NORTH CAROLINA JOURNAL OF LAW & TECHNOLOGY
VOLUME 25, ISSUE 2: DECEMBER 2023

NO NEED TO REINVENT THE WHEEL: WHY THE U.S. SHOULD IMPLEMENT CO-REGULATORY MECHANISMS TO REGULATE DEEPFAKE CONTENT ON INTERNET PLATFORMS

Abe Loven[*]

"Most people, in fact, will not take trouble in finding out the truth, but are much more inclined to accept the first story they hear."
—Thucydides[1]

Artificial intelligence systems create hyper-realistic fake media—deepfakes—that are completely indistinguishable from authentic media. As a result, people cannot believe the things they see or hear; it is impossible to determine whether the subject-matter portrayed in any medium is real or fake. This is the beginning of an information apocalypse.

The U.S. has done nothing to deal with deepfakes, but it must do something—and it must do it fast. This Article argues the U.S. should adopt legislation modeled off a co-regulatory system developed by the EU containing two main parts: (1) a broadly worded statute requiring social media platforms to implement best practices for detecting and labelling deepfakes, and (2) a regularly updated voluntary industry code specifying how platforms can demonstrate compliance with the statute.

TABLE OF CONTENTS

[*] J.D. Candidate, University of North Carolina School of Law, 2025.

[1] THUCYDIDES, HISTORY OF THE PELOPONNESIAN WAR 47 (Rex Warner trans., Penguin Books, 1972).

I. INTRODUCTION

It was 8:42 a.m. on May 22, 2023, and a Twitter (now known as X) user had just reported alarming news: the Pentagon was on fire.[2] The post included a photo of a massive cloud of black smoke billowing out from the lawn next to a building.[3] At 10:03 a.m., a Russian news outlet posted the image to its Twitter account,[4] and the post went viral—including among investors.[5] Stocks moved

[2] Emmanuelle Saliba, *How Verified Accounts Helped Make Fake Images of a Pentagon Explosion Go Viral*, ABC NEWS (May 23, 2023, 7:59 PM), https://abcnews.go.com/US/verified-accounts-helped-make-fake-images-pentagon-explosion/story?id=99541361 [https://perma.cc/GG2L-AV4K].

[3] Nick Waters (@N_Waters89), X (May 22, 2023, 10:19 AM), https://twitter.com/N_Waters89/status/1660651721075351556 [https://perma.cc/4TRW-H3GN] (showing edited screenshots of the original posts, which have since been removed).

[4] Saliba, *supra* note 2.

[5] Philip Marcelo, *Fact Focus: Fake Image of Pentagon Explosion Briefly Sends Jitters Through Stock Market*, ASSOCIATED PRESS (May 23, 2023, 2:02 PM), https://apnews.com/article/pentagon-explosion-misinformation-stock-

immediately, and a sell-off began.[6] But at 10:09 a.m., authorities shared an unexpected development: there in fact was not a fire at the Pentagon.[7] The panic settled, and the stock market rebounded.[8]

What just happened? The truth is, there never was a fire at the Pentagon; the viral image was a deepfake created by artificial intelligence ("AI").[9] Nevertheless, in spite of its poor quality,[10] this one viral image duped countless social media users[11] and even news outlets.[12] It also caused the S&P 500 to decrease 30 points, erasing billions of dollars in wealth instantly.[13]

With their increasing deceptiveness and numerosity, deepfakes are becoming more and more harmful to society.[14] "[D]eepfakes can

market-ai-96f534c790872fde67012ee81b5ed6a4 [https://perma.cc/QL3J-M9TX].

[6] Saliba, *supra* note 2 (noting that stocks and other investments "moved in ways that typically occur when fear enters the market," including sell-offs of stocks and purchases of U.S. Treasury Bonds and gold).

[7] *Id.* (timeline included in video on website).

[8] Andrew Ross Sorkin et al., *An A.I.-Generated Spoof Rattles the Markets*, N.Y. TIMES: DEALBOOK NEWSL. (May 23, 2023), https://www.nytimes.com/2023/05/23/business/ai-picture-stock-market.html [https://perma.cc/6W5U-LYF9].

[9] Shannon Bond, *Fake Viral Images of an Explosion at the Pentagon Were Probably Created by AI*, NPR (May 23, 2023, 6:19 PM), https://www.npr.org/2023/05/22/1177590231/fake-viral-images-of-an-explosion-at-the-pentagon-were-probably-created-by-ai [https://perma.cc/5M7B-ALHN].

[10] Bill McCarthy, *Fake Pentagon Explosion Image Spreads Online*, AFP: FACT CHECK (May 22, 2023, 4:03 PM), https://factcheck.afp.com/doc.afp.com.33FV4BU [https://perma.cc/KAE3-TS44] (noting the "columns on the building are mismatched sizes," a lamppost "appears disjointed," and the sidewalk "seems to blend in with the street, grass and fence").

[11] Saliba, *supra* note 2 (citing a study that found at least "3,785 [Twitter] accounts had mentioned the falsehoods, including dozens of accounts verified with Twitter's blue ribbon").

[12] Sorkin et al., *supra* note 8 (noting that the news outlets RT and ZeroHedge both shared the photo); McCarthy, *supra* note 10 (noting that an Indian television station aired the photo in a breaking news broadcast).

[13] Davey Alba, *How a Fake AI Photo of a Pentagon Blast Wiped Billions Off Wall Street*, SYNDEY MORNING HERALD (May 24, 2023, 7:49 AM), https://www.smh.com.au/business/markets/how-a-fake-ai-photo-of-a-pentagon-blast-wiped-billions-off-wall-street-20230524-p5daqo.html [https://perma.cc/TFW4-TV5X].

[14] Among other things, deepfakes can be used for bank fraud; destroying a person's image and credibility; harassing and humiliating people and

create an environment in which nothing is believed, causing a breakdown in trust associated with social organizations, government entities, religious groups, and almost everything else."[15] While some deepfakes seem relatively harmless, like an image of Pope Francis wearing a puffer jacket,[16] others have the potential to be very harmful, such as a video of Ukrainian President Volodymyr Zelenskyy ordering troops to surrender to Russia.[17]

Once deepfakes are published, it is very difficult—if not impossible—for consumers to tell whether the content is real or fake. Humans are largely unable to tell deepfakes apart from real media,[18] and even AI-powered deepfake detectors fail the task[19]—

organizations; extorting and blackmailing; making fraudulent documents; spreading fake news; influencing public opinion; inciting acts of violence; and polarizing societal groups. Tom Olzak, *Adversarial AI: What It Is and How To Defend Against It?*, SPICEWORKS (June 28, 2022), https://www.spiceworks.com/tech/artificial-intelligence/articles/adversarial-ai-attack-tools-techniques/ [https://perma.cc/48NU-HL97].

[15] *Id.*

[16] James Vincent, *The Swagged-Out Pope is an AI Fake — and an Early Glimpse of a New Reality*, THE VERGE (Mar. 27, 2023, 9:25 AM), https://www.theverge.com/2023/3/27/23657927/ai-pope-image-fake-midjourney-computer-generated-aesthetic [https://perma.cc/F6FY-XQBV].

[17] Bobby Allyn, *Deepfake Video of Zelenskyy Could be "Tip of the Iceberg" in Info War, Experts Warn*, NPR (Mar. 16, 2022), https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia [https://perma.cc/FL2U-LATV]; *see also* William Corvey, *Media Forensics (MediFor) (Archived)*, DEF. ADVANCED RSCH. PROJECTS AGENCY, https://www.darpa.mil/program/media-forensics [https://perma.cc/R9RC-4LS4] (last visited Nov. 18, 2023) ("While many manipulations are benign, performed for fun or for artistic value, others are for adversarial purposes, such as propaganda or misinformation campaigns.").

[18] Klaire Somoray et al., *Providing Detection Strategies to Improve Human Detection of Deepfakes: An Experimental Study*, COMPUTS. HUM. BEHAV., Dec. 2023, at 4 (finding that study participants' determinations of whether a video was real or a deepfake were, on average, accurate only about 60% of the time).

[19] Ben Dickson, *AI Tools Can Detect Deepfakes, But for How Long?*, PC MAG (Aug. 29, 2019), https://www.pcmag.com/news/ai-tools-can-detect-deepfakes-but-for-how-long [https://perma.cc/8JDX-KT5F]; Alex O'Brien, *How to Spot an AI Cheater*, BBC: FUTURE NOW (July 20, 2023), https://www.bbc.com/future/article/20230720-how-to-spot-an-ai-cheater-artificial-intelligence-large-language-models [https://perma.cc/G2HB-UMHV]

especially as deepfake generators develop new techniques to avoid detection.[20] Moreover, even if a deepfake generator voluntarily sought to aid detection by embedding identifying markers, this technology has yet to be developed.[21] As a result of these technological limitations, deepfakes are becoming nearly undetectable.

In addition to becoming increasingly undetectable, deepfakes are also becoming easier to create.[22] One journalist, for example, made a deepfake video of himself using just one photograph and a sixty-second audio recording.[23] The technology was cheap and easy to use; he spent only eleven dollars and eight minutes creating the deepfake.[24]

As deepfakes become harder to detect and easier to create, it is vital that policies be adjusted to respond. In the United States ("U.S."), current policies do not require online services to do anything to mollify the harms created by deepfakes hosted on their platforms.[25] In fact, one law—Section 230 of the Communications Decency Act ("Section 230"),[26] which grants online services strong

(noting that "technology alone won't be enough to respond" to the rise in AI-generated text).

[20] Ioana Patringenaru, *Deepfake Detectors Can Be Defeated, Computer Scientists Show for the First Time*, UC SAN DIEGO: TODAY (Feb. 8, 2021), https://today.ucsd.edu/story/defeating_deepfake_detectors [https://perma.cc/8V42-WXT7].

[21] Shannon Bond, *AI-Generated Deepfakes are Moving Fast. Policymakers Can't Keep Up*, NPR (April 27, 2023, 6:11 PM), https://www.npr.org/2023/04/27/1172387911/how-can-people-spot-fake-images-created-by-artificial-intelligence [https://perma.cc/3455-47Y2].

[22] Shannon Bond, *It Takes a Few Dollars and 8 Minutes to Create a Deepfake. And That's Only the Start*, NPR (Mar. 23, 2023, 5:00 AM), https://www.npr.org/2023/03/23/1165146797/it-takes-a-few-dollars-and-8-minutes-to-create-a-deepfake-and-thats-only-the-sta [https://perma.cc/YK9A-6QFP] ("Concerns about deepfakes have been around for years. What's different now is technology has advanced and become accessible to anybody with a smartphone or computer.").

[23] *Id.* (citing Ethan Mollick (@emollick), X (Feb 10, 2023, 9:21 AM), https://twitter.com/emollick/status/1624050928092340238 [https://perma.cc/G2BW-6QUM]).

[24] *Id.*

[25] *See infra* Part II-C.

[26] 47 U.S.C. § 230.

immunities—enables online platforms to host deepfake content without fear of even being exposed to private tort actions.[27] As a result, online services in the U.S. can host deepfake content with impunity, despite the immense social harm this causes. This Article argues that the U.S. should act by implementing a co-regulatory scheme—modeled on one recently adopted by the European Union ("EU")—to foster the creation of codes requiring large online platforms to adopt best practices for the screening and labeling of deepfake content.

This Article will proceed in several parts. Part II will provide an overview of deepfakes and various deepfake technologies, including deepfake generators, deepfake detectors, and provenance authenticators; it will also discuss current deepfake policy and the harms presented by deepfakes. Part III will provide an analysis of the EU legislation mentioned above, and Part IV will assess this legislation. Part V will analyze the effectiveness and legitimacy of co-regulatory mechanisms like the one described in the EU legislation. Lastly, Part VI details why removing Section 230 immunity is not a desirable alternative to pursuing the action proposed in this Article—that is, to pass legislation modeled off the EU's.

II. AN OVERVIEW OF DEEPFAKES & DEEPFAKE TECHNOLOGY

While the term "deepfake" is vague and "still in flux,"[28] it generally refers to a photo, video, or audio recording "that seems real but has been manipulated with artificial intelligence technologies."[29] Like other types of synthetic media, deepfakes do not depict reality, but rather "depict made-up events, sometimes

[27] *Id*; *see infra* Part VI.

[28] James Vincent, *Why We Need a Better Definition of 'Deepfake'*, THE VERGE (May 22, 2018, 2:53 PM), https://www.theverge.com/2018/5/22/17380306/deepfake-definition-ai-manipulation-fake-news [https://perma.cc/3EKD-VLWG]; *accord* Jia Wen Seow et al., *A Comprehensive Overview of Deepfake: Generation, Detection, Datasets, and Opportunities*, 513 NEUROCOMPUTING 351, 351 ("The deepfake definition has been broadened over the years.").

[29] U.S. GOV'T ACCOUNTABILITY OFF., GAO-20-379SP, SCIENCE & TECH SPOTLIGHT: DEEPFAKES 1 (2020).

quite realistically."[30] What differentiates deepfakes from other types of synthetic media is that they are created using AI.[31]

To recognize the impact deepfakes have on society, it is important to understand the technology used to create deepfakes; the effectiveness and availability of deepfake generators, deepfake detectors, and content authentication technology; the scope of current domestic deepfake regulation; and the harms presented by deepfakes. This Article considers each in turn.

A. Deepfake Technology

Deepfakes are created using a type of AI called deep learning.[32] In fact, the word "deepfake" is a portmanteau of the terms "deep learning" and "fake."[33] Deep learning is best thought of as a subset of machine learning, which itself is a subset of AI.[34] Generally, machine learning emulates human learning in order to enable an AI system to "adapt to uncertain or unexpected conditions."[35] Because machine learning is a broad term, a single machine learning system might employ one of many different techniques.[36] Deep learning,

[30] Jon Bateman, Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios 4 (July 2020) (unpublished manuscript) (on file with Carnegie Endowment for Int'l Peace).

