in the 1920s-30s, the street theatre of the 1960s, the anti-globalization video art and "culture-jamming" pranks of the 1990s, along with more institutionalized outreach efforts by nonprofits and NGOs in the 2000s.[9]

Deepfakes can humanize problems that might seem distant or abstract. The Pakistani climate-change initiative Apologia Project depicted world leaders apologising from the year 2032 for their previous inaction on environmental crises. The rhetorical power of the project derives from the seeming sincerity of the leaders' remorse and the knowledge that so much more could have been done under their watch. As part of their HIV Treatment4all campaign, the French charity Solidarité Sida showed Trump bragging about how he eradicated AIDS. The video at once satirizes the former U.S. president's tendency to take undeserved credit for things he wasn't involved in or that never happened. At the same time, the video calls out world leaders to take a more active role in helping to combat the still-pressing disease, as the effort to offer medical treatment at scale remains unfinished.

Malaria No More UK joined forces with Synthesia AI to create a short film featuring football legend David Beckham speaking out against malaria. Spearheading the organization's "Malaria Must Die, So Millions Can Live" campaign, the video depicts Beckham effortlessly shifting between nine languages, his lips perfectly synchronized to layered audio tracks from different voices. He stares at the camera and emphasizes the severity of the disease, communicating that there needs to be greater global coordination to end it.

The Belgian collective Extinction Rebellion (XR) made a deepfake of then-Prime Minister Sophie Wilmès delivering a formal address about how Covid-19, SARS, Ebola, and the Swine Flu were all a direct result of the climate crisis. The aim was to pressure the government to acknowledge this link, and to look beyond just the current pandemic to the ongoing dangers of human-generated climate change.

Artists also have partnered with advocacy organizations to depict alternative futures.

---

9   See, for example, Marilyn DeLaure, Mortiz Fink, and Mark Dery, eds., Culture Jamming: Activism and the Art of Cultural Resistance (New York: NYU Press, 2017).

These aspirational visions suggest what could happen if power brokers and institutions possessed a stronger civic will or were held more accountable for their actions. Bill Posters teamed up with anonymous Brazilian activists for the fictional Amazon Prime series, Green Heart. The deepfake promo depicts Amazon CEO Jeff Bezos announcing his commitment to protecting the Amazon rainforest on the occasion of the company's 25th anniversary. Playing on the double meaning of the company's name, the project asks what might be possible if one of the world's foremost extractors of human labor and natural resources shifted instead to conservation and sustainability.

Left: A deepfake of former Prime Minister Sophie Wilmès speaking about the link between the health and climate crisis. Right: A deepfake of David Beckham where he speaks nine languages to support the "Malaria Must Die, So Millions Can Live" campaign.

The UNICEF Innovation Office, MIT, and the Scalable Cooperation Group aimed to build understanding for Syrian war refugees through the AI project Deep Empathy. It drew on synthetic image translation with pictures from the war-torn city of Homs to create vivid renderings of what San Francisco, London, or Tokyo would look like if they faced similar devastation. DeepEmpathy brings the war home for many Western viewers, while compelling them to reflect on how race, class, and geography shape how certain crises are prioritized over others.

Some creative endeavors have proven to be too hot for mainstream media. AI-generated PSAs from the pro-democracy group Represent US were pulled by TV networks before they were scheduled to air after the first 2020 U.S. presidential debate. The deepfakes featured Vladimir Putin and Kim Jong-un assuring viewers that rival powers didn't need

to destroy U.S. democracy, because "America is doing that to itself." While the networks

provided no rationale for pulling the advertisements, a spokesperson for Represent Us said that they'd been "too bold" to air.

Controversy has surrounded the use of AI tools to synthetically "resurrect" deceased victims of injustice as a technique to demand change. The group Propuesta Civica "brought back" the murdered Mexican journalist Javier Valdez Cárdenas to call for an end to state-backed violence against the press. In the U.S., the nonprofit Change for Ref in collaboration with the McCann Health campaign created a video featuring Joaquin Oliver, a victim of the mass shooting in Parkland, Florida, to promote gun-control legislation. And in Lebanon, the independent station MTV news "brought back" victims of the Beirut explosion to call for justice on the one year anniversary of the August 2020 event. Even though the filmmakers obtained consent from the victims' families, some observers considered it tasteless and disturbing to present deceased people delivering politically charged messages.

Three years after his death, Propuesta Civica's deepfake "brought back" Mexican journalist Javier Valdez Cárdenas to call for an end to state-backed violence against the press.

Framing the Conversation: The Ethics and Aesthetics of Persuasion
Makers of advocacy-oriented media are constantly evaluating the most effective means to reach, engage, and mobilize viewers. A speech delivered directly to the camera might be efficient at communicating a message, but viewers may tune out because of didactic delivery or lack of storytelling. On the other hand, an argument packaged within a

compelling narrative or visually stunning environment might move viewers emotionally, but could cloud the specific call to action.

The visceral realism of synthetically generated media can help raise attention and awareness, but over-the-top stylization and shock effects can create discomfort or distraction from the message. This was a criticism of graphic and emotionally wrenching drunk driving or anti-smoking PSAs on U.S. and U.K. television.

Synthetic resurrection must be treated with heightened sensitivity. Animating images of the dead may be disturbing for viewers, especially those close to the individual. The project's aims and the wishes of the person's family and friends must be carefully considered, as well as the details of the treatment. Deepfakes that exploit the dead in a hasty effort to score political points could overwhelm or derail the cause.

Tools, Memes, and Tough Cases
The development of accessible, easy-to-use apps has helped to spur deepfakes' proliferation across the Internet. Wombo creates lip-synching face portraits that can perform an array of songs. Reface superimposes a face onto existing gifs. MyHeritage's "deep nostalgia" feature animates eyes, faces, and mouths from old photographs. Sway uses motion filters to transform a static pose into a dancing or stunt-filled sequence. FaceApp allows users to age and contort the image of a face. Zao draws on a film and TV clip library to enable voice modulation and faceswaps.

Apps like Wombo, MyHeritage, Reface and FaceApp have helped to spur deepfakes' proliferation across the Internet

These apps have facilitated the synthetic "memeification" of both well-known figures and everyday people using them on their friends and family. Boasting the amusement of variation and relying on just enough insider knowledge, deepfake memes attract a subculture of interest among like-minded viewers and fan communities, especially on

platforms such as TikTok. Making these memes requires as little as one image of the targeted individual, and the user interface is designed to feel intuitive.