[31] *See* Vincent, *supra* note 28; Nate Lanxon, *Deepfakes: What Are Fake AI Video Dangers, and How to Spot Them*, BL (Sept. 20, 2023, 1:21 PM), https://www.bloomberglaw.com/product/blaw/bloombergterminalnews/bloomberg-terminal-news/S0ZF3FDWLU68 [https://perma.cc/33TG-U3A6] ("While manipulation of digital files using Photoshop and other apps is nothing new, deepfakes are accomplished using a form of AI.").

[32] Bateman, *supra* note 30, at 4.

[33] Nick Barney, *Definition: Deepfake AI (Deep Fake)*, TECHTARGET (last updated Mar. 2023), https://www.techtarget.com/whatis/definition/deepfake [https://perma.cc/AP8X-ZBWL].

[34] John Paul Mueller & Luca Mueller, *What is Deep Learning?*, FOR DUMMIES (last updated July 16, 2019), https://www.dummies.com/article/technology/information-technology/ai/machine-learning/what-is-deep-learning-262737/ [https://perma.cc/HR3K-CWB5].

[35] *Id.*

[36] *Id.* ("Machine learning relies on different paradigms such as using statistical analysis, finding analogies in data, using logic, and working with symbols.").

though, refers specifically to machine learning systems that use "artificial neural networks to learn from data."[37]

An artificial neural network ("ANN") is an "adaptive system that learns by using interconnected . . . neurons in a layered structure that resembles a human brain."[38] As a result of their brain-like structures, ANNs "can learn and make intelligent decisions on their own."[39] Within an ANN, neurons are arranged in layers[40]—often hundreds of layers in modern deep learning systems.[41] Regardless of which layer it is located in, each neuron receives data inputs from the previous layer, processes that data, and delivers an output to neurons in the next layer.[42] Each layer within the ANN is responsible for detecting a broad pattern from the data set, and each neuron is responsible for detecting a specific feature of the data.[43] After receiving an output from the previous layer, the present layer will further "refine and optimize" the data, thereby "building upon" the previous layer's work.[44] Or, to state it simply, because each layer's neurons "get more and more specific" and receive increasingly processed data,[45] an ANN "learns more and more about the data as

[37] *What is Deep Learning?*, GOOGLE, https://cloud.google.com/discover/what-is-deep-learning [https://perma.cc/XR5W-JWUT] (last visited Oct. 20, 2023).

[38] *What is a Neural Network?*, MATHWORKS, https://www.mathworks.com/discovery/neural-network.html [https://perma.cc/928G-7RYL] (last visited Nov. 18, 2023).

[39] Grace Shao, *What "Deepfakes" Are and How They May Be Dangerous*, CNBC (last updated Jan. 17, 2020, 2:47 AM), https://www.cnbc.com/2019/10/14/what-is-deepfake-and-how-it-might-be-dangerous.html [https://perma.cc/ED7R-4NHP].

[40] saumyasaxena2730, *Introduction to Deep Learning*, GEEKSFORGEEKS (Apr. 14, 2023), https://www.geeksforgeeks.org/introduction-deep-learning/ [https://perma.cc/9YUX-C4VJ].

[41] *What is a Neural Network?*, *supra* note 38; Mueller & Mueller, *supra* note 34 ("[T]he term *deep* is appropriate; it refers to the large number of layers potentially used for analysis.").

[42] saumyasaxena2730, *supra* note 40.

[43] *What is Deep Learning?*, *supra* note 37; *see* saumyasaxena2730, *supra* note 40.

[44] *What is Deep Learning?*, IBM, https://www.ibm.com/topics/deep-learning [https://perma.cc/7RY3-8EDD] (last visited Nov. 15, 2023).

[45] Ryan Thelin, *What is Deep Learning? A Tutorial for Beginners*, EDUCATIVE (Nov. 10, 2020), https://www.educative.io/blog/deep-learning-beginner-tutorial [https://perma.cc/CD2C-PBUU].

it moves from one [layer] to another."[46] For example, in a hypothetical ANN meant to perform image recognition, the first layer of neurons might identify edges while the second layer identifies shapes and the third layer identifies objects.[47] This system of refining and optimizing input data allows the ANN to recognize patterns and eventually classify objects.[48]

An important feature of ANNs is that they can learn from their own work. Just like human brains, ANNs learn by processing huge amounts of data[49] and rearranging the relations between their own neurons.[50] The ANN calculates errors in its outputs[51] and then adjusts "the weights on the connections between the [neurons] . . . so that the [ANN] can better classify" data inputs in the future.[52] This revision process enables the ANN to teach itself without the aid of a human programmer, allowing it to learn much faster and more accurately.[53] Once an ANN has been trained on a large set of training data, it can be used to make predictions on any new data it receives.[54] To put it another way, the deep learning process enables an ANN "to do what comes naturally to humans: learn by example."[55]

[46] saumyasaxena2730, *supra* note 40.

[47] *What is Deep Learning?*, *supra* note 37.

[48] *What is a Neural Network?*, *supra* note 38.

[49] Mueller & Mueller, *supra* note 34.

[50] *What is Deep Learning?*, ORACLE, https://www.oracle.com/artificial-intelligence/machine-learning/what-is-deep-learning/ [https://perma.cc/597C-HTXG] (last visited Nov. 18, 2023); *What is Deep Learning?*, *supra* note 37.

[51] *What is Deep Learning?*, supra note 44.

[52] *What is Deep Learning?*, *supra* note 37.

[53] Thelin, *supra* note 45 ("[Deep learning] also increases accuracy because the algorithm can detect all features rather than just those recognizable to the human eye.").

[54] *What is Deep Learning?*, *supra* note 37.

[55] *What is Deep Learning? 3 Things You Need to Know*, MATHWORKS, https://www.mathworks.com/discovery/deep-learning.html [https://perma.cc/UZ4N-LYJ4] (last visited Nov. 18, 2023); Ed Stacey, *Can Startups Solve the Threat of Deepfakes?*, FORBES (Oct. 28, 2019, 12 :21 PM), https://www.forbes.com/sites/edstacey/2019/10/28/can-startups-solve-the-threat-of-deepfakes/ ?sh=722c5fc125c0 [https://perma.cc/3C23-G2GQ] ("In simple terms, [a deep learning system] learns from real data (audio, visual or textual information) to produce original content.").

Deep learning systems have been used in a wide range of industries for myriad functions, including automated driving, medical research, language translation, and stock trading.[56] They have performed at a high level on a wide variety of tasks, such as passing the multi-state bar exam in the 90th percentile of actual test takers.[57] A deep learning system even solved a problem that had vexed the medical community for the last half-century: how to predict a protein's three-dimensional shape using only its sequence of amino acids.[58] Because of the breadth of industries AI has revolutionized, President Biden has called AI "the most consequential technology of our time."[59]

The same deep learning systems that have enabled these revolutionary advancements have, however, also enabled the creation of hyper-realistic deepfakes.[60]

[56] *What is Deep Learning?*, *supra* note 44; *What is Deep Learning?*, *supra* note 50.

[57] Karen Sloan, *Bar Exam Score Shows AI Can Keep Up with "Human Lawyers," Researchers Say*, REUTERS (Mar. 15, 2023, 2:17 PM), https://www.reuters.com/technology/bar-exam-score-shows-ai-can-keep-up-with-human-lawyers-researchers-say-2023-03-15/ [https://perma.cc/54V8-UGA3].

[58] Will Douglas Heaven, *AI for Protein Folding*, MIT TECH. REV. (Feb. 23, 2022), https://www.technologyreview.com/2022/02/23/1044957/ai-protein-folding-deepmind/ [https://perma.cc/5ZYP-USEJ].

[59] *Remarks by President Biden and Vice President Harris on the Administration's Commitment to Advancing the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, WHITE HOUSE (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/10/30/remarks-by-president-biden-and-vice-president-harris-on-the-administrations-commitment-to-advancing-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ [https://perma.cc/4WQ5-XDV2] ("AI is all around us. Much of it is making our lives better.").

[60] Geraint Rees, *Here's How Deepfake Technology Can Actually Be a Good Thing*, WORLD ECON. FORUM (Nov. 25, 2019), https://www.weforum.org/agenda/2019/11/advantages-of-artificial-intelligence/ [https://perma.cc/54ZX-RXDK] ("While questions are rightly being asked about the consequences of deepfake technology, it is important that we do not lose sight of the fact that artificial intelligence (AI) can be used for good, as well as ill."). As recently recognized by the White House itself,

> [a]rtificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve

B. The Availability and Effectiveness of Deepfake Generators and Detectors

"In one of the first deepfakes to go viral,"[61] comedian and filmmaker Jordan Peele created a public service announcement on the dangers of deepfake technology by delivering a speech in the guise of President Barack Obama.[62] This video is surprisingly realistic despite having been created in 2018[63]—ancient by the evolutionary standards of AI, which has been doubling in computational power every few months.[64]

While fake media is not a new concept, generative AI has recently made it much easier to create and harder to detect.[65] As AI

> urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud . . . and disinformation . . . and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor demands a society-wide effort that includes government, the private sector, academia, and civil society.

Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 Fed. Reg. 75191, 75191 (Oct. 30, 2023).

[61] Ryan S. Gladwin, *Meet the Company That's Helping Governments Detect AI Deepfakes*, EMERGE (Oct. 18, 2023), https://decrypt.co/202181/deepfake-detection-deepmedia-ai-government [https://perma.cc/ND4L-H4KX].

[62] Buzzfeed, *You Won't Believe What Obama Says In This Video!*, YOUTUBE (Apr. 17, 2018), https://www.youtube.com/watch?v=cQ54GDm1eL0 [https://perma.cc/W5LN-CX79].

[63] Gladwin, *supra* note 61.

[64] Jeff Brown, *Growth in Artificial Intelligence Is Beyond Exponential*, LEGACY RSCH. GRP.: DAILY CUT (Oct. 22, 2020), https://www.legacyresearch.com/the-daily-cut/growth-in-artificial-intelligence-is-beyond-exponential/ [https://perma.cc/ZQ25-36XD] (finding that "AI computing power has doubled every 3.4 months" since 2012, "dwarf[ing] Moore's Law").

[65] To explain:

> While the concept of disinformation has been around for centuries, recently, those wishing to spread it have taken advantage of social media and easy-to-use editing technologies to do so at an alarming pace. . . .
>
> And as artificial intelligence (AI) continues to advance, it will become even easier to manipulate all types of media—and even more difficult to detect manipulation when it occurs. Think altered photos, videos, audio—all with the intent to mislead.

Dana Rao, *Deepfake Task Force: The Danger of Disinformation Needs a New Collaboration*, ADOBE: BLOG (Aug. 23, 2021), https://blog.adobe.com/en/publis technologies have grown at exponential rates, the deepfake technology scene has been changing rapidly. This section considers the current state of three types of technologies related to deepfakes: deepfake generators, which create deepfakes; deepfake detectors, which identify deepfakes; and content authenticators, which identify authentic media.

1. Deepfake Generators

In recent years, rates of deepfake media creation have risen exponentially. For example, it is predicted that more deepfake pornographic videos "will have been produced in 2023 than the total number of every other year combined."[66] One deepfake detection company predicts that roughly 500,000 video and voice deepfakes will have been shared on social media globally by the end of 2023.[67] In the last year alone, more than 15 billion deepfake images—the equivalent of one-third of the entire amount of images ever uploaded to Instagram—have been created using simple text-to-image algorithms.[68]

The proliferation of deepfakes is enabled in part by the fact that deepfake-generating technology ("deepfake generators" or "generators") has become significantly less expensive. Whereas

h/2021/08/23/deepfake-task-force-danger-of-disinformation-needs-new-collaboration [https://perma.cc/B9E5-9NW4].

[66] Matt Burgess, *Deepfake Porn Is Out of Control*, WIRED (Oct. 16, 2023), https://www.wired.com/story/deepfake-porn-is-out-of-control/ [https://perma.cc/B6N7-G4PE] (finding that of the at least 244,625 deepfake pornographic videos on top deepfake porn websites as of October of 2023, 113,000 of them were uploaded in the first nine months of 2023).

[67] Alexandra Ulmer & Anna Tong, *Deepfaking It: America's 2024 Election Collides with AI Boom*, REUTERS (May 30, 2023), https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-30/ [https://perma.cc/ZG4K-SH4H]. Note that the 500,000 video and voice deepfakes predicted to be shared by the end of 2023 does not even include deepfake images.

[68] Alina Valyaeva, *AI Has Already Created as Many Images as Photographers Have Taken in 150 Years. Statistics for 2023*, EVERYPIXEL J. (Aug. 15, 2023), https://journal.everypixel.com/ai-image-statistics?fbclid=IwAR3Su07k8NJPE4Xd3e2x9VFgbhYrX18FESM4HQBuUac4x7NTqduB7iyOJRk [https://perma.cc/7JP8-RDKH] ("To put this in perspective, it took photographers 150 years, from the first photograph taken in 1826 until 1975, to reach the 15 billion mark.").

creating a deepfake voice recording, for example, used to cost over $10,000, it now costs just a few dollars.[69] Generators have also become exceedingly easy to use.[70] The issue is further compounded by the fact that deepfakes are also becoming much more convincing[71]—a trend that will continue as deep learning systems become able to process more data.[72]

[69] Ulmer & Tong, *supra* note 67.

[70] Consider:

> Artificial intelligence allows virtually anyone to create complex artworks, like those now on exhibit at the Gagosian art gallery in New York, or lifelike images that blur the line between what is real and what is fiction. Plug in a text description, and the technology can produce a related image — no special skills required.

Tiffany Hsu & Steven Lee Myers, *Can We No Longer Believe Anything We See?*, N.Y. TIMES (Apr. 8, 2023), https://www.nytimes.com/2023/04/08/business/media/ai-generated-images.html [https://perma.cc/9XKC-EECG].

[71] To summarize:

> Artificial intelligence has improved greatly over the past year, allowing nearly anyone to create a persuasive fake by entering text into popular A.I. generators that produce images, video or audio — or by using more sophisticated tools. When a deepfake video of President Volodymyr Zelensky of Ukraine was released in the spring of 2022, it was widely derided as too crude to be real; a similar faked video of President Vladimir V. Putin of Russia was convincing enough for several Russian radio and television networks to air it [in June 2023].

Tiffany Hsu & Stuart A. Thompson, *A.I. Muddies Israel-Hamas War in Unexpected Way*, N.Y. TIMES (Oct. 30, 2023), https://www.nytimes.com/2023/10/28/business/media/ai-muddies-israel-hamas-war-in-unexpected-way.html [https://perma.cc/23VX-8SPX].

[72] Lanxon, *supra* note 31 ("The bigger the library of content a deep-learning algorithm is fed with, the more realistic the phony can be."). To understand how powerful this phenomenon will soon become, consider the following:

> Apple recorded 10 to 20 hours of speech to create Siri. Actor-director Jordan Peele made a minute-long deepfake in 2018 appearing to show former US President Barack Obama [*supra* notes 61–63 and accompanying text] . . . . Peele imitated Obama's voice and used 56 hours of sample video recordings of the former president. Those sample sizes are infinitesimal compared with what AI companies are now applying their new tools to: the entire corpus of material freely available on the web, from YouTube to Wikipedia to stock image libraries. The simplest way to understand the difference this makes is to refer back to that viral Obama clip: A person had to manipulate a video that already existed and provide a real vocal performance; today, someone can

While several of the largest deepfake generators have tried to implement safeguards on their services to prevent them from being used for certain tasks, many of these have proven ineffective. For example, the owners of the popular image generator DALL-E claim to have put on safeguards preventing the program from generating images of "public figures."[73] Although this safeguard succeeded in preventing users from generating images of Presidents Joe Biden and Donald Trump, it failed to prevent the generation of images depicting other notable figures like Vice President Mike Pence,[74] exemplifying the difficulty generators face in creating effective safeguards for their systems.

Because deepfake generators are becoming more effective and accessible to users, it is likely the meteoric rise in deepfake generation will continue over the next several years. As deepfake generators produce an unlimited variety of deepfakes on myriad subjects, their rudimentary "safeguards" will remain inadequate for preventing the creation of deepfakes that are harmful. The situation can already be considered dire, but these factors indicate this is only the tip of the iceberg. The time to act is now—before the problem gets out of control.

2. *Deepfake Detectors*

To help identify deepfakes, many companies are designing AI-powered deepfake detection tools ("deepfake detectors" or "detectors") that, like generators, utilize deep learning systems.[75] Detectors search media content for evidence of manipulation usually not discernable to a human, such as resolution inconsistencies

> simply ask a machine to create a video of the former president and it will appear.

Id.

[73] Ulmer & Tong, *supra* note 67.

[74] *Id.*

[75] John Villasenor, *Artificial Intelligence, Deepfakes, and the Uncertain Future of Truth*, BROOKINGS INST. (Feb. 14, 2019), https://www.brookings.edu/articles/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/ [https://perma.cc/TP3P-EH85].

between a base video and a spliced-on face within a deepfake video.[76]

While this technology holds promise, there are currently many problems with deepfake detectors. One is that they work well only when the media being processed is of a high resolution[77]—a major problem given how uncommon it is for media to be high-resolution.[78] As a result, sometimes detectors fail "even when an image is obviously fake."[79] A more systemic problem is that the development of detectors lags behind the development of generators,[80] and no comprehensive deepfake detection system exists.[81]

[76] *Id; accord* Matthew Hutson, *Detection Stays One Step Ahead of Deepfakes—For Now: The Spread of AI-Generated Content is Keeping the Tech Designed to Spot it on its Toes*, INST. OF ELEC. AND ELECS. ENG'RS: SPECTRUM (Mar. 6, 2023), https://spectrum.ieee.org/deepfake [https://perma.cc/G6TF-4TDX] (describing a deepfake video detector that "studies color changes in faces to infer blood flow").

[77] Stuart A. Thompson & Tiffany Hsu, *How Easy Is It to Fool A.I.-Detection Tools?*, N.Y. TIMES (June 28, 2023), https://www.nytimes.com/interactive/2023/06/28/technology/ai-detection-midjourney-stable-diffusion-dalle.html [https://perma.cc/V2JE-C543] (finding that deepfake detectors "struggled" to detect deepfake images after "just a bit of grain was introduced" to the photos).

[78] "[O]ne drawback with the current A.I. detectors" is that

> [t]hey tend to struggle with images that have been altered from their original output or are of low quality, according to Kevin Guo, a founder and the chief executive of Hive, an image-detection tool.
>
> When A.I. generators like Midjourney create photorealistic artwork, they pack the image with millions of pixels, each containing clues about its origins. "But if you distort it, if you resize it, lower the resolution, all that stuff, by definition you're altering those pixels and that additional digital signal is going away," Mr. Guo said. . . .
>
> . . . .
>
> . . . Such shortfalls can undermine the potential for A.I. detectors to become a weapon against fake content. As images go viral online, they are often copied, resaved, shrunken or cropped, obscuring the important signals that A.I. detectors rely on.

Id.

[79] *Id.* (showing that a low-resolution deepfake photo of "a towering, Yeti-like beast next to a quaint couple" fooled the five top deepfake detectors).

[80] Corvey, *supra* note 17 ("[T]he digital imagery playing field . . . currently favors the manipulator.").

[81] *Id.* ("The forensic tools [for deepfake detection] used today lack robustness and scalability, and address only some aspects of media authentication; an

The gap between detectors and generators is unlikely to close since generators are constantly being improved to craft deepfakes that detectors cannot detect.[82] In fact, several companies incorporate detectors into their generators so the system can catch its own errors and fix them.[83] When the system's detector "detect[s] flaws in the forgery," its generator can respond by developing "improvements addressing the flaws."[84] Because deep learning systems learn by analyzing their own mistakes, generators benefit when detectors identify the mistakes to analyze.[85] "By harnessing the potential of pitting [ANNs] against each other," this technology has "revolutionized" deepfake generation.[86] As a result, deepfakes made by generators employing this process can be highly effective, evading over 99% of detection under ideal circumstances.[87]

The competition between generators and detectors is creating a "deepfakes arms race" in which "even the best detection methods

end-to-end platform to perform a complete and automated forensic analysis does not exist.").

[82] *Forbes* identifies the problem as such:

> There is a fundamental issue with using [AI-powered deepfake detection technology]. The more widely the algorithms designed to detect certain tell-tale characteristics are used, the quicker they become outdated, as developers creating deepfakes will always be able to find a way to adjust to changes. This creates an unsustainable virus/anti-virus dynamic–because, like antivirus software, it cannot guarantee permanent protection as new viruses are created every day.

Stacey, *supra* note 55; *accord* Thompson & Hsu, *supra* note 77 ("The generators are designed to be able to fool a detector.").

[83] Stacey, *supra* note 55; Hutson, *supra* note 76 ("Synthetic-media creation and detection is an arms race, one in which each side builds on the other."); Thompson & Hsu, *supra* note 77 ("Every time somebody builds a better generator, people build better discriminators, and then people use the better discriminator to build a better generator.").

[84] Shao, *supra* note 39. For a longer explanation on the process, see DEP'T OF HOMELAND SEC., INCREASING THREAT OF DEEPFAKE IDENTITIES 12 (2021).

[85] Hutson, *supra* note 76 ("Given a new detection method, someone can often *train* a generation algorithm to become better at fooling it." (emphasis added)).

[86] Nils Köbis et al., *Fooled Twice: People Cannot Detect Deepfakes But Think They Can*, 24 ISCIENCE 1, 2 (2021).

[87] Patringenaru, *supra* note 20; *see also* Hutson, *supra* note 76 ("Given a new detection method, someone can often train a generation algorithm to become better at fooling it.").

will often lag behind the most advanced creation methods."[88] This problem is further compounded by the fact that there are far fewer people working to improve detectors than there are working to improve generators.[89] Therefore, while it is likely that generators and detectors "will become locked in a perpetual back-and-forth as both sides become more sophisticated,"[90] the "capacity to generate deepfakes is proceeding much faster than the ability to detect them."[91]

Since detectors identify deepfakes better than humans can, they should be utilized whenever possible,[92] and legislation should be crafted to ensure this happens quickly and effectively. Nevertheless, because deepfake detectors are currently unreliable,[93] some industry

[88] Villasenor, *supra* note 75. However, at least one deepfake detection company believes it can stay ahead of the curve by continuously training its deepfake detector system with the newest deepfake content created by its own state-of-the-art deepfake generator program:

> "Our DubSync platform is essentially a deepfake generator. We have to build a generator in order to know what a good deepfake is," [Emma] Brown [the company's cofounder] explained. "And that's what feeds our deepfake detection."
>
> Brown claims that DubSync deepfake generation stays "about six to 12 months ahead of anyone else," in order to ensure that the firm has the most cutting-edge data to train from. This is done with the aim of preventing bad actors from creating deepfakes that are more advanced than their AI can detect. But it's a constant battle to keep that lead.
>
> "It's a cat-and-mouse game, for sure."

Gladwin, *supra* note 61. "This is done with the aim of preventing bad actors from creating deepfakes that are more advanced than their AI can detect. But it's a constant battle to keep that lead." *Id.*

[89] William A. Galston, *Is Seeing Still Believing? The Deepfake Challenge to Truth in Politics*, BROOKINGS INST. (Jan. 8, 2020), https://www.brookings.edu/articles/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/ [https://perma.cc/XB57-33Z5] ("[I]dentifying fake media has long received less attention, funding, and institutional support than creating it.").

[90] Hutson, *supra* note 76.

[91] Galston, *supra* note 89.

[92] Villasenor, *supra* note 75 ("[T]echnological solutions will have no impact when they aren't used.").

[93] Thompson & Hsu, *supra* note 77. For example, in the deepfake of the Pentagon referenced at the beginning of this Article, four of the five leading deepfake detectors mistook this image as not being a deepfake, *id.* (scroll to the

leaders are pursuing alternative means of differentiating deepfakes from real content.[94]

3. *Provenance Technology for Real Media*

Separate from detection methods, deepfakes can also be combatted by technology assuring the authenticity of real media.[95] Various systems are being developed to embed into a media file information telling the consumer the provenance—especially the creation and edit history—of the media file.[96] These systems are fundamentally different from deepfake detectors in that they do not rely on detection at all; rather, they rely on validation—whether that be validating the authenticity of real media[97] or the artificiality of deepfakes.[98]

second image under "A selection of test results")—even though the image was of poor quality, *see* McCarthy, *supra* note 10.

[94] Deepfake detection tools "provide a false solution to a much more complex and difficult-to-solve problem."

> "Proving what's fake is going to be a pointless endeavor and we're just going to boil the ocean trying to do it," said Chester Wisniewski, an executive at the cybersecurity firm Sophos. "It's never going to work, and we need to just double down on how we can start validating what's real."

Hsu & Thompson, *supra* note 71.

[95] Hutson, *supra* note 76 ("Short-term, . . . we need detection algorithms. Long-term, we also need protocols that establish provenance, perhaps involving watermarks or blockchains.").

[96] *E.g.,* CONTENT AUTHENTICITY INITIATIVE, https://contentauthenticity.org/ [https://perma.cc/GC29-NA6B] (last visited Nov. 22, 2023). To provide a more technical definition, consider how the White House recently defined "watermarking," a term used to refer to provenance authentication technology:

> The term "watermarking" means the act of embedding information, which is typically difficult to remove, into outputs created by AI — including into outputs such as photos, videos, audio clips, or text — for the purposes of verifying the authenticity of the output or the identity or characteristics of its provenance, modifications, or conveyance.

Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 Fed. Reg. 75191, 75196 (Oct. 30, 2023).

[97] Rao, *supra* note 65 ("[R]ather than taking on the impractical (and, quite frankly, impossible) task of catching every bad actor [with deepfake detectors], a provenance-based solution creates a place for good actors to be trusted.").

[98] Xuandong Zhao et al., Invisible Image Watermarks Are Provably Removable Using Generative AI 1 (Aug. 6, 2023) (unpublished manuscript) (on file with

Because of its nascency, provenance authentication technology has not yet become reliable.[99] However, such technology, if developed, could potentially play a major role in ensuring real media is distinguishable from deepfake media. Beyond working in conjunction with detection systems, authentication technology could work independently, thereby reducing—maybe even eliminating—the need for detectors.[100] The promise of authentication technology is exemplified by the fact that the White House recently stated it would "help develop effective labeling and content provenance mechanisms, so that Americans are able to determine when content is generated using AI and when it is not."[101] Congress should act on this promise by passing legislation requiring online platforms to adopt provenance authentication technology.

* * *

Deepfakes are becoming increasingly rampant for several reasons. Chief among them is the rapid advancement of generator technology, which can make increasingly higher quality deepfakes cheaper and faster than ever before. The harms posed by deepfakes are mounting due to the current ineffectiveness of detectors, provenance authenticators, and generator safeguards. The most glaring problem, though, is the complete lack of U.S. laws mandating the development and adoption of these technologies. The government must act promptly to ensure these systems are developed and implemented.

arXiv) ("[M]ajor AI companies such as Google, Microsoft, Meta, and OpenAI have pledged to add watermarks to the content generated by their AI products.") (emphasis omitted).