This level of accessibility does come with constraints. Most apps have a limited array of pre-selected templates, which make it easy for viewers to replicate and modify pre-existing media. Some makers and viewers consider the low production values of such app-generated deepfakes to be part of the aesthetic. The glitchy, grainy image quality feeds into the absurdist fantasies that the memes conjure, particularly when applied to well-known Hollywood stars or cartoon characters.

Software packages such as DeepFaceLab facilitate the making of more elaborate and polished videos. This approach tends to require more expertise, training data, and expensive hardware. The maker plays an active role during each step of production rather than simply uploading a picture into an app. Nonetheless, the prime interest with most of these more professionally rendered videos remains the same: inserting movie stars, musicians, and politicians into well-known slices of screen culture.

Nicolas Cage was one of the first celebrities to surface as a deepfake meme in 2018. The actor's manic eccentricities, span of roles (including his starring performance in the all-too-relevant Face/Off), and off-screen antics allowed him to build on his highly visible presence within Internet culture. This flurry of "Cagefakes" reflects the entertainment value of AI-enabled media and the ease with which it can be made by enthusiasts. Another popular fancentric deepfake meme along these lines was Back to the Future featuring Tom Holland and Robert Downey Jr.

A more elaborate and participatory project is Baka Mitai, in which any face can be animated to sing along to the Karaoke song "Baka Mitai" from the Xbox game Yakuza 0. It was initially created using open-source software and YouTube tutorials, but after going viral on TikTok, an app version emerged. Almost anyone can now create Baka Mitai-esque videos lip-synced to other popular songs or featuring custom face movements.

While beloved by fans, deepfake entertainment can cause confusion. In early 2021, strikingly realistic deepfakes of Tom Cruise created by expert VFX designer Chris Ume and professional Tom Cruise impersonator Miles Fisher went viral on TikTok. The videos of Cruise playing golf, performing a magic trick, and imitating a snapping turtle were not intended to deceive and were collectively published on the account

DeepTomCruise. However, the precision of Fisher's impersonations, combined with Ume's expert AI work, tricked some viewers into thinking the videos were authentic when they began to circulate more widely.

In other instances, deepfake tools have stoked legal controversy. The Canadian psychologist Jordan Peterson threatened a lawsuit over NotJordanPeterson, a text input-to-AI enabled speech web app that could make the combative right-wing culture warrior and free-speech advocate say anything users wanted. The site owners quickly abdicated. On another occasion, Jay-Z unsuccessfully sought the removal of deepfakes from YouTube that simulated him reciting Shakespearean soliloquies and singing Billy Joel's "We Didn't Start the Fire." The hip hop artist's DMCA takedown notice claiming that the deepfakes violated copyright was ultimately rejected by Google. The creator, Vocal Synthesis, had previously made similar deepfakes depicting Queen Elizabeth II reading lyrics by the punk band the Sex Pistols and conservative pundit Tucker Carlson reading the Unabomber Manifesto.
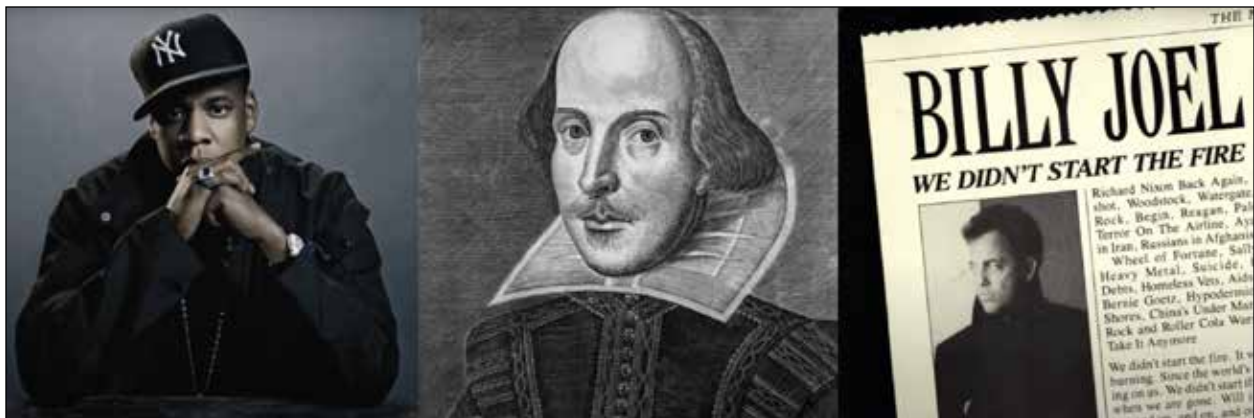
In a voice deepfake created by Vocal Synthesis, we can hear Jay-Z rapping Shakespeare's "To Be or Not To Be" and Billy Joel's "We Didn't Start the Fire." The hip hop artist unsuccessfully sought the removal of the videos claiming copyright infringement.

Framing the Conversation: Use and Abuse of Accessible Design
Given the popularity of homemade deepfakes and the ease with which they can circulate, these apps call for more attention. It is perhaps ironic that Peterson, a free-speech absolutist, reacted so strongly to the technologies themselves, claiming that "the sanctity of your voice, and your image, is at serious risk." Still, NotJordanPeterson's seemingly limitless possibilities do raise questions of ethical design. As Henry Ajder and Nina Schick wrote in an article for Wired, the more open an app, the more susceptible it is to malicious uses.

In confronting questions of design protocol, apps will need to reconcile with Terms of Service Agreements, potential limitations on likeness, and restrictions concerning who can make deepfakes. This is especially true since media literacy remains underdeveloped among so many young and old makers and consumers alike. Finding the right balance between creative interests and individuals' rights and restrictions over production will be crucial. WITNESS has begun to workshop a draft code of ethics that incorporates rights, potential harms, technological advances, and creative freedoms.

Jay-Z's resistance to being re-created via AI also points to how, in a different scenario, bad actors could dismantle the legal protections granted to deepfakes (and other forms of expressive art and parody) by claiming that they violate copyright infringement. This is a particularly acute problem when these complaints are filed through automated

social media moderation systems. For example, YouTube's Content ID protocol struggles to distinguish between "fair use" and illegitimate cases. Threats of legal action could be used to intimidate creators who cannot afford costly court fights. To adjudicate all these cases adequately, the major platforms would need to hire and train many more moderators than they are currently willing to pay for.

Art and Documentary Horizons

Practitioners of documentary art are continuously pushing the boundaries of nonfiction storytelling, encouraging a more media-literate and socially engaged public. Work using synthetic media increasingly extends beyond short-form videos into larger projects. Bill Posters and Daniel Howe's Spectre installation features deepfakes of celebrities, politicians, and tech entrepreneurs boasting about how they manipulate users by harvesting their data. These absurdist performances draw viewers' attention to the ways that techno-utopian celebrations of a networked world mask pernicious forms of exploitation. As Posters and Howe explain, the project enables "audiences to experience the inner workings of the Digital Influence Industry" and "feel what is at stake when the data taken from us in countless everyday actions is used in unexpected and potentially dangerous ways."

Bill Posters and Daniel Howe's Spectre installation features deepfakes of celebrities, politicians, and tech entrepreneurs boasting about how they manipulate social media users. Some of those featured are Kim Kardashian, Mark Zuckerberg and Freddy Mercury.

Stephanie Lepp's Deep Reckonings features synthetic re-creations of men who have abused power attempting to grapple with their own words and actions. Vignettes range from U.S. Supreme Court Justice Brett Kavanaugh at a press conference to Alex Jones in an interview with Joe Rogan. Their confessions seem to be a complete inversion of their public lives, casting them in a moral light. These videos could be interpreted as conjuring fantasies of remorse and reconciliation, but they might also be seen as placing the real-world personas of these men in sharp relief, creating a more truthful account of their egregious words and actions. Lepp views Deep Reckonings as a way to use "our synthetic selves to elicit our better angels." Viewing a synthetic version of oneself addressing personal struggles, Lepp argues, could be a means for deepfakes to elicit self-betterment and personal growth. This is part of a broader wave of "Deepfakes for Good," including projects focused on mental health, education, and social justice.

The art installation In Event of Moon Disaster presents a counterfactual history of the 1969 Apollo 11 moon landing, encouraging viewers to think critically about how we construct narratives of the past and how we understand our current information ecosystem. Directed by Francesca Panetta and Halsey Burgund in collaboration with the MIT Center for Advanced Virtuality, the installation features a period American living room in which a fabricated TV news report plays on loop. At the heart of the report is a deepfaked Richard Nixon reading the contingency speech his administration had prepared in case the Apollo 11 space mission failed and the astronauts became stranded on the moon.

Complete with a "reveal" component of the installation and an accompanying educational website, In Event of Moon Disaster serves as both revelation and warning. According to Panetta and Burgund, the project was designed to raise awareness of "how new technologies can obfuscate the truth around us" but also to demonstrate how the same technologies can be used constructively.

Left: Richard Nixon reading a contingency speech after the Apollo 11 space mission "failed." Right: Museum installation features a period American living room in which a fabricated TV news report plays Nixon's contingency speech on loop. According to the artists, the project was designed to raise awareness of "how new technologies can obfuscate the truth around us."

That charge is also taken up by the feature-length documentary Welcome to Chechnya. Director David France and VFX supervisor Ryan Laney created digital faces to shield twenty three persecuted members of the Chechen LGBTQ+ community. France and Laney found queer activists in NYC to "lend their faces" to the project as an activist gesture. They were photographed with a nine-camera setup that captured their faces from every possible angle, then matched with the film's subjects through a deep-learning process, tweaked with meticulous effects work. France and Laney originally considered other methods such as blurring the subjects' faces or casting them in shadow, but opted for this bespoke form of faceswapping. This way, identities are protected while preserving the subjects' humanity, allowing them to express themselves more fully to viewers. A slight halo surrounds their heads, signaling that their faces have indeed been altered, but the effect doesn't distract from the action. Here, the deepfakery (a term that France himself bristles at) serves a practical, narrative function along with a higher ethical purpose.

In the film Welcome to Chechnya, deepfakes are used to protect the identity of persecuted LGBTQ+Chechens.

Framing the Conversation: Labeling and Disclosure

Signaling explicitly to viewers that they are watching a deepfake can be a way of ensuring transparency. In more top-down, legacy news publications and broadcasting, it was easy enough to frame material with a caption or a parenthetical note, a label above a headline, or a program host's wry introducion. Online, and especially on social media, such transformations and parodies can more easily reach a viewer as decontextualized fragments. The dilemma is that explicit labels (watermarks, pre-roll warnings, etc.) might neutralize the satire's impact, and subtler markers (or none at all) might lead to misinterpretation by viewers or platform moderators. The protocols of such "semantic signposting" are far from clear. And there is no one-size-fits-all solution; as illustrated by the debate about a use of synthetic voice in Roadrunner, a recent documentary about the late chef and television host Anthony Bourdain, questions of audience expectations and genre conventions vary even within the documentary space.

Who is responsible for providing context for deepfakes, and what might be an appropriate marker? The websites for Deep Reckoning and Spectre frame their

projects as works of art, but their respective creators, Lepp and Posters, have different perspectives regarding labeling that they shared in a Deepfakery video episode. Lepp added watermarks along with introductory and concluding disclaimers, believing that they emphasize "the power of the medium, that you can know it is fake and it will still influence or move you." By contrast, Posters prefers not to use prefatory text, believing that it undermines the videos' rhetorical power. He insists that the realism of the performances, along with the way they might challenge gatekeepers to assess and categorize them, is part of the project's point.

For documentary, disclosure poses an important ethical issue. The slight halo or "afterblur" that France created around the subjects' heads in Welcome to Chechnya continuously reminds viewers that the faces have been altered. In Roadrunner, director Morgan Neville put no indicators around the synthetic audio of the deceased culinary adventurer "reading aloud" a despairing email to friend and artist David Choe. In addition, Neville claimed to have gotten consent from Bourdain's estate via his literary executor and widow Ottavia Busia. However, Busia took to Twitter to deny it. Many understood Neville's comments about the scenes in a New Yorker interview — "we can have a documentary ethics panel about it later" — to be flippant and dismissive.[10]

Given the ease with which deepfakes can be made, altered, and shared, social media companies need to take seriously how they're managed. A nuanced and interpretive approach to moderation would assess the implications of different forms of sound and image fabrication. Too light a touch could confuse viewers or lead to the spread of online misinformation. Too aggressive an approach could result in deepfake art being unable to find a platform.

A series of articles by First Draft and The Partnership on AI asserts the need for more conscientious labeling, but more research is needed on the potential risks and benefits. Legal and policy guardrails can also impact these processes. Countries that already heavily police their public sphere, such as China, will likely respond to deepfakes in the same sweeping and punitive fashion as they do other forms of sociopolitical critique. In the United States, social media companies' hands-off approach to the content they host

---

10   Helen Rosner, "The Ethics of a Deepfake Anthony Bourdain Voice," New Yorker, July 17, 2021 https://www. newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice; Helen Rosner, "A Haunting New Documentary About Anthony Bourdain," New Yorker, July 15 2021 https://www.newyorker.com/ culture/annals-of-gastronomy/the-haunting-afterlife-of-anthony- bourdain; Interview with Justin Hendrix and Sam Gregory, "Voice Clone of Anthony Bourdain Prompts Synthetic Media Ethics Questions," Tech Policy Press, July 16, 2021, https://techpolicy.press/voice-clone-of-anthony-bourdain-prompts-synthetic-media-ethics-questions/

has been shaped by Section 230 of the Communications Decency Act, which states: "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." Given the outsized roles these platforms now play in how media and information circulate, many watchdogs, critics, and politicians have argued that this framework needs to be revisited.