[99] *Id.* at 2 (finding that one malicious software was 93-99% effective in removing "a particularly resilient watermark" embedded in deepfake images).

[100] Rao, *supra* note 65 (arguing that provenance-based solutions "provide[] a critical backstop if AI-based detection tools cannot keep up with AI-based creations").

[101] Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 88 Fed. Reg. at 75191. The White House also ordered the Office of Management and Budget to develop recommendations on how federal agencies should implement "reasonable steps to watermark or otherwise label output from generative AI." *Id.* at 75219.

C. *The Scope of Existing Deepfake Policy Solutions in the U.S.*

In the U.S., there is no federal legislation regulating deepfakes, and executive branch actions on AI do very little to address deepfakes. Although the Biden administration issued a sprawling Executive Order[102] on AI in October 2023, this order does nothing to regulate industry behavior regarding deepfakes.[103] The only federal executive action addressing deepfakes is a voluntary agreement[104] among industry leaders. Even this, however, does very little to address deepfakes, stating simply that the signees "agree to develop robust mechanisms, including provenance and/or watermarking systems" to help consumers determine whether a piece of media is a deepfake.[105] These commitments are purely voluntary, and the agreement itself explicitly recognizes the need for enforceable legislation.[106]

[102] *Id.* at 75191.

[103] So far, the White House has simply ordered the Secretary of Commerce to do two things: (1) "submit a report . . . identifying the existing standards, tools, methods, and practices, as well as the potential development of further science-backed standards and techniques" for "authenticating content and tracking its provenance," "labeling synthetic content, such as using watermarking," and "detecting synthetic content"; and (2) "issue guidance to [federal] agencies for labeling and authenticating such content that they produce or publish" in order to "strengthen[] public confidence in the integrity of official United States Government digital content." *Id.* at 75202–03.

[104] *Ensuring Safe, Secure, and Trustworthy AI*, WHITE HOUSE, https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf [https://perma.cc/9CTD-JVM8] (last visited Nov. 18, 2023).

[105] *Id.* at 3. It specifies that the "watermark or provenance data should include an identifier of the service or model that created" the deepfake but does not need to include information identifying the user who created the deepfake. *Id.* "More generally," it adds, "companies making this commitment pledge to work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish" deepfakes from non-deepfake content. *Id.*

[106] The agreement reads:

> These voluntary commitments are only a first step in developing and enforcing binding obligations to ensure safety, security, and trust. Realizing the promise and minimizing the risk of AI will require new laws, rules, oversight, and enforcement.
>
> . . . .

A few states do have laws regulating deepfakes in some manner.[107] However, all these laws focus entirely on just two narrow facets of deepfakery: nonconsensual deepfake pornography and political deepfakes.[108] There is no state law regulating deepfakes comprehensively.

* * *

Nowhere in the U.S. is there a statute or regulation requiring the development or adoption of technology that can detect deepfakes or authenticate real media. In fact, there is no law comprehensively regulating deepfakes at all. This has enabled the unfettered proliferation of deepfakes on online platforms in the U.S., which will have dire consequences.

D. Harms Presented by Deepfakes

As generators grow increasingly complex, deepfakes are becoming nearly indistinguishable from authentic media—especially to humans, who are unable to perceive many of the indicia deepfake detectors rely on. In a 2023 lab study, participants presented with audio clips could correctly identify them as being authentic or deepfake only 70% of the time[109]—a likely overestimation given the ideal circumstances of the study.[110] In a

> . . . These voluntary commitments are . . . designed to advance a generative AI legal and policy regime. Companies intend these voluntary commitments to remain in effect until regulations covering substantially the same issues come into force.

Id. at 1–2.

[107] Isaiah Poritz, *Deepfake Porn, Political Ads Push States to Curb Rampant AI Use*, BL (June 20, 2023, 5:00 AM), https://www.bloomberglaw.com/product/blaw/bloomberglawnews/bloomberg-law-news/X5BNBPGO000000 [https://perma.cc/Q7X6-Q6RR] (showing that, as of June 2023, nine states had some type of law regulating deepfakes, and four others were considering proposed deepfake laws).

[108] *Id*; Natalie Lussier, *Nonconsensual Deepfakes: Detecting and Regulation the Rising Threat to Privacy*, 58 IDAHO L. REV. 353, 371–75 (2022).

[109] Kimberly T. Mai et al., *Warning: Humans Cannot Reliably Detect Speech Deepfakes*, PLOS ONE, Aug. 2, 2023, at 8–9.

[110] Public Library of Science, *Study Shows Speech Deepfakes Frequently Fool People, Even After Training on How to Detect Them*, SCIENCE X NETWORK: PHYS.ORG (Aug. 2, 2023), https://phys.org/news/2023-08-speech-deepfakes-frequently-people.html [https://perma.cc/Y9LC-4HWG] ("Because participants

different study involving videos, participants were only 58% accurate in their determinations despite being told how much of the content would be deepfake.[111] The study showed participants were biased towards assuming each video was authentic[112] and were overconfident in their own detection capabilities.[113] "Taken together, these two biases suggest that people adopt a seeing-is-believing heuristic. Namely, people tend to take videos at face value unless they find clear-cut evidence of it being fake."[114] The deceptiveness of deepfakes is further compounded by the fact that many people do not even know what a deepfake is.[115] As of 2023, only 42% of Americans know what a deepfake is.[116] For people who are not college graduates, the figure is even lower: 28%.[117]

were aware that some of the clips would be deepfakes—and because the researchers did not use the most advanced speech synthesis technology—people in real-world scenarios would likely perform worse than the study participants.").

[111] Köbis et al., *supra* note 86, at 6–7 ("Also, looking at the videos separately reveals that only for 5 of the 16 videos, participants' guesses are significantly more accurate than flipping a coin.").

[112] *Id.* at 8 ("[O]ur participants are very conservative when reporting that a video is a deepfake, i.e., people have a tendency toward guessing authentic.").

[113] *Id.* (noting also that "overconfidence is particularly pronounced among those who perform worse" at detection).

[114] *Id.* at 11 (citation omitted); *accord* Hsu & Thompson, *supra* note 71 ("People will believe anything that confirms their beliefs or makes them emotional."); *AI Image Wins Historic Photography Competition*, ABSOLUTELY AI (Feb. 2, 2023), https://www.absolutelyai.com.au/post/ai-image-wins-historic-photography-competition [https://perma.cc/V4UB-3Q3J] (telling how a deepfake image won a photography competition).

[115] *How To Protect Against Deepfakes – Statistics and Solutions*, IPROOV (Aug. 26, 2022), https://www.iproov.com/blog/deepfakes-statistics-solutions-biometric-protection [https://perma.cc/J86K-ZVTM] ("[I]f people don't know what [deepfakes] are, they are less likely to be prepared to identify when they are being spoofed.").

[116] Olivia Sidoti & Emily A. Vogels, *What Americans Know About AI, Cybersecurity and Big Tech*, PEW RSCH. CTR. (Aug. 17, 2023), https://www.pewresearch.org/internet/2023/08/17/what-americans-know-about-ai-cybersecurity-and-big-tech/ [https://perma.cc/7PZR-8NC3]; *cf. How To Protect Against Deepfakes – Statistics and Solutions*, *supra* note 115 (finding, in 2022, that only 29% of people worldwide knew what a deepfake is).

[117] Sidoti & Vogels, *supra* note 116.

It is important to acknowledge that deepfake generating technology is not used solely for nefarious activities.[118] It has, for example, been used to show long-deceased artist Salvador Dalí talk about his artwork at a museum[119] and "record" one of Martin Luther King Jr.'s unrecorded speeches.[120] It can help translate videos, as demonstrated in an advertisement featuring soccer superstar David Beckham speaking in nine languages.[121] It has given people the ability to speak after losing their voice to disease[122] and allowed people to "animate" old family photos.[123] It has even helped bring a Parkland shooting victim back to life so he could advocate for gun control.[124]

Other deepfakes, though, have had immensely harmful impacts. For example, in 2019, a deepfake voice impersonating a CEO was used to trick a British energy company to send €220,000 to a scammer.[125] More broadly, the technology has enabled the creation

[118] Dominic Lees, *Deepfakes Are Being Used for Good – Here's How*, THE CONVERSATION (Nov. 4, 2022, 12:58 PM), https://theconversation.com/deepfakes-are-being-used-for-good-heres-how-193170 [https://perma.cc/UQ35-YR9G].

[119] Dami Lee, *Deepfake Salvador Dalí Takes Selfies with Museum Visitors*, THE VERGE (May 10, 2019, 8:50 AM), https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum [https://perma.cc/VWZ4-2385].

[120] *Virtual Martin Luther King, Jr. Project*, N.C. STATE UNIV., https://vmlk.chass.ncsu.edu/ [https://perma.cc/QB56-TW34] (last visited Nov. 18, 2023).

[121] Guy Davies, *David Beckham 'Speaks' 9 Languages for New Campaign to End Malaria*, ABC NEWS (Apr. 9, 2019, 12:52 PM), https://abcnews.go.com/International/david-beckham-speaks-languages-campaign-end-malaria/story?id=62270227 [https://perma.cc/4JRJ-7DL6].

[122] Isobel Asher Hamilton, *This Tech Company Used AI to Give a Radio Host His Voice Back After It Was Robbed by a Rare Medical Disorder*, BUS. INSIDER (June 15, 2018, 7:13 AM), https://www.businessinsider.com/tech-firm-cereproc-uses-ai-to-give-jamie-dupree-his-voice-back-2018-6 [https://perma.cc/49JM-AVTU].

[123] MYHERITAGE, https://www.myheritage.com/deep-nostalgia [https://perma.cc/KU69-7LZK] (last visited Nov. 22, 2023).

[124] Tim Nudd, *The Pain and Triumph of Change the Ref's 'Unfinished Votes'*, MUSE BY CLIO (June 1, 2021, 10:45 AM), https://musebycl.io/creative-brief/pain-and-triumph-change-refs-unfinished-votes [https://perma.cc/K658-Y9F6] (video available at https://cdn.musebycl.io/2020-10/UnfinishedVotes.com_.mp4).

[125] Margi Murphy, *The Next Wave of Scams Will Be Deepfake Video Calls from Your Boss*, BLOOMBERG (Aug. 25, 2023),

of an entire deepfake pornography industry that "predominantly targets women and is produced without people's consent or knowledge."[126] In the geopolitical space, deepfake videos have been employed in systematic disinformation campaigns, such as those done by China.[127] On the border between the innocuous and the nefarious are some deepfakes done for artistic purposes, such as a video of Supreme Court Justice Brett Kavanaugh admitting to having done some acts that, "by today's standards," would be considered sexual assault,[128] or a viral deepfake song impersonating

https://www.bloomberg.com/news/articles/2023-08-25/deepfake-video-phone-calls-could-be-a-dangerous-ai-powered-scam [https://perma.cc/2PS3-APQ9].

[126] Ulmer & Tong, *supra* note 67.

[127] *See* GRAPHIKA, DEEPFAKE IT TILL YOU MAKE IT: PRO-CHINESE ACTORS PROMOTE AI-GENERATED VIDEO FOOTAGE OF FICTITIOUS PEOPLE IN ONLINE INFLUENCE OPERATION 1 (2023).

[128] The artist described the "greatest untapped potential" of deepfake media is using it

> to envision and elicit the change we wish to see. We see this capacity in the possibility of using synthetic media to envision . . . more morally courageous versions of our public figures. . . .
>
> This capacity also leverages what I consider to be a superpower of synthetic media — that we can know they're fake and they still affect us. We can know that a synthetic video of our future sober self is fake and still have it encourage us into recovery, or that a synthetic video of Brett Kavanaugh is fake and still have it move us to advocate for gender equality in a more compassionate way. We don't have to sacrifice responsible production in order to leverage this source of prosocial potential: we make deepfakery explicit as part of envisioning more skilled, healed, courageous, and otherwise better versions of ourselves and our world.

Deepfakes for Good: The Prosocial Potential of Synthetic Media, DEEP RECKONINGS (last visited Oct. 20, 2023), https://www.deepreckonings.com/statement.html [https://perma.cc/26V5-XHPK]. In support of why such artistic pursuits can still be harmful, consider how a different deepfake artistic creation—in this case, "a series of images depicting satanic rituals inside libraries"—"was found circulating on far-right social media, where users claimed it depicted a genuine event." Thompson & Hsu, *supra* note 77 (scroll to fourth image under "A selection of test results").

famous pop artists Drake and The Weeknd that raised concerns about the continued viability of the music industry.[129]

Perhaps the most concerning thing about deepfakes is the grave impact their proliferation can have on the way people consume even authentic media.[130] Expounding on how modern humans rely on videos, images, and audio for information, one philosopher opined: “In order to survive and flourish, people need to constantly acquire knowledge about the world. And since we do not have unlimited time and energy to do this, it is useful to have sources of information that we can simply trust without a lot of verifying.”[131] Before deepfakes, people used to be able to trust that videos, photographs, and audio recordings would largely convey truthful information, eliminating the need for verification.[132] But now, “as a result of deepfakes, we are heading toward an ‘infopocalypse’ where we cannot tell what is real from what is not.”[133]

As deepfakes become increasingly impossible to detect, “seeing will no longer be believing, and we will have to decide for ourselves—without reliable evidence—whom or what to believe.”[134] The cloud of undetectable deepfakes threatens to erode

[129] Joe Coscarelli, *An A.I. Hit of Fake ‘Drake’ and ‘The Weeknd’ Rattles the Music World*, N.Y. TIMES (April 24, 2023), https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html [https://perma.cc/82VD-CG4G].

[130] Hsu & Thompson, *supra* note 71 (“[T]he mere possibility that A.I. content could be circulating is leading people to dismiss genuine images, video and audio as inauthentic.”).