Specific pieces of legislation on the books involving synthetic media include California's AB 730, which bans political deepfakes 30 days before an election, unless they come with an explicit disclaimer or are clearly works of satire/parody. Policies of this kind strive to ensure campaign equity and protect the democratic process.

PART II: DISINFORMATION AND THE WEAPONIZATION OF HUMOR

Malicious deepfakes resonate with a long history of media designed to confound, discredit, or oppress. Totalitarian leaders have been notorious in this regard. In the 1930s and 1940s, Joseph Stalin "erased" from photographs individuals who fell out of party favor. This process of altering images was a way to "de-person" individuals from official records, literally "cutting them out of history." The realism of present-day audio-visual manipulation and the ease with which deepfakes can be made and shared make the threat more urgent. In particular, media made under the guise of "comedy" serves as both a weapon against opponents and a shield to dismiss accusations of ill intent.

Playful Propaganda

Amateur makers and state outlets have enlisted deepfakes to inflate the public image of authoritarian leaders by light-heartedly depicting them as Hollywood heroes; for example, Filipino President Rodrigo Duterte appearing as Iron Man. There's no direct relationship to the details of the Marvel films; it's simply a quick, spectacular impression of heightened masculinity, casting Duterte as a super-human hero.

Supporters of Vladimir Putin, likewise, have morphed the Russian strongman president into a slew of on-screen stars, from The Witcher to James Bond. The main state broadcaster, Russia Today, has published two provocative deepfakes affirming Russia's global power and mocking accusations of illegal meddling in foreign elections. In one video Trump diligently works for the broadcaster after losing the election. In another, Trump along with Angela Merkel, Emmanuel Macron, Boris Johnson, and Joe Biden complain to a therapist in a paranoid fashion about Russia Today's election interference. The tagline reads, "They're crazy about us."

Politician-as-superhero deepfakes are used to inflate the public image of authoritarian leaders and dismiss their detractors.

Beware the "Liar's Dividend"

Deepfakes' very existence can be exploited by those in power to cast doubt on credible news sources, allowing bad actors to dismiss demonstrably true evidence as the supposed results of digital fabrication. Legal scholars Danielle Citron and Robert Chesney say this rhetorical move yields a "liar's dividend," letting individuals dodge responsibility for their statements and actions. This tactic can have a corrosive effect on credible journalism, as accused figures can dismiss the efforts of media outlets aiming to hold them accountable for wrongdoing as simply "fake news." Citron and Chesney note that this confusion (and even skepticism) surrounding truthful reporting creates a climate ripe for the rise of authoritarian leaders.[11]

Philosopher Regina Rini observes that the increasing production and circulation of deepfakes could heighten a general public distrust in audio-visual media. Rini writes that audio-visual recordings have served democratic societies in providing a useful "epistemic backstop" against which a wide breadth of claims concerning utterances and actions could be evaluated. However, the rise of deepfakes could lead to a pervasive crisis of confidence, making new methods and protocols (especially at the legal and political level) necessary for discerning fact from fiction.[12]

---

11   Robert Chesney and Danielle Keats Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security" (July 14, 2018). 107 California Law Review 1753 (2019), 1785-86.
12   Regina Rini, "Deepfakes and the Epistemic Backstop," Philosophers' Imprint 20, no. 24 (August 2020): 2-9.

In 2019, a poorly made video of a New Year's address by the allegedly incapacitated Gabonese president, Ali Bongo, was declared a deepfake and part of a cover-up by opposition leader Bruno Ben Moubamba. Although the video was not, in fact, a deepfake, it contributed to growing unrest and an attempted military takeover. After the 2021 coup d'état in Myanmar, there were widespread claims to deepfakery concerning the video confession of the former chief minister Phyo Min Thein, who alleged corruption by ousted leader Aung San Suu Kyi. AI experts deduced that the video was most likely a coerced confession rather than a deepfake.

In the U.S., Donald Trump has claimed repeatedly that the Access Hollywood recording in which he bragged about grabbing women's genitals was "not authentic." Additionally, some of his supporters decried his concession speech following the January 6 Capitol siege as a deepfake. Republican congressional candidate Winnie Heartstrong also called the video depicting the 2020 killing of George Floyd a mashup of "digital composites."

The 2019 New Year's address by Gabon's Ali Bongo sparked national unrest after it was wrongly believed to be a deepfake.

As WITNESS' Sam Gregory wrote in Wired, the phenomenon of the liar's dividend and growing skepticism towards established media outlets presents challenges to journalists and human rights activists. How can these groups prove that their reporting is credible? How, in turn, can these groups identify whether state broadcasting is fact-based and

trustworthy? The pressure will only intensify to prove to citizens and stakeholders alike which pieces of media are real and ought to be believed.

Gaslighting and "Humor"

The concept of gaslighting as a form of psychological manipulation has expanded from interpersonal relationships into politics: A powerful figure spreads a false narrative while trying to weaken opponents' confidence by denying their contrary perceptions and experiences. This can serve as a kind of strategic disorientation, sometimes with high stakes. When Vladimir Putin deployed Russian troops to the Crimean Peninsula in 2014, he simultaneously denied their presence, insisting they were simply "local self-defense forces" doing humanitarian aid.

It can also be a form of gaslighting to spread malicious media and then claim that people "didn't get the joke," or that it was taken out of context. Members of U.K.'s Conservative party made the claim that a video deceptively edited to make the opposition leader appear incapable of answering a question was simply "humorous" and "satirical." Likewise, the Philippines' Duterte frequently covers for controversial statements by claiming later they were jokes, such as his threat to pull his country out of the U.N.

Humor very often serves as both a sword and a shield. Misleading media, often laden with racist, conspiratorial, and xenophobic messages, can be packaged or retroactively explained away as harmless bits of sarcasm and irony that shouldn't be taken too seriously. Defenders will say that this media is simply meant to push the boundaries of pop culture and to fly in the face of overly sensitive notions of "political correctness." The white nationalist commentator Nick Fuentes, known for his anti-Semitic views, acknowledged this tactic, saying, "Irony is so important for giving a lot of cover and plausible deniability for our views."