[131] Don Fallis, *The Epistemic Threat of Deepfakes*, 34 PHIL. & TECH. 623, 624 (2020); *cf.* Jeffrey Gottfried, *About Three-Quarters of Americans Favor Steps to Restrict Altered Videos and Images*, PEW RSCH. CTR. (June 14, 2019), https://www.pewresearch.org/short-reads/2019/06/14/about-three-quarters-of-americans-favor-steps-to-restrict-altered-videos-and-images/ [https://perma.cc/WUK9-LXJA] (“About six-in-ten U.S. adults (61%) say it is too much to ask of the average American to be able to recognize altered videos and images, while fewer than half (38%) say the public *should* be able to recognize them.”).

[132] Fallis, *supra* note 131, at 624.

[133] *Id.* at 623.

[134] Galston, *supra* note 89; Hsu & Thompson, *supra* note 71; Thompson & Hsu, *supra* note 77 (claiming that deepfakes are “threatening society’s ability to separate fact from fiction”).

trust in the veracity of all media—even authentic media.[135] Or, to put it simply: "If any image can be manufactured— and manipulated —how can we believe anything we see?"[136]

As deepfakes become better and easier to create, "there will come a day when nothing you see on the internet can be believed."[137] This would have profound social consequences.[138] As one commentator said: "The man in front of the tank at Tiananmen Square moved the world. Nixon on the phone cost him his presidency. Images of horror from concentration camps finally moved us into action. If the notion of . . . believing what you see is under attack, that is a huge problem."[139] Moreover, in a phenomenon called the "liar's dividend," a person could reduce the impact of unfavorable media content by simply denying its authenticity.[140]

[135] Danielle K. Citron & Robert Chesney, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1779 (2019); Galston, *supra* note 89 ("As a consequence of this, even truth will not be believed.") (citation and quotation marks omitted); Fallis, *supra* note 131, at 625 ("When fake videos are widespread, people are less likely to believe that what is depicted in a video actually occurred. Thus, as a result of deepfakes, people may not trust genuine videos . . . .") (footnote omitted).

[136] Hsu & Myers, *supra* note 70.

[137] *Id.*

[138] Consider the following:

> In 2008, Barack Obama was recorded at a small gathering saying that residents of hard-hit areas often responded by clinging to guns and religion. In 2012, Mitt Romney was recorded telling a group of funders that 47% of the population was happy to depend on the government for the basic necessities of life. And in 2016, Hillary Clinton dismissed many of Donald Trump's supporters as a basket of deplorables. The accuracy of these recordings was undisputed. [Now], however, campaign operatives will have technological grounds for challenging the authenticity of such revelations. . . .

Galston, *supra* note 89.

[139] *Id*; Hsu & Thompson, *supra* note 71 ("What happens when literally everything you see that's digital could be synthetic? . . . . That certainly sounds like a watershed change in how we trust or don't trust information.").

[140] Gladwin, *supra* note 61; Cade Metz, *Internet Companies Prepare to Fight the 'Deepfake' Future*, N.Y. TIMES (Nov. 24, 2019), https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html [https://perma.cc/8SL5-3SYU]; Kelley M. Sayler & Laurie A. Harris, CONG. RSCH. SERV., IF11333, DEEP FAKES AND NATIONAL SECURITY 1 (2023) ("[T]he Liar's Dividend could become

Political candidates, for example, could "dismiss accurate but embarrassing representations" of things they did by claiming the recordings are deepfakes—and such evasions would "be hard to disprove."[141]

These problems are not mere hypotheticals: they are already reality.[142] For example, a candidate for the U.S. House of Representatives recently claimed the video depicting the murder of George Floyd was a deepfake.[143] Similarly, many real photos and videos from the 2023 Israel-Hamas war have been widely accused of being fake.[144] In fact, allegations of deepfakery have led to actual

more powerful as deep fake technology proliferates and public knowledge of the technology grows.").

[141] Galston, *supra* note 89; *see also* Metz, *supra* note 140 ("[Deepfakes] have allowed people to claim that video evidence that would otherwise be very convincing is a fake."); *Fallis*, *supra* note 131 ("As deepfakes become more prevalent, it may be epistemically irresponsible to simply believe that what is depicted in a video actually occurred. Thus, even if one watches a genuine video of a well-known politician taking a bribe and comes to believe that she is corrupt, one might not *know* that she is."). *But see* Kaylyn Jackson Schiff et al., The Liar's Dividend: The Impact of Deepfakes and Fake News on Trust in Political Discourse 37–39 (Oct. 19, 2023) (unpublished manuscript) (on file with EconPapers) (finding that claims of an authentic video being deepfaked were largely ineffective at minimizing its impact while acknowledging that "[m]ore research is warranted to evaluate whether this effect for video persists . . . as deepfakes become popularized" in the time after the study was conducted).

[142] *See* Mack DeGeurin, *8 Times 'Deepfake' Videos Were Actually Real*, GIZMODO (June 10, 2023), https://gizmodo.com/ai-deepfake-8-times-deepfake-videos-were-actually-real-1850520257 [https://perma.cc/7YPA-6KCU].

[143] Zack Budryk, *GOP House Candidate Publishes 23-Page Report Claiming George Floyd Death was Deepfake Video*, THE HILL (June 24, 2020), https://thehill.com/homenews/house/504429-gop-house-candidate-publishes-23-page-report-claiming-george-floyd-death-was/ [https://perma.cc/S5ZC-W88M].

[144] As one article described it,

> the mere possibility that A.I. content could be circulating is leading people to dismiss genuine images, video and audio as inauthentic.
>
> On forums and social media platforms like X, Truth Social, Telegram and Reddit, people have accused political figures, media outlets and other users of brazenly trying to manipulate public opinion by creating A.I. content, even when the content is almost certainly genuine.

violence and political instability. After the president of Gabon suffered a stroke in 2018, the Gabonese government attempted to prove his healthiness by posting a video of him delivering a speech.[145] Opponents, however, claimed it was a deepfake, leading the military to initiate a coup d'état.[146] This phenomenon has even made its way into the American court system. One legal team argued a video in evidence was a deepfake,[147] and another claimed the prosecution could not prove a video in evidence was not a deepfake.[148] Now, a former federal district judge is advocating for "a change to the federal rules of evidence that would allow courts to weigh whether evidence is the product of generative artificial intelligence."[149]

> . . . "The specter of deepfakes is much, much more significant now — it doesn't take tens of thousands, it just takes a few, and then you poison the well and everything becomes suspect."

Hsu & Thompson, *supra* note 71.

[145] Ali Breland, *The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink*, MOTHER JONES (Mar. 15, 2019), https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/ [https://perma.cc/2K7N-SEE9].

[146] Interestingly, there is no consensus among experts of whether the video is a deepfake. Regardless, though, as one expert said: "In some ways it doesn't matter if it's fake. That's not the underlying issue. It can be used to just undermine credibility and cast doubt." *Id.*

[147] Shannon Bond, *People Are Trying to Claim Real Videos Are Deepfakes. The Courts Are Not Amused*, NPR (May 8, 2023, 5:01 AM), https://www.npr.org/2023/05/08/1174132413/people-are-trying-to-claim-real-videos-are-deepfakes-the-courts-are-not-amused [https://perma.cc/VH9F-4ZS2].

[148] Zoe Tillman, *The Defense in The First Jan. 6 Trial's Closing Argument: Maybe the Evidence Is Fake*, BUZZFEED NEWS (Mar. 7, 2022, 8:04 PM), https://www.buzzfeednews.com/article/zoetillman/guy-reffitt-capitol-riot-trial-defense [https://perma.cc/CJL7-KK3R].

> These scenarios . . . creates [sic] a particularly insidious situation in jury trial settings where attorneys for either the defense or prosecution simply need to instill some doubt into a jury's mind. If deepfake are indistinguishable from reality and present everywhere one looks, how can anyone confidently claim any single video is true?

DeGeurin, *supra* note 142.

[149] Jacqueline Thomsen, *Ex-Judge Urges US Courts to Change Rules on AI Evidence*, BL (Oct. 27, 2023, 3:38PM), https://www.bloomberglaw.com/product/blaw/bloomberglawnews/business-and-practice/BNA%2000000018b-7269-d389-abcf-7f6f0f460001 [https://perma.cc/8TGH-X9FX].

* * *

Deepfake technology has reached a critical tipping point, and the impact of widespread public mistrust in authentic media is just beginning to be felt.[150] To state the problem simply: "Creating deepfakes is easier than ever, yet detecting them becomes increasingly difficult,"[151] and the harms stemming from prolific spread of deepfakes is immense. Given these facts, it is perhaps unsurprising that 77% of Americans believe the U.S. should take steps to restrict deepfakes.[152] Fortunately, the U.S. does not need to create a deepfake regulation system from scratch. Instead, it can imitate a legal mechanism the EU has already implemented for dealing with deepfakes.

III. AN OVERVIEW OF HOW THE EU TACKLES THE DEEPFAKE PROBLEM

While the U.S. has done very little to regulate the proliferation of deepfakes online,[153] the EU has taken strong actions that can—and should—serve as a model for regulations in the U.S. This Section provides an analysis of the EU's regulatory scheme for deepfakes.

[150] Hsu & Thompson, *supra* note 71 ("The specter of deepfakes is much, much more significant now.").

[151] Köbis et al., *supra* note 86, at 9.

[152] Amy Mitchell et al., *Americans Think Made-Up News and Videos Create More Confusion Than Other Types of Misinformation*, PEW RSCH. CTR. (June 5, 2019), https://www.pewresearch.org/journalism/2019/06/05/3-americans-think-made-up-news-and-videos-create-more-confusion-than-other-types-of-misinformation/ [https://perma.cc/2ETF-53VA] (reporting results from a 2019 study); *accord Poll Shows Voters Want Rules on Deep Fakes, International Standards, and Other AI Safeguards*, A.I. POL'Y INST., https://theaipi.org/poll-shows-voters-want-rules-on-deep-fakes-international-standards-and-other-ai-safeguards/ [https://perma.cc/V8F2-XC4S] (last visited Nov. 18, 2023) ("76% of voters want AI-generated images to be required to contain proof they were generated by a computer.").

[153] Rob Chesnut, *The EU Is Making the Rules for Big Tech as the US Watches*, BL (Sept. 5, 2023, 4:00 AM), https://www.bloomberglaw.com/product/blaw/bloomberglawnews/bloomberg-law-news/X9PA0TO000000 [https://perma.cc/6JTZ-AMSF] ("When it comes to actually regulating big internet companies, the US has been largely silent.").

The Digital Services Act ("DSA"),[154] passed in 2022,[155] is a broad piece of EU legislation that regulates illegal content, advertising, and disinformation, among other things.[156] The DSA imposes different obligations on companies depending on their size and the types of services they provide.[157] It also imposes additional requirements on the largest companies—specifically "online platforms and online search engines" with more than 45 million monthly users in the EU.[158] Only nineteen companies currently meet these criteria.[159] These "very large online platforms" and "very large online search engines" (collectively "VLOPs") "must comply with the most stringent rules of the DSA" because of "their size and the potential impact they can have on society."[160]

Of the DSA requirements unique to VLOPs, many establish co-regulatory mechanisms, which are "governance structure[s] where government involvement exists, but is limited, and most of the actions are taken by other stakeholder groups, usually under the oversight of one or more governmental bodies."[161] The DSA creates

[154] Regulation 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act), 2022 O.J. (L 277) [hereinafter Digital Services Act].

[155] *The Digital Services Act Package*, EUR. COMM'N (last updated Sept. 25, 2023), https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package [https://perma.cc/SSV2-SEHT].

[156] *Id.*

[157] *The Digital Services Act: Ensuring a Safe and Accountable Online Environment*, EUR. COMM'N, https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en [https://perma.cc/9AHG-K8ZR] (last visited Nov. 18, 2023).

[158] Digital Services Act, *supra* note 154, at 63 (discussing art. 33(1)).

[159] *DSA: Very Large Online Platforms and Search Engines*, EUR. COMM'N (Apr. 15, 2023), https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops [https://perma.cc/LSA8-LAFJ].

[160] *Id.*

[161] David Morar, *The Digital Services Act's Lesson for U.S. Policymakers: Co-Regulatory Mechanisms*, BROOKINGS INST. (Aug. 23, 2022), https://www.brookings.edu/articles/the-digital-services-acts-lesson-for-u-s-policymakers-co-regulatory-mechanisms/ [https://perma.cc/S76D-NNPL].

three types of co-regulatory mechanisms for VLOPs: risk assessments, mitigation measures, and audits.[162]

A. Risk Assessments

The DSA requires VLOPs to conduct a risk assessment at least once per year "and in any event prior to deploying functionalities that are likely to have a critical impact on the risks identified" in the Act.[163] Each risk assessment must "diligently identify, analyse [sic] and assess any systemic risks in the [EU] stemming from the design or functioning of [the VLOP's] service and its related systems, including algorithmic systems, or from the use made of their services."[164] The risk assessment must be "specific to [the VLOP's] services and proportionate to the systemic risks, taking into consideration their severity and probability."[165] Among other systemic risks, the risk assessment must account for "any actual or foreseeable negative effects on civic discourse and electoral processes, and public security."[166] VLOPs are required to take into account how these systemic risks are affected by "the design of their recommender systems and any other relevant algorithmic system" and "their content moderation systems."[167]

Beyond these requirements, the DSA does not impose much structure for risk assessments.[168] The DSA "is not descriptive in how it defines the risk assessments, and VLOPs are to create these assessments on their own."[169] However, the government is still involved to an extent: An EU commission can request a copy of a VLOP's risk assessment at any time,[170] and an EU board is required

[162] *Id.*

[163] Digital Services Act, *supra* note 154, at 64 (discussing art. 34(1)).

[164] *Id.*

[165] *Id.*

[166] *Id.* (discussing art. 34(1)(c)).

[167] *Id.* (discussing art. 34(2)(a)–(b)).

[168] Morar, *supra* note 161 (noting that the risk assessments are only "loosely structured by the government").

[169] *Id.*

[170] Digital Services Act, *supra* note 154, at 65 (discussing art. 34(3)).

to publish an annual report identifying and assessing the "most prominent and recurrent systemic risks" reported by VLOPs.[171]

Importantly, risk assessments "are not done for their own sake. Rather, [VLOPs] are supposed to—based on them—establish mitigation measures."[172]

B. Mitigation Measures

VLOPs must "put in place reasonable, proportionate[,] and effective mitigation measures, tailored to the specific systemic risks identified" in the risk assessments.[173] While the DSA does not explicitly state what form mitigation measures must take, it gives several examples.[174] Of particular relevance is the DSA's endorsement of mitigation measures ensuring a deepfake "is distinguishable through prominent markings" on the VLOP's online platform and "providing an easy to use functionality" enables users to indicate this information.[175] More broadly, the DSA also endorses adopting "content moderation processes"[176] and "adapting the design, features[,] or functioning of [a VLOP's] services, including their online interfaces" to address deepfakes.[177]

The DSA leaves much discretion to VLOPs themselves to determine how to mitigate harms presented by their platforms.[178] More than anything, these mitigation requirements can be thought of as creating an obligation of due diligence rather than mandating specific mitigatory activities.[179] However, because a VLOP's due

[171] *Id.* (discussing art. 35(2)(a)).

[172] Morar, *supra* note 161.

[173] Digital Services Act, *supra* note 154, at 65 (discussing art. 35(1)).

[174] *Id.* (discussing art. 35(1)(a)–(k)).

[175] *Id.* (discussing art. 35(1)(k)).

[176] *Id.* (discussing art. 35(1)(c)).

[177] *Id.* (discussing art. 35(1)(a)).

[178] Morar, *supra* note 161.

[179] *See* Rachel Griffin & Carl Vander Maelen, Codes of Conduct in the Digital Services Act: Exploring the Opportunities and Challenges 2 (Sept. 6, 2023) (unpublished manuscript) (on file with Social Science Research Network) (characterizing many of the DSA's regulations specific to VLOPs as "creat[ing] 'due diligence' obligations"); *see also* Digital Services Act, *supra* note 154, at 48 (titling chapter three, which contains VLOP obligations, "Due Diligence Obligations for a Transparent and Safe Online Environment").

diligence obligations are not clearly defined within the DSA,[180] the Act authorizes the creation of various codes to flesh out due diligence requirements.[181]

Accordingly, the DSA says the EU will "encourage and facilitate the drawing up of voluntary codes of conduct."[182] Where significant risks

> emerge and concern several [VLOPs], the Commission may invite the [VLOPs] concerned . . . as well as relevant competent authorities, civil society organisations [sic] and other relevant stakeholders, to participate in the drawing up of codes of conduct, including by setting out commitments to take specific risk mitigation measures, as well as a regular reporting framework on any measures taken and their outcomes.[183]

In essence, the "primary overarching purpose" of these codes is "clarifying and supplementing" the due diligence obligations the DSA imposes on VLOPs.[184] The codes will "supplement broad, abstract obligations" such as the DSA's mitigation requirements, "with more specific, concrete commitments."[185]

While the EU delegates significant authority over to the creators of these codes,[186] the codes are still subject to meaningful government oversight.[187] The DSA states the government will ensure these codes "clearly set out their specific objectives, contain key performance indicators to measure the achievement of those objectives[,] and take due account of the needs and interests of all

[180] *See* Griffin & Maelen, *supra* note 179*,* at 7 (noting that the systemic risks identified in DSA 34… are "vague and abstract").

[181] *See id.* at 6 (characterizing these codes as "*de facto* regulatory obligations" because they "offer[] perhaps the most straightforward way to demonstrate compliance" with the DSA's mitigation requirements).

[182] Digital Services Act, *supra* note 154, at 76 (discussing art. 45(1)).

[183] *Id.* (discussing art. 45(2)).

[184] Griffin & Maelen, *supra* note 179, at 2.

[185] *Id.* at 3 tbl.1.

[186] *Id.* at 6 ("DSA codes are envisaged as . . . delegating significant regulatory power to private auditors.").

[187] Morar, *supra* note 161 ("While drafting is to be done by these participants, the government would take an active role to ensure the codes of conduct are written properly, and especially consider the needs and interests of EU citizens. The codes of conduct thus involve significant government oversight . . . .").

interested parties, and in particular citizens."[188] The government will also "ensure that participants report regularly . . . on any measures taken and their outcomes, as measured against the key performance indicators that they contain."[189]

One such code of conduct is the Code of Practice on Disinformation. The 2022 Strengthened Code of Practice on Disinformation ("CPD" or "Code")[190] is an update to the world's first self-regulatory code on disinformation.[191] In total, the CPD contains forty-four commitments and 128 specific measures for implementing these commitments.[192] Like its predecessor, the CPD was not written by the EU itself[193] and is nominally voluntary.[194] However, because abiding by the terms of the CPD qualifies as a mitigation measure under the legally binding DSA,[195] VLOPs have a major incentive to adhere to the Code. Perhaps as a result, the CPD has forty-four signatories.[196]

The CPD contains several measures that would likely have a meaningful impact on the proliferation of deepfake media online if implemented in the U.S. Broadly, it requires signatories to "put in

[188] Digital Services Act, *supra* note 154, at 76 (discussing art. 45(3)).

[189] *Id.*

[190] EUR. COMM'N, THE STRENGTHENED CODE OF PRACTICE ON DISINFORMATION 2022 (2022) [hereinafter CODE OF PRACTICE ON DISINFORMATION].

[191] Brooke Tanner, *EU Code of Practice on Disinformation*, BROOKINGS INST. (Aug. 5, 2022), https://www.brookings.edu/articles/eu-code-of-practice-on-disinformation/ [https://perma.cc/6L7U-29EZ].

[192] *Id.*

[193] *The 2022 Code of Practice on Disinformation*, EUR. COMM'N (July 4, 2022), https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation [https://perma.cc/BPJ4-R6SH].

[194] *See id.* ("It is for the signatories to decide which commitments they sign up to and it is their responsibility to ensure the effectiveness of their commitments' implementation.").

[195] Natasha Lomas, *Europe Wants Platforms to Label AI-Generated Content to Fight Disinformation*, TECHCRUNCH (June 6, 2023, 5:09 AM), https://techcrunch.com/2023/06/06/eu-disinformation-code-generative-ai-labels/ [https://perma.cc/Z6M9-KNJE]; *see infra* notes 219–20 and accompanying text.

[196] *Signatories of the 2022 Strengthened Code of Practice on Disinformation*, EUR. COMM'N (June 16, 2022), https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation [https://perma.cc/L88P-9BBU].

place or further bolster policies to address both misinformation and disinformation across their services," including deepfakes.[197] They must also "adopt, reinforce[,] and implement clear policies regarding impermissible manipulative behaviours [sic] and practices on their services, based on the latest evidence on the conducts and tactics . . . employed by malicious actors."[198] Furthermore, signatories must create policies "for countering prohibited manipulative practices" for deepfakes, "such as warning users and proactively detect [sic] such content."[199] VLOPs must ensure the algorithms used to detect, moderate, and sanction deepfakes are "trustworthy,"[200] and they must "operate channels of exchange" between other VLOPs "in order to proactively share information about cross-platform influence operations . . . with the aim of preventing dissemination and resurgence on other services."[201]

In addition to requiring signatories to implement deepfake policies and monitor for deepfakes on their platforms, the CPD also requires signatories to provide users with several tools for identifying deepfakes. CPD signatories must "empower users with tools to assess the provenance and edit history or authenticity or accuracy" of the digital content they consume.[202] Signatories must also "help users benefit from the context and insights provided by independent fact-checkers or authoritative sources" by means such as "labels indicating fact-checker ratings."[203] Furthermore, they must develop tools to send "warnings or updates" to users who have interacted with media that has since been identified as deepfake.[204] All labeling and warning systems must be designed "in accordance with up-to-date scientific evidence and with analysis of [the

[197] CODE OF PRACTICE ON DISINFORMATION, *supra* note 190, at 15.

[198] *Id.* at 16.

[199] *Id.* at 17.

[200] *Id.*

[201] *Id.*

[202] *Id.* at 21.

[203] *Id.* at 22.

[204] *Id.*

signatory's] users' needs on how to maximise [sic] the impact and usefulness of such interventions."[205]

Additionally, CPD signatories must "facilitate user access to tools and information to assess the trustworthiness of information sources"[206] by providing access to "indicators of trustworthiness," such as those developed by independent third parties focusing on a source's integrity.[207] Should an indicator of trustworthiness prove to be incorrect, the signatory has a duty to correct it.[208] Signatories must also allow users to flag content they believe to be false.[209]

With their breadth and robustness, the terms of the CPD seem adequate for withstanding an imminent onslaught of deepfakes. It requires VLOPs to establish policies regarding the acceptability of deepfake content, design and implement detectors, allow for user reporting of deepfakes, label detected deepfakes as such, and implement provenance authentication technology. It even requires platforms to notify users if content they interacted with previously later turns out to be deepfake. Importantly, the CPD requires that all these policies and technologies meet the best possible industry standards.

C. Audits

In addition to risk assessments and mitigation measures like the CPD, the DSA also contains several provisions relating to audits. A VLOP's state of compliance with the terms of the DSA—and with codes like the CPD when applicable—is to be determined by annual audit.[210] These audits are performed not by the government, but rather by some other independent organization.[211] The DSA does

[205] *Id.* at 23.

[206] *Id.*

[207] *Id.*

[208] *Id.* at 24.

[209] *Id.* at 25. A user must be allowed to appeal if they believe their content has been improperly flagged as false. *Id.*

[210] Digital Services Act, *supra* note 154, at 67 (discussing art. 37(1)(a)-(b)).

[211] *Id.* at 68 (discussing art. 37(3)).

not state how the audit must be conducted,[212] instead encouraging standardization bodies to develop voluntary standards for the audits.[213]

The annual audit must create a written opinion assessing the VLOP's compliance with the DSA and relevant codes.[214] In the event of noncompliance, the audit must provide "operational recommendations on specific measures to achieve compliance,"[215] and the VLOP must implement the recommended measures and generate a report detailing the implementation.[216] If the VLOP refuses to implement the recommended measures, it must generate a report to "justify . . . the reasons for not doing so and set out any alternative measures that they have taken to address any instances of non-compliance identified."[217]

For the purposes of an audit, a VLOP's "adherence to and compliance with a given code of conduct" like the CPD "may be considered as an appropriate risk mitigating measure" for systemic risks, thereby indicating compliance with the DSA's due diligence requirements.[218] Similarly, "refusal without proper explanations" to comply with a code can be considered when determining if a VLOP has violated the DSA.[219]

In addition to its co-regulatory measures, the DSA contains a powerful enforcement mechanism: Any VLOP that fails to comply with the DSA can incur a fine worth up to 6% of the VLOP's annual global revenue.[220]

* * *

As discussed above, the DSA contains three main provisions: risk assessments; mitigation measures, such as the CPD; and audits.

[212] Morar, *supra* note 161 ("While the legislation has specific guidelines on who the auditors should be, and a very general minimal framework for the audit reports themselves, the DSA does not design the audits in any way.").

[213] Digital Services Act, *supra* note 154, at 75. (discussing art. 44(1)(e)).

[214] *Id.* at 68 (discussing art. 37(4)(g)).

[215] *Id.* (discussing art. 37(4)(h)).

[216] *Id.* at 69 (discussing art. 37(6)).

[217] *Id.*

[218] *Id.* at 29 (discussing recital 104).

[219] *Id.*

[220] *Id.* at 94 (discussing art. 74(1)).

This Article continues by providing an assessment of the DSA and CPD, asking whether they are likely to be effective at handling the problem of deepfake media.

IV. ASSESSING THE EU'S POLICIES

The DSA will strongly incentivize VLOPs to implement rigorous measures to counter the spread of deepfakes. The Act's greatest strength is its use of co-regulatory mechanisms. By their nature, co-regulatory mechanisms hold major advantages over direct government regulations because they "provide flexibility that traditional legislation usually lacks, either systematize or publicize industry action through assessments, invite in civil society to help shape mitigation strategies in codes of conduct, or ensure independent audits of platforms."[221] Here, the DSA's co-regulatory mechanisms perform all these functions.

By requiring VLOPs to file their risk assessments with the EU, which then publishes an annual report on the risks commonly reported, the public is made aware of the most pressing issues relating to deepfakes, resulting in "transparent and standardized industry action" on the issue of deepfakes.[222] Furthermore, the Act's strong enforcement mechanism—which provides steep penalties for noncompliance and protects auditors' independence to ensure robust enforcement—is also very likely to motivate VLOPs to ensure strict compliance with the DSA. With such a steep penalty for noncompliance, risk-averse VLOPs will almost certainly seek to avoid a noncompliance determination by proactively engaging in the required mitigatory activities.[223] Even when faced with uncertainty, the VLOP will likely err on the side of mitigating.

[221] Morar, *supra* note 161.

[222] *Id.*

[223] To put it one way:

> The EU isn't kidding around. The potential fines for failure to comply with the DSA are breathtakingly large—up to 6% of a company's global (not just EU) revenue (not profit) or a complete ban from the EU.
>
> Google generated almost $280 billion in revenue in 2022, and a file [sic] of 6% would amount to almost $17 billion, over a quarter of their annual profits. . . . That will get your attention.