Donald Trump and his supporters used these techniques throughout his first run for the presidency, his time in the White House, and his 2020 re-election campaign. An anonymously created Joe Biden website displayed all kinds of falsehoods about the then-candidate; it claimed to be "merely parodying" Biden's official homepage. Trump also said that he was speaking in jest when he requested that Russia find Hilary Clinton's emails and that he was being "sarcastic," but not "that sarcastic" when he claimed that Barack Obama was "the founder of Isis." Trump supporters circulated images that distorted Biden's face or recontextualized things

he said or did in order to caricature him as "sleepy" or "delirious." [13]

In 2020, Trump's deputy chief of staff Dan Scavino tweeted a video that appeared to show Biden asleep right before the start of a TV interview. The video was actually taken from an interview the same anchor conducted with singer Harry Belafonte from 2011, with Biden's still face overlaid onto that of the singer. In another incident, Trump's campaign posted a short clip featuring spliced audio in which Biden appeared to say "coronavirus is a hoax." The popular pro-Trump Twitter account Trump War Room frequently shared this kind of media, and then declared that it "triggered journalists who can't take a joke about their candidate." Several high-profile examples were created by the conservative meme-maker Carpe Donktum (aka Logan Cook), whose videos won a competition from the far-right media outlet, Infowars, and were often retweeted by Trump and other prominent Republicans. In response to criticism that a deceptive video had fooled supporters who believed it was real, Trump's head of campaign communications Tim Martaugh claimed the video was "obviously a parody." Republican figures were also targeted in this way—for example, a video of Mike Pence carrying boxes of medical equipment was re-cut to make it sound like he admitted the boxes were actually empty—but far less often.

Crossing the Aisle, Blurring Boundaries
Other media outlets operate in what might seem like a gray area between malicious media and more legitimate, socially engaged art. The self-described "Christian news satire" hub The Babylon Bee poses a challenge of interpretation. Founded in 2016 by aspiring pastor Adam Ford, the publication made a name for itself by poking fun at evangelical dogma as well as tensions between Trump and the Republican establishment. The article "Psychopathic Megalomaniac Somehow Garnering Evangelical Vote" drew a lot of attention in the publication's early years. The Bee was then sold to entrepreneur Seth Dillon in 2018, who has been moving it more towards a pro-Trump, anti-liberal stance, aiming to take on what it sees as the threat of an outsized left-leaning force.

The Bee's outlandish claims have caused Facebook, Twitter, and the New York Times to label their articles as "misinformation," which the Bee has pushed back on, garnering

---

13   For more on disinformation within the right-wing media ecology in the U.S., see Yochai Benkler, Robert Faris, and Hal Roberts, Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics (Oxford: Oxford University Press, 2018). W. Lance Bennet and Steven Livingston, The Disinformation Age: Politics, Technology, and Disruptive Communication in the United States (Cambridge: Cambridge University Press, 2021).

support from both grassroots and elite conservative circles via the #FreeTheBee campaign. The New York Times issued a formal apology after it claimed that the Bee "trafficked in misinformation under the guise of satire."

Still, it would be a mistake to see the Bee as simply an Onion of the right, a cultural organ unfairly scrutinized by the "liberal media" for its Christian-conservative brand. Such a comparison glosses over the targets of the humor as well as the style in which it is fashioned. Ridiculing the powerless constitutes what humor scholars such as Caty Borum Chattoo and Lauren Feldman call "punching down" rather than "punching up," and often veers into the territory of misogynistic and anti-LGBTQIA+ attacks.[14] Even as the Bee mocks claims that its content is intentionally deceptive, its articles reinforce readers' world views. Trump used to tweet stories from the platform as true, as when he shared his astonishment that Twitter had intentionally shut down their network to slow the spread of anti-Biden stories. As journalist Parker J. Bach describes in Slate, the Bee's articles might not in themselves be a form of mis/disinformation, but they strengthen and reinforce content that is deceptive. From a media-platform perspective, the question is not just how to label individual articles published and distributed by the Bee or others, but how to handle posts containing these articles that are circulated and commented on by other people.

---

14   Caty Borum Chattoo and Lauren Feldman, A Comedian and an Activist Walk Into a Bar: The Serious Role of Comedy in Social Justice (Berkeley: University of California Press, 2020).

PART III: THE LEGAL VIEW

Reflections from Professor Evelyn Aswad on Deepfakes and International Human Rights

Professor Aswad is the Herman G. Kaiser Chair in International Law at the University of Oklahoma, where she also serves as the Director of the Center for International Business and Human Rights. The following builds on her contribution to the Deepfakery web series.

The international law standard on freedom of expression is set forth in Article 19 of the widely ratified International Covenant on Civil and Political Rights (ICCPR). This covenant provides broad protections for freedom of expression, including different forms of art (such as satire) and human rights activism. Freedom of expression may be limited, but only if the government can demonstrate that a restriction passes the three-part test of legality, legitimacy and necessity/proportionality. If a government wants to limit or otherwise burden speech, including speech through the use of AI-enabled media, it bears the burden of proving that all three conditions are met. The UN Human Rights Committee's General Comment No. 34 further clarifies how these tests should be applied.

Although these human rights principles are designed for state actors, the international community has endorsed the UN Guiding Principles for Business and Human Rights, which call on companies to respect internationally recognized human rights in their business operations. The UN Special Rapporteur on Freedom of Opinion and Expression has called upon social media companies to use this framework in order to avoid infringing on freedom of expression and to address adverse impacts.

These tests might be understood, in attempting to regulate AI-enabled media, as follows:

Legality: Any proposed legislation or policies that restrict or burden expression must not, among other things, be vague or overly broad. In the case of deepfakes and other forms of AI-enabled media, this includes defining clearly what specific forms of media and their uses are being restricted. This practice gives users clear guidelines, limits the discretion of those implementing the policy (which helps to avoid selective and discriminatory enforcement), and avoids restricting practices that do not pose risks of harm.

Legitimacy: The reason for limiting expression must be a legitimate public interest objective, as set forth in ICCPR Article 19, such as protecting the rights of others, national security, or public health. Protecting a regime, head of state, or government official from the kinds of deepfake satire detailed in this report would not constitute legitimate grounds for limiting expression.