The most impactful part of the DSA is its extensive use of codes—especially of the CPD. Codes like the CPD have a wide range of advantages over traditional regulation. One is their flexibility, which allows them to be regularly updated to "address emerging and quickly-evolving risks."[224] This allows the CPD to target the features of specific platforms and deal with problems not fully addressed in the DSA.[225] By including concrete commitments, key performance indicators, and reporting standards, the CPD facilitates oversight by auditors and regulators.[226] Additionally, because the DSA allows the government to invite a diverse range of actors to participate in the development of the CPD, the process can also be remarkably inclusive.[227]

The great potential codes harbor as a mechanism for countering deepfake proliferation is exemplified by the robustness of the CPD's current provisions. Already, the CPD requires VLOPs to implement effective algorithms for identifying and moderating deepfakes; flag potentially deepfake media as such; allow users to report suspected deepfakes; develop protocols to allow users to probe and prove the provenance and authenticity of media; and coordinate with fact-checkers and other VLOPs to identify deepfake campaigns across their platforms.[228] These requirements are specific and strict, attacking the deepfake problem on several fronts. The fact the CPD has already produced such meaningful requirements demonstrates the effectiveness of codes generally as a mechanism for combatting deepfakes.

Chesnut, *supra* note 153.

[224] Griffin & Maelen, *supra* note 179, at 7.

[225] *See id.*

[226] *See id.*

[227] *See* Morar, *supra* note 161 (noting that co-regulatory mechanisms "can be more inclusive than the normal policy-making process").

[228] *See supra* notes 198–210 and accompanying text.

V. A DEFENSE OF THE LEGITIMACY & WISDOM OF CO-REGULATORY MECHANISMS

Some commentators question whether co-regulatory mechanisms are in fact effective at their goals.[229] Others have bemoaned the "ambiguous" legal status of soft law in general.[230] To this effect, some have questioned the legitimacy of the codes of practice established by co-regulatory mechanisms.[231] While co-regulatory mechanisms do present some unique challenges, they are certainly legitimate forms of governance and are ideal for situations like the regulation of deepfakes.

Co-regulatory mechanisms offer a number of advantages over traditional command-and-control legislation.[232] A major advantage is that they "combine the efficiency and adaptability of industry self-regulation" with the "public accountability and transparency" of traditional government regulation."[233] As a result, co-regulatory

[229] *See, e.g.*, Griffin & Maelen, *supra* note 119, at 6 ("Developing de facto regulatory obligations addressing contested, politicised [sic] issues – such as online speech regulation, media freedom, and public health and security – through informal, opaque negotiations between companies and unelected regulatory agencies, raises concerns about legitimacy and accountability.").

[230] Corina Andone & Florin Coman-Kund, *Persuasive Rather Than Binding' EU Soft Law? An Argumentative Perspective on the European Commission's Soft Law Instruments in Times of Crisis*, 10 THEORY & PRAC. LEGIS. 22, 30 (2022); *accord* Carl Vander Maelen, *Hardly Law or Hard Law? Investigating the Dimensions of Functionality and Legalisation of Codes of Conduct in Recent EU Legislation and the Normative Repercussions Thereof*, 47 EUR. L. REV. 752, 756 ("[T]here is no meaningful legal framework at the EU level that specifies [codes of conduct's] legal nature, the conditions for their validity, their precise areas of application, their links with hard law, and their limits.").

[231] Maelen, *supra* note 230, at 771 (claiming codes of conduct "have a difficult relationship with legitimacy," especially when considering "accountability and the transparency, inclusiveness and openness of the governance process").

[232] These advantages include "greater flexibility and adaptability," "potentially lower compliance and administrative costs," "an ability to harness industry knowledge and expertise to address industry-specific and consumer issues directly," and "quick and low-cost complaints-handling and dispute resolution mechanisms." AUSTRALIAN COMMC'NS AND MEDIA AUTH., OPTIMAL CONDITIONS FOR EFFECTIVE SELF- AND CO-REGULATORY ARRANGEMENTS 5 (2011).

[233] Helen Cheng, *The Rise of Co-Regulation, From GDPR to Canada's Bill C-11*, INT'L ASS'N OF PRIV. PROS. (Dec. 17, 2020), https://iapp.org/news/a/the-rise-

mechanisms are especially valuable in rapidly advancing fields like AI[234] because a "common challenge[] with emerging technology governance" is designing a regulatory system "flexible enough to keep pace with changing technologies."[235] The process of creating traditional legislation "is slow and, once enacted, major statutes tend to remain in place for decades without substantial revisions or updates."[236] For this reason, legislation "could easily become obsolete" if it is worded with too much specificity.[237] Co-regulatory mechanisms are important because they allow legislators to enact "broadly worded legislation" more likely to withstand the test of time[238] by "operationaliz[ing] . . . and translating it into concrete indicators" that businesses can understand and comply with.[239]

In addition to having great longevity, broadly worded statutes are advantageous because they grant companies the flexibility to

of-co-regulation-from-gdpr-to-canadas-bill-c-11/ [https://perma.cc/G7WR-3L4A]; Maelen, *supra* note 230, at 769–70 (arguing that co-regulatory mechanisms are a "best of both worlds" solution that combines the "inclusiveness and flexibility" of self-regulatory mechanisms with the "uniformity and rigidity" of traditional regulation, leading to "legal certainty, effectiveness, and accountability").

[234] To explain:

> [C]onsider AI-based solutions to address online disinformation. This is an area where the landscape changes too quickly for technology-specific regulation. Thus, in some AI domains, a better approach is to use regulation to address higher-level issues—such as the need for transparency—and allow soft law frameworks to be developed by experts who are much closer to the technology than regulators.

John Villasenor, *Soft Law as a Complement to AI Regulation*, BROOKINGS INST. (July 31, 2020), https://www.brookings.edu/articles/soft-law-as-a-complement-to-ai-regulation/ [https://perma.cc/FNQ2-Z5JW]; *accord* Griffin & Maelen, *supra* note 179, at 7 (noting that the CPD developed under the DSA has proven effective at regulating deepfakes despite the fact "that the DSA did not anticipate the rapidly increasing availability of generative AI").

[235] Jonas J. Monast, *Emerging Technology Governance in the Shadow of the Major Questions Doctrine*, 24 N.C. J.L. & TECH. 1, 4 (2023).

[236] *Id.* at 6.

[237] Cheng, *supra* note 233.

[238] *Id.*

[239] *Id*; *accord* Griffin & Maelen, *supra* note 179, at 3 tbl.1 ("Codes supplement broad, abstract obligations . . . with more specific, concrete commitments.").

implement requirements in ways they find most optimal[240]—which has the additional benefit of incentivizing innovation.[241] Alternatively, if the lack of standards inherent in broadly worded laws makes it "burdensome" for a company to prove it has complied with the law,[242] it can comply with the concrete, specific standards set forth by the co-regulatory mechanism.[243] In sum, the flexibility of co-regulatory mechanisms allows regular updates, thereby "reducing the risk of outdated detailed standards," while also enabling "targeting specific platforms or technical features."[244]

Scholars correctly point out that the difficulty in proving compliance with a broadly worded statute makes a nominally "voluntary" code almost mandatory for businesses.[245] This, however, is not a bad thing; it is a feature, not a bug. Since a diverse range of stakeholders are invited to directly participate in the creation of a code's standards, concrete regulations emerging from co-regulatory mechanisms are quite sound. By involving collaboration between public and private actors, co-regulatory mechanisms ensure a "high level of compliance" while "reducing administrative costs" for the businesses subject to the

[240] Francesco Vigna, Co-regulation Approach for Governing Big Data: Thoughts on Data Protection Law, 15th Int'l Conf. on Theory & Prac. of Elec. Governance (ICEGOV 2022) 59 (Nov. 18, 2022) (on file with Association for Computing Machinery).

[241] *Id.* at 60.

[242] *Id.* at 59.

[243] *See* Griffin & Maelen, *supra* note 179, at 3.

[244] *Id.* at 7.

[245] Maelen, *supra* note 230, at 766 (referring to the phenomenon as the "'hardening' of soft law instruments"); Griffin & Maelen, *supra* note 179, at 1 (arguing that "codes create de facto legal obligations").

requirements.[246] They also give many interest groups a seat at the table.[247]

Co-regulatory mechanisms like the DSA do not delegate away ultimate responsibility for ensuring people are protected by the law. Rather, they still recognize the government as the party "ultimately responsible for protecting the public interest, with baseline statutory requirements serving as an incentive for businesses to undertake proactive . . . measures."[248] In contrast to self-regulatory mechanisms, which involve no government oversight, co-regulatory mechanisms do have meaningful government oversight.[249] And while "industry-led codes inherently delegate regulatory authority from democratically-legitimated institutions to private actors," "this is not inherently negative."[250] Rather, "it can enable more detailed and regularly-updated standards, and enhance legitimacy by enabling pluralistic stakeholder participation."[251] Admittedly, care must be taken to ensure the co-regulatory mechanism does not fall victim to two potential traps. One is the risk of regulatory

[246] Vigna, *supra* note 240, at 59; *accord* Griffin & Maelen, *supra* note 179, at 7 ("[C]odes can develop stronger institutional accountability mechanisms. Concrete commitments, KPIs and reporting standards facilitate oversight by regulators, auditors and external stakeholders, including comparisons between platforms and over time. Codes could also develop additional oversight structures. . . .") (footnote omitted).

[247] Griffin & Maelen, *supra* note 179, at 4 tbl.1 ("Code development should give user groups, civil society and other relevant stakeholders opportunities to shape DSA implementation.").

[248] Cheng, *supra* note 233.

[249] Vigna, *supra* note 240, at 60 ("The role of public authorities and government in co-regulation may consist in either active collaboration with private stakeholders during the drafting of the co-regulatory instrument, or just the final endorsement of the co-regulatory measure."); Maelen, *supra* note 230, at 754 ("Co-regulation is different [from self-regulation] in that it encompasses constructions that link non-state regulatory systems to state regulation, by relying on private entities to perform a variety of government functions while state authorities provide oversight and enforcement."); *accord* Griffin & Maelen, *supra* note 179, at 6 ("Under the DSA, regulators can identify systemic risks that codes should address and invite preferred stakeholders to participate – ultimately determining whether commitments suffice to comply with [the DSA].").

[250] Griffin & Maelen, *supra* note 179, at 9.

[251] *Id.*

capture[252]—that is, that the co-regulatory mechanism might become dominated by the industries it is charged with regulating.[253] Another is the risk of not being transparent or inclusive in the development of its codes.[254] These concerns are legitimate, and great care should be taken to ensure they do not occur. They are not, however, insurmountable.

The DSA creates a co-regulatory mechanism that will likely be effective at mitigating the harms caused by deepfakes in the EU. Its requirement for assessments ensures that VLOPs, the government, and the public can identify risks and can act proactively to address problems. The DSA's mitigation requirements are worded broadly enough to ensure long-term viability no matter how quickly deepfake technology changes. The CPD itself contains excellent measures requiring VLOPs to combat deepfakes on a variety of fronts, all using the best available technology and tactics. An audit system ensures compliance with these terms, although care must be taken to ensure such audits are rigorous and do not fall victim to regulatory capture.

In light of how effective the EU legislation will likely be, a similar policy should be adopted in the U.S. The next Section considers why a much-discussed alternative—removing Section 230 immunity for VLOPs—is dangerous and ill-advised.

VI. CONSIDERING THE ALTERNATIVES: WHY CO-REGULATORY MECHANISMS ARE BETTER THAN REMOVING SECTION 230 IMMUNITY

Because current laws do not require VLOPs to do anything about deepfakes on their platforms,[255] private tort actions are the only means of forcing VLOPs to act. Indeed, a variety of tort actions can be brought to address deepfakes, and these can be brought against a variety of defendants. However, these suits almost never succeed

[252] AUSTRALIAN COMMC'NS AND MEDIA AUTH., *supra* note 232, at 5; Griffin & Maelen, *supra* note 179, at 9.

[253] Will Kenton, *Regulatory Capture Definition with Examples*, INVESTOPEDIA (Mar. 1, 2021), https://www.investopedia.com/terms/r/regulatory-capture.asp [https://perma.cc/9B6E-RAM3].

[254] Maelen, *supra* note 230, at 770.

[255] *See supra* notes 102–08 and accompanying text.

against VLOPs because of one law: Section 230 of the Communications Decency Act.[256] As a result, many commentators concerned with the proliferation of deepfakes have called for Section 230 to be repealed or modified so as to allow VLOPs to be held liable for hosting deepfakes.[257] This, however, is not an acceptable solution; Section 230 is too important to be done away with or modified.

To understand the importance of Section 230, it is worth remembering why it was enacted. Long before the internet existed, the Supreme Court held it was unconstitutional to impose liability on a bookseller who distributes a book with illegal content when the bookseller was unaware of that book's contents.[258] In this decision, the Court emphasized that to hold otherwise would impose an unconstitutional chilling effect on speech.[259] The case drew a clear

[256] 47 U.S.C. § 230.

[257] *E.g.*, Nicholas O'Donnell, Note, *Have We No Decency?: Section 230 and the Liability of Social Media Companies for Deepfake Videos*, 2021 ILL. L. REV. 701, 738 (2021) (calling for a "narrow, deepfake-specific amendment to Section 230"); Jared de Guzman, *Saving Face: Lessons from the DMCA for Combating Deepfake Pornography*, 58 GONZ. L. REV. 109, 136 (2022) (arguing that Section 230 "should be reasonably amended to protect the individual liberties violated by deepfake pornography"); Anne Pechenik Gieseke, *"The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography*, 73 VAND. L. REV. 1479, 1493, 1509 (2020) (proposing "amending section 230 to allow victims to sue platforms for unlawfully hosting deepfake pornography"); *see also* Alex Engler, *Fighting Deepfakes When Detection Fails*, BROOKINGS INST. (Nov. 14, 2019), https://www.brookings.edu/articles/fighting-deepfakes-when-detection-fails/ [https://perma.cc/QD99-G6QC] ("A growing number of scholars are calling for qualifications on Section 230."); Morar, *supra* note 161 ("[G]utting Section 230 has unfortunately risen to the top of the pile of legislative propositions in the United States.").