Necessity and proportionality: This test should be applied using interdisciplinary, multi-stakeholder input to determine when it is truly necessary to limit the use of AI-enabled media. This condition includes asking:

> **1.** Are there non-censorial approaches that could be deployed to achieve the public-interest objective? If effective non-censorial methods are available, it may not be necessary to burden speech to achieve the objective. In the context of AI-enabled media, there are a number of questions that should be considered. For example, are governments or social media platforms investing enough in digital literacy and other media-education initiatives to build societal resiliency with respect to the use of such technologies? Can fact-checking operations be effective? Are governments or social media platforms investing in technology that empowers consumers to know if they are looking at a deepfake?
> **2.** If non-censorial methods are insufficient to achieve the legitimate objective, what is the least intrusive measure that can be pursued to accomplish the goal? Speech regulators should develop a continuum of options to achieve the objective and then select the one that achieves the public-interest objective with the least burden on speech. With regard to AI-enabled media, labelling content would be less intrusive than deleting it.
> **3.** If the least intrusive measure is implemented, does it actually work? If the selected measures are ineffective or counterproductive, they cannot be upheld as they do not serve to achieve the public-interest objective. In the case of AI-enabled media, if governments and social media platforms implement measures

that burden or limit expression, they would need to monitor and be transparent about whether the measures are effective in achieving the legitimate public-interest goal (e.g., preventing intimate image abuse, disinformation harms, etc.)

Social media platforms need to ensure this three-part test is applied to all aspects of content moderation, including platform speech codes as well as human and automated moderation of speech. These companies should also be transparent with the public about the measures they are applying to regulate speech.

Reflections from Attorney Matthew F. Ferraro on Satire, Fair Use, and Free Speech

A former U.S. intelligence officer, Matthew F. Ferraro is an attorney at WilmerHale, where he works at the intersection of cybersecurity, national security, and crisis management.

Disclaimer: The following are general statements of legal principles, not legal advice and not intended as such. I'm speaking only for myself, and not on behalf of my firm or clients.

Protections Under the U.S. Constitution

Satire is generally protected under the First Amendment to the U.S. Constitution. This principle was enunciated in a well-known U.S. Supreme Court case, Hustler v. Falwell, 485 U.S. 46 (1988). Adult-media mogul Larry Flynt provoked a lawsuit by satirizing the Christian televangelist Jerry Falwell in a cartoon Campari ad in his sexually explicit Hustler magazine. The ad showed an illustrated Falwell "recalling" his first sexual experience—with his mother, in an outhouse. A disclaimer noted that it was an "ad parody not to be taken seriously." Falwell sued and won damages from the lower courts for intentional infliction of emotional distress. However, the U.S. Supreme Court reversed the lower courts' award and held unanimously that it was, in fact, a protected parody.

The ruling held that "public figures" such as Falwell may not recover damages for the intentional infliction of emotional distress without showing that the offending publication contained a false statement of fact, made with "actual malice." This requires knowledge that the statement was false or made with reckless disregard as to whether or not it was true.

U.S. defamation law distinguishes between public figures and private persons. For someone of public interest or familiarity—like a government official, politician, celebrity, business leader, actor, or athlete—to litigate a defamation claim successfully, they must not only prove that the statement was defamatory (a false statement of fact), but also that it was made with "actual malice." The higher standard is because the Constitution aims to promote debate involving public issues. Private persons outside of the public spotlight generally only need to prove that the defamer acted "negligently"—that they failed to behave with the level of care that someone of ordinary prudence would have exercised under the same circumstances.

Balancing First Amendment Rights with Fair Use and Defamation Laws
"Fair use" is the ability to use someone else's copyrighted intellectual property under certain permissible circumstances, such as satire, without authorization or monetary payment. Congress wrote the principles of fair use into law in the Copyright Act of 1976, stating that such purposes as criticism, news reporting, teaching, and scholarship should not constitute an "infringement of copyright." The Act articulates a four-factor balancing test to determine whether a use is a fair one. Critical aspects include the purpose and character of the use, the amount and substantiality of the portion used, and the effect of the use on the potential market value of the copyrighted work.

Since the passing of the Copyright Act, fair use has increasingly hinged on the question of whether the use was "transformative," and the extent to which the aggregate amount of material from the original media object is appropriate to the new use or application. The Center for Media and Social Impact has some excellent resources in this area.

For defamation, after determining whether the person suing is a private or public figure, the court will ask the following general questions: Is the statement in question indeed false, but purporting to be fact? Is the false statement somehow published, or communicated to a third person? Does the defendant show the appropriate level of fault (negligence if the plaintiff is a private person, or actual malice if a public figure)?

Lastly, the courts will ask if the false statements could and did cause the plaintiff to suffer damages such as loss of money or potential earnings, harm to their reputation, or difficulties in relations with third persons whose views might have been affected

Addressing Deepfakes

Several of the laws adopted by U.S. states around deepfakes specifically exempt satire from their prohibitions:

- In New York, a new right-to-publicity deepfake law (S5959/A5605C) establishes a postmortem right to protect performers' likenesses—including digitally manipulated likenesses—from unauthorized commercial exploitation for 40 years after death. However, the law includes broad fair-use-type exceptions for parody or criticism, or educational or newsworthy value. Notably, the law provides safe harbor. If the digital replica carries "a conspicuous disclaimer" that its use was not authorized by the rights holder, then the use "shall not be considered likely to deceive the public into thinking it was authorized."
- This same New York law also bars most nonconsensual deepfake pornography. In these cases, a disclaimer is no defense. The law requires written consent from the depicted individual.
- A law in California (AB-730) prohibits anyone from distributing, within 60 days of an election, any materially deceptive audio or visual media depicting any candidate on the ballot with "actual malice," i.e., the intent to injure the candidate's reputation or to deceive voters—unless the media carry an explicit disclosure that it has been manipulated. Again, this law exempts satire and parody, as determined by the courts. It also provides exemptions from liability for broadcasting stations and websites that label the media in a way that foregrounds the fact of alteration, or for airing paid political advertisements that contain materially deceptive audio or video.

Combating the Malicious Uses of AI-enabled Media

As I argue in a Washington Post article co-written with Jason C. Chipman, titled "Fake News Threatens Our Businesses, Not Just Our Politics," although free-speech rights protect opinion, businesses and individuals may have legal recourse—especially when third parties defame private individuals or benefit financially from spreading lies. Here are some of the possible remedies.

- State and federal laws bar many kinds of online hoaxes.
- Laws that could be applicable to deepfakes may include: defamation, trade libel, false light, violation of the right to publicity, intentional infliction of emotional distress, negligent infliction of emotional distress, and right to publicity.
- Manipulated media that harm a victim's commercial activities may also be

actionable under widely recognized economic and equitable torts, including: unjust enrichment, unfair and deceptive trade practices, and tortious interference with prospective economic advantage.

Federal laws may be applicable if the media misappropriates intellectual property:

- The Lanham Act prohibits the use in commerce, without the consent of the registrant, of any "registered mark in connection with the sale, offering for sale, distribution, or advertising of any goods" in a way that is likely to cause confusion. The Lanham Act also prohibits infringing on unregistered, common-law trademarks.
- The Copyright Act of 1976, which protects original works of authorship, may provide remedies if the manipulated media is copyrighted. (See the article I co-wrote with Jason C. Chipman and Stephen W. Preston for Pratt's Privacy & Cybersecurity Law Report.)

No court that I know of has weighed in so far on social media platforms' obligations to label or remove deceptive media that claims to be satirical. As noted above, labelling is relevant to some of the deepfake laws and, as we saw in the Hustler case, to some free-speech law, too. The major concern is "line drawing"—one person's satire is another's harmful defamation. In the United States, we've tended to err on the side of allowing free speech.

Strategies for Satirists

Generally speaking, I've observed that public figures are afforded less protection than private persons, so satirists are on firmer ground when they focus on the former rather than on everyday citizens. Opinions are usually protected, while stating false claims as fact or vice versa (e.g., X person did Y thing) are not. Thus, satirists will find themselves in safer territory to the extent that they focus on opinions and not on specific misleading or spurious assertions.

PART IV: SOCIAL MEDIA PLATFORMS, POLICIES, AND PITFALLS

Major social media platforms have specific policies for deceptively manipulated media as well as synthetic audio and video. A number of platforms are revising their rules and protocols for these areas, as well as trying to clarify their rules on satire and parody. However, these policies are often not consistently enforced, particularly on a global scale, and often fail to adequately account for the full range of false or misleading information, as well as artistic works (such as satire), that appear on these platforms. There is a lack of coherence and transparency regarding how decisions are made, as well as avenues of appeal available to users and others.

Civil-society advocates note that one core issue is that major social media companies do not adequately resource content moderation globally, nor understand local context sufficiently. In some national contexts, such as the U.S., Cambodia, and India, companies have been perceived as too closely involved with government entities and political actors. As WITNESS has noted elsewhere, human rights activists (e.g., in Myanmar, Brazil, Sri Lanka, Ethiopia, the Philippines, Hungary, and India) frequently call out the failure of platforms to act when social media is used to incite violence or amplify crises—often coordinated, commercialized and directed by governments. In the United States, activists have criticised Facebook and other platforms for failing to address racialized hate and misinformation. Although some progress has been made, it hasn't been enough.

Platforms vary to the degree that they consider synthetic media manipulation separate from that of shallowfakes or cheapfakes. They also differ in how they understand and evaluate manipulation, deception, and the likelihood that a particular work of media may cause harm. In many cases, deepfakes will be covered under other principles that are format-agnostic; for example, media that takes the form of hate speech, election interference, COVID-19 misinformation, incitement to violence, and non-consensual sexual images. As legal scholar Evelyn Aswad notes, policies on deceptively manipulated media should be designed with human rights principles in mind, an approach adopted

by Facebook's Oversight Board when examining a recent case that raised questions about free expression and satire.

Facebook's own recent explanation of how they are trying to better handle satire as well as humor in general, based on a set of interviews the company conducted with experts, offers a good primer on the key dynamics for any platform: "Stakeholders noted that humor and satire are highly subjective across people and cultures, underscoring the importance of human review by individuals with cultural context. Stakeholders also told us that 'intent is key,' though it can be tough to assess. Further, true satire does not 'punch down': the target of humorous or satirical content is often an indicator of intent. And if content is simply derogatory, not layered, complex, or subversive, it is not satire. Indeed, humor can be an effective mode of communicating hateful ideas." However, none of the platforms have a robust explanation of how they assess satire in a nuanced and contextualized manner.

CURRENT POLICIES

Facebook

Facebook's policy (as of October 2021) states that the platform will remove content if it has been "edited or synthesized—beyond adjustments for clarity or quality—in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say." Content will also be removed if it is the "product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic."

The policy's focus on AI-enabled media points to deepfakes along with forms of skillfully edited video. It does not include other forms of deceptively manipulated media such as shallowfakes and cheapfakes. However, Facebook's other Community Guidelines still apply to relevant issues concerning the credible threat of violence, agnostic to manipulation type, where action is taken based on "hate speech" "or "misinformation and unverifiable rumors that contribute to the risk of imminent violence or physical harm." Additionally, third-party fact checkers (who are from independent organizations contracted by Facebook) can flag shallowfake content as "false" or at least partially false. This results in viewers receiving a warning when they encounter this kind of media, and leads to duplicates and near-duplicates of the relevant images or video being identified with automation. This can help ensure that misinformation does not spread, but it also leads to the take-down of inaccurately classified shallowfakes and deepfakes.

Facebook does grant explicit exceptions for "content that is parody or satire" and "video that has been edited solely to omit or change the order of words." Despite this caveat, until recently Facebook did not make clear how considerations of satire would be made. A June 2021 Facebook Oversight Board decision has compelled the company to clarify how they incorporate satire into their assessments. In turn, this calls for Facebook to pursue a contextual understanding of media grounded in local context, and address how users themselves might indicate satirical intent. As of April 2021, Facebook has also started experimenting with adding labels to satire pages on users' newsfeeds.

Twitter
Twitter's policy states: "You may not deceptively promote synthetic or manipulated media that are likely to cause harm. In addition, we may label Tweets containing synthetic and manipulated media to help people understand their authenticity and to provide additional context."

The social media platform uses three key questions to decide if and how to label their content or decide to remove it. Is the content "synthetic" (created through AI tools) or is it deceptively manipulated? Is the content shared in a deceptive manner? Is the content likely to impact public safety or cause serious harm? Notably, the policy does not explicitly provide exemptions for satirical media, but bases moderation decisions on whether "the context in which media are shared could result in confusion or misunderstanding or suggests a deliberate intent to deceive people about the nature or origin of the content."

TikTok
TikTok's Community Guidelines on misinformation prohibits "digital forgeries (Synthetic Media or Manipulated Media) that mislead users by distorting the truth of events and cause harm to the subject of the video, other persons, or society."

However, the guidelines also make "exceptions under certain circumstances," including "artistic content, satirical content, content in fictional settings, and counterspeech." TikTok does not provide a transparent account of how often this exception is applied, whether the decisions made with this rule are applied to similar content algorithmically, or whether users can contest decisions.

YouTube
Youtube's Misinformation Policies (referenced October 2021) prohibit deceptively

manipulated media, which the platform defines as "content that has been technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm." As is the case with other platforms, deceptively manipulated media on YouTube will often be considered unacceptable under other policies on hate, violence, or specific circumstances such as elections. YouTube has rules on so-called "educational, documentary, scientific or artistic content," which allow exceptions for satire. For example, their policies on COVID-19 misinformation note that they "may make exceptions if the purpose of the content is to condemn, dispute, or satirize misinformation that violates our policies."

Coming Up Short

One perceived flaw with several social media companies' policies, including Facebook, is that they concentrate on deepfakes without directly addressing many of the other common forms of deceptively manipulated media, such as shallowfakes. They have been taken to task for only providing warning labels or reduced distribution for shallowfakes if they are flagged by third-party fact checkers. YouTube's deceptive-practices policy similarly makes an exception for clips taken out of context. Platforms that have deepfake-specific policies have argued that they aim to address the novelty of these AI-enabled production techniques and the difficulty for the public to detect them, while harmful shallowfakes are often covered by rules about impact and outcomes. Critics note that such rules are often inconsistently and incompletely applied.

(Mis)labelling Satire

Narrowly conceived or overly stringent practices of fact checking and labeling satirical media (and other forms of digital art) can have a stifling impact on creative expression, free speech, and open debate. This occurred in Cameroon when a local academic and activist shared a clearly fabricated video of the French ambassador telling Cameroonians that they never really achieved independence from France's colonial exploitation. Facebook's third-party fact checkers at the French broadcaster France 24 labelled the video partially false, thus nullifying some of the rhetorical power of the critique.

As Internet Sans Frontières director Julie Owono observed in the Deepfakery web series session on global human rights, labelling this video as "partially false" means "whoever shares this clip will be labeled as someone sharing "fake news." In attempting to "fight fake news [in this way], platforms actually contribute in silencing important debates and an expression in countries that desperately need those debates and expression."

This approach contrasts with Facebook's lack of intervention in Gabon when the opposition leader falsely claimed a video of president Ali Bongo was a deepfake, as discussed above. Social media policies need to be aware of the liar's dividend and false claims that media is manipulated, and ensure these are addressed with specific policies.

Local and Global Context

The Facebook Oversight Board's ruling on satire made it clear that the platform should address context. More specifically, the Board stated that Facebook should provide: "content moderators with: (i) access to Facebook's local-operation teams to gather relevant cultural and background information, and (ii) sufficient time to consult with Facebook's local-operation teams and to make the assessment. Facebook should ensure that its policies for content moderators incentivise further investigation or escalation where a content moderator is not sure whether a meme is satirical or not."

As noted earlier, all too frequently, the major social media companies do not adequately resource their global content moderation to provide the time, capacity, and expertise to make sound judgements on the media they host.

Inconsistent Enforcement and the Failure of Automation

Even when platforms' policies seem to be comprehensive, they are too frequently applied in erratic fashion. Twitter has been accused of inconsistency in their labeling practices, as was the case during the recent U.S. presidential election. The platform did not label ambiguously edited footage of Trump in a Biden campaign advertisement or deceptively edited audio of Biden from Trump's campaign.

Even when a company has determined that a work of media has violated its policies, many struggle to effectively remove it from their platform. TikTok faced criticism along these lines when they only took down a select few reposts of a widely circulated, deceptively edited video of Biden during the 2020 presidential election. The video depicted Biden saying, "We can only re-elect Donald Trump." Automated systems are used to identify duplicate or near-duplicate examples of policy-violating media, but these can perpetuate errors from false positives. They can also overlook media that is not in violation according to one protocol (a satirical work clearly contextualized as such), but is in violation of another (a satirical work re-contextualized to remove satirical intent and used maliciously).

PART V: KEY QUESTIONS, NEXT STEPS

There is a need for more sustained and nuanced reflection on the use and misuse of synthetic media, and its place in a broader media ecosystem. The following questions, grouped by subject heading, foreground the civic possibilities and potential harm of synthetic media, pointing toward future discussions and research agendas.

In the coming months, we aim to convene individuals from across disciplines, sectors, fields of expertise, and walks of life to help explore some answers. Please join the conversation.

Consent and the Targets of Satirical Deepfakery

1. In what cases, if any, is consent needed to target individuals in positions of power?

2. Should deepfakes' visceral and photorealistic nature require consent in ways that other forms of parody or impersonation do not?

3. Can someone meaningfully consent to their likeness being used by a third party in whatever way they choose?

4. Under what circumstances is deepfaking the deceased acceptable?

5. Under what circumstances should it be completely unacceptable to deepfake another person?

6. Are there practical steps available to distinguish better between satirical content and malicious manipulation masquerading as humor?

Disclosure, Labelling, and Intent

1. Should creators be required to label all deepfakes? How might this be implemented and enforced? Could the explicit labelling of deepfake satire and other forms of AI-enabled art undermine their resonance?

2. How can synthetic- media makers give insight to consumers about the process

of creation without necessarily neutralizing the rhetorical or aesthetic power of the object?

3. How could the responsibility for labelling and disclosing this kind of media be shared between the tool developers, creators, online platforms, and other stakeholders?

4. Do poorly executed deepfakes that happen to fool some viewers need to be treated by both platforms and lawmakers as functionally identical to those that explicitly aim to harm or deceive?

5. How can platforms' interfaces and user experience make it harder or easier for viewers to understand the origin, style, and intention of the media they encounter?

Weaponization and Differential Impact

1. What is the differential impact of weaponized deepfakes on countries in the Global South as well as marginalized communities in the West/Global North?

2. How can a human rights framework be applied for practical decision-making around manipulated media and deepfakes at a platform and governmental level?

Content Moderation and Countermeasures

1. What is the appropriate content-moderation approach for deepfakes and synthetic media that works globally and respects clear normative frameworks such as international human rights law?

2. What are the ways that platforms can ensure AI-enabled art (including deepfakes) are not discredited by automated moderation or misguided fact checking?

3. How do social media platforms' differing approaches to moderating this kind of media impact specific global regions more than others, or differ in impact based on culturally specific understandings of satire?

4. Should deepfake apps be legally required to sign onto a common "code of practice" to minimize potential for misuse? Might users be required to sign something similar?

Law and Policy

1. What is the role of government in regulating how different forms of media circulate online? Should legislation dictate protocols for categorizing and vetting AI-enabled media specifically, and/or set standards for accountability?

2. Should the use of deepfakes in political campaigns and political advertisements be prohibited in certain or all contexts?

3. What are the ways in which copyright and defamation laws could be used as a pretense to suppress creative and socially charged uses of AI-enabled media? Are there particular guardrails or anticipatory legal measures that could be taken to prevent this?

4. How can democratic societies avoid the weaponization of "deepfake panic" and "the liar's dividend" as well as governmental or commercial policies that overly restrict satire as well as truthful journalism and fact-finding?