[258] Smith v. California, 361 U.S. 147, 153–55 (1959).

[259] The Court explained that a statute imposing strict liability on distributors for the contents of the books they distribute would have an unconstitutional chilling effect on constitutionally protected speech:

> [I]f the bookseller is criminally liable without knowledge of the contents, . . . he will tend to restrict the books he sells to those he has inspected . . . . Every bookseller would be placed under an obligation to make himself aware of the contents of every book in his shop. It would be altogether unreasonable to demand so near an approach to omniscience. And the bookseller's burden would become the public's

distinction between speech publishers and speech distributors; while publishers were liable for the content of the speech, distributors were not.[260]

While the line between publisher and distributor was easy enough to identify at the time, it became much harder to identify with the advent of the internet.[261] If, for example, a user makes a defamatory post on a social media platform, should the platform be considered a publisher or distributor of that post? The answer came from a pair of mid-1990's cases:[262] Whether a platform would be deemed publisher or distributor depended on how much control the platform exerted over user-generated posts.[263] These precedents incentivized websites to stop moderating user content entirely

> burden, for by restricting him the public's access to reading matter would be restricted.

Id. at 153–54 (quotation marks, footnotes, and citations omitted).

[260] Ashley Johnson & Daniel Castro, *Overview of Section 230: What It Is, Why It Was Created, and What It Has Achieved*, INFO. TECH. & INNOVATION FOUND. (Feb. 22, 2021), https://itif.org/publications/2021/02/22/overview-section-230-what-it-why-it-was-created-and-what-it-has-achieved/ [https://perma.cc/2X3A-BR4T] (noting Smith v. California drew a sharp "distinction between publishers, which are liable for the statements they circulate, and distributors—such as a bookstore or a newsstand—which are not").

[261] *Id.* ("The rise of online services—from early blogs and forums to search engines, social media, online marketplaces, and more—complicated the issue because they blurred the lines between distributors and publishers.").

[262] Cubby, Inc., v. CompuServe, Inc., 776 F. Supp. 135 (S.D.N.Y. 1991); Stratton Oakmont, Inc. v. Prodigy Servs. Co., No. 31063/94, 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995).

[263] Christopher Cox, *The Origins and Original Intent of Section 230 of the Communications Decency Act*, RICH. J.L. & TECH. (Aug. 27, 2020), https://jolt.richmond.edu/2020/08/27/the-origins-and-original-intent-of-section-230-of-the-communications-decency-act/ [https://perma.cc/ND8L-DL8Q] (stating the differentiating factor between the two sites was that one adopted content guidelines while the other did not). To explain further, the website in *Cubby* was considered a mere distributor of a user-generated newsletter because it "had no firsthand knowledge of the contents of [the newsletter], had no control over its publication, and had no opportunity to review the newsletter's contents." Johnson & Castro, *supra* note 260. By contrast, the website in *Stratton Oakmont* was considered a publisher of a user-generated post because it "had a set of content guidelines that outlined rules for user-generated content, a software program that filtered out offensive language, and moderators who enforced the content guidelines." *Id.*

because, if they did, they would expose themselves to additional liability.[264]

Section 230 was enacted to undo these precedents.[265] It has two main provisions: one that protects VLOPs from liability when they choose to remove third-party content, and one that protects them when they do not.[266] In addition to removing the legal barriers that disincentivized VLOPs from moderating content, Section 230's immunity shield helped "encourage continued growth and development of the [i]nternet in its early years."[267] By creating "clear rules of legal liability," Section 230 offered meaningful protection to websites hosting user-generated content.[268] The impact this had on the development of the internet cannot be overstated. Without Section 230's immunity shield, "websites would be exposed to lawsuits for everything from users' product reviews to

[264] Adi Robertson, *Why the Internet's Most Important Law Exists and How People are Still Getting it Wrong*, THE VERGE (June 21, 2019), https://www.theverge.com/2019/6/21/18700605/section-230-internet-law-twenty-six-words-that-created-the-internet-jeff-kosseff-interview [https://perma.cc/2KVS-LQWP] (claiming these two cases created a "really weird rule where these online platforms can reduce their liability by not moderating content"); Cox, *supra* note 263 ("If allowed to stand, this jurisprudence would have created a powerful and perverse incentive for platforms to abandon any attempt to maintain civility on their sites.").

[265] As Christopher Cox, the author and co-sponsor of Section 230, said:

> If allowed to stand, this jurisprudence would have created a powerful and perverse incentive for platforms to abandon any attempt to maintain civility on their sites. A legal standard that protected only websites where "anything goes" from unlimited liability for user-generated content would have been a body blow to the internet itself. Rep. Wyden and I were determined that good faith content moderation should not be punished, and so the Good Samaritan provision in [Section 230] was born.

Cox, *supra* note 263.

[266] Johnson & Castro, *supra* note 260.

[267] *Id.*

[268] Cox, *supra* note 263.

book reviews."[269] But with it, websites have been able to host user-generated content for myriad purposes.[270]

While Section 230 has been incredibly important for enabling the internet to become what it is today, it has also created some harm. One is that it almost completely prevents a plaintiff from holding a VLOP liable for hosting deepfake content appropriating the plaintiff's likeness.[271] As a result, many commentators concerned

[269] To explain further:

> In 21st century terms, this would mean that Yelp would be exposed to lawsuits for its users' negative comments about restaurants, and Trip Advisor could be sued for a user's disparaging review of a hotel. Indeed any service that connects buyers and sellers, workers and employers, content creators and a platform, victims and victims' rights groups — or provides any other interactive engagement opportunity one can imagine — would face open-ended liability if it continued to display user-created content.

Cox, *supra* note 263.

[270] Because of Section 230,

> user-created content . . . has come to typify the modern internet. Not only have billions of internet users become content creators, but equally they have become reliant upon content created by other users. Contemporary examples abound. In 2020, without user-created content, many in the United States contending with the deadliest tornado season since 2011 could not have found their loved ones. Every day, millions of Americans rely on "how to" and educational videos for everything from healthcare to home maintenance. During the COVID-19 crisis, online access to user-created pre-K, primary, and secondary education and lifelong learning resources has proven a godsend for families across the country and around the world. More than 85% of U.S. businesses with websites rely on user-created content, making the operation of Section 230 essential to ordinary commerce. The vast majority of Americans feel more comfortable buying a product after researching user generated reviews, and over 90% of consumers find user-generated content helpful in making their purchasing decisions. User generated content is vital to law enforcement and social services. Following the rioting in several U.S. cities in 2020, social workers were able to match people with supplies and services to victims who needed life-saving help, directing them with real-time maps.

Id. (footnotes omitted).

[271] Leslie Y. Garfield Tenzer, Defamation in the Age of Artificial Intelligence (Aug. 18, 2023) (unpublished manuscript) (on file with Social Science Research Network) (noting that Section 230 would effectively bar a defamation suit brought against a VLOP for hosting defamatory AI-generated content); Gieseke, *supra*

with the proliferation of deepfakes have called for Section 230 to be repealed or modified so as to allow VLOPs to be held liable for hosting deepfakes.[272] Often, these critics argue that the internet has changed significantly since Section 230 was passed[273] and that Congress never intended for the law to be applied so broadly.[274]

To the extent these critics argue VLOPs should be strongly incentivized—if not required—by law to stem the spread of

note 257, at 1493, 1495 (noting that the "[h]yperimmunity" given by Section 230 to VLOPs creates "nearly insurmountable barriers" for victims of deepfake pornography seeking redress); Lindsey Joost, *The Place for Illusions: Deepfake Technology and the Challenges of Regulating Unreality*, 33 U. FLA. J.L. & PUB. POL'Y 309, 329 (2023) (stating that "Section 230 essentially leaves [deepfake] victims with no practical recourse"); Guzman, *supra* note 257, at 136 (describing the "super-immunity" that Section 230 gives VLOPs as an "almost impenetrable . . . shield"); O'Donnell, *supra* note 257, at 715 ("Existing legislation and categories of civil liability focus entirely on [deepfakes'] creators, with no responsibility imposed on social media companies."). In fact, this impact of Section 230 is so widely known that deepfake victims rarely even attempt to sue VLOPs for hosting such content. Tenzer, *supra*, at 18 n.128 ("Today, section 230 is so commonly understood that plaintiffs' causes of actions against individuals posting defamatory statements typically do not name the social media sites as parties to the lawsuit.").

[272] *E.g.*, O'Donnell, *supra* note 257, at 738 (calling for a "narrow, deepfake-specific amendment to Section 230"); Guzman, *supra* note 257, at 136 (saying Section 230 "should be reasonably amended to protect the individual liberties violated by deepfake pornography"); Gieseke, *supra* note 257, at 1509 (proposing "amending section 230 to allow victims to sue platforms for unlawfully hosting deepfake pornography"); *see also* Engler, *supra* note 257 ("A growing number of scholars are calling for qualifications on Section 230."); Morar, *supra* note 161 ("[G]utting Section 230 has unfortunately risen to the top of the pile of legislative propositions in the United States.").

[273] *Section 230*, ELEC. FRONTIER FOUND., https://www.eff.org/issues/cda230 [https://perma.cc/ADR9-VNC2] (last visited Nov. 18, 2023) ("When Section 230 was passed in 1996, about 40 million people used the Internet worldwide. By 2019, more than 4 billion people were online, with 3.5 billion of them using social media platforms. In 1996, there were fewer than 300,000 websites; by 2017, there were more than 1.7 billion.").

[274] Gieseke, *supra* note 257, at 1495 (claiming that "section 230 has been lionized to mythic status" to offer VLOPs "near-unlimited immunity from any user posts"); Guzman, *supra* note 257, at 136 ("[T]he internet has grown beyond [Section 230]'s original intent.").

deepfake content,[275] these critics are correct.[276] However, their arguments fail to account for the fact that modifying Section 230 could have dramatic downstream effects for the internet as we know it.[277]

It is improvident to modify Section 230 because this law has always been, and will continue to be, vital for the development of the internet as we know it.[278] Section 230's immunity shield has almost singlehandedly enabled the development of platforms that host user-generated content.[279] Section 230 is necessary because it would be impossible for VLOPs to accurately monitor and filter the user-generated content they host.[280] Without this shield, VLOPs

[275] O'Donnell, *supra* note 257, at 727 ("Ideally, any law would hold liable the medium through which deepfakes are distributed (i.e., online platforms). It is impossible to hold platforms liable, however, if Section 230 continues to grant them immunity.").

[276] This Article, after all, is calling on Congress to require VLOPs to take actions addressing the spread of deepfakes.

[277] Villasenor, *supra* note 75 ("[A]ttempts to weaken section 230 of the CDA in the name of addressing the threat posed by deepfakes would create a whole cascade of unintended and damaging consequences to the online ecosystem.").

[278] *Section 230*, *supra* note 273 ("The free and open internet as we know it couldn't exist without Section 230.").

[279] Johnson & Castro, *supra* note 260 ("[I]t is clear that many common features of websites, such as user reviews and comments, exist because of the liability protection offered by Section 230.").

[280] Consider the following:

> With how active Internet users are on social media, it's nearly impossible for these platforms to remove every potentially unlawful or offensive post. For example, every sixty seconds, Facebook's billions of users post an average of 510,000 comments, 292,000 status updates, and 136,000 photos. Facebook has 15,000 content reviewers working around the world to remove posts that violate the website's community standards, yet still some posts manage to slip through the cracks. . . .
>
> In a world without Section 230 protections, online services would have to be aware of every post on their websites and make the correct content moderation decision 100 percent of the time to avoid liability for their users' speech. If, as the Supreme Court ruled . . . it's unreasonable to expect a bookseller to know the contents of a few hundred books, it's even more unreasonable to expect the operators of a social media platform to know the contents of thousands, millions, or even billions of posts, even when using thousands of human moderators and advanced algorithms.

would risk facing innumerable lawsuits for hosting content they did not create.[281] "The mere prospect of such lawsuits would force [VLOPs] to reduce or entirely prohibit user-generated content."[282] That would mean no social media, no product reviews, and no Wikipedia.[283] Even email services would be at risk.[284] The profundity of the implications stemming from a modification of Section 230 cannot be overstated.

By implementing a measure akin to the EU's, the U.S. would avoid all the problems associated with modifying or eliminating Section 230. Given the importance of these "26 words that created the internet,"[285] it is unwise to modify Section 230 when other, less risky measures are available for preventing the harmful spread of deepfakes.

Johnson & Castro, *supra* note 260 (footnotes omitted).

[281] Barbara Ortutay, *What You Should Know About Section 230, The Rule that Shaped Today's Internet*, PBS NEWS HOUR (Feb. 21, 2023, 10:55 AM), https://www.pbs.org/newshour/politics/what-you-should-know-about-section-230-the-rule-that-shaped-todays-internet [https://perma.cc/49V6-95AS] ("[Section 230] shields companies that can host trillions of messages from being sued into oblivion by anyone who feels wronged by something someone else has posted . . . .").

[282] Daniel Funke, *Meet Section 230: 'The Most Important Law Protecting Internet Speech'*, POYNER INST.: POLITIFACT (Mar. 1, 2021), https://www.politifact.com/article/2021/mar/01/meet-section-230-most-important-law-protecting-onl/ [https://perma.cc/B953-5VM6]; *Section 230*, *supra* note 273 ("Without Section 230's protections, many online intermediaries would intensively filter and censor user speech, while others may simply not host user content at all."); Ortutay, *supra* note 281 (noting that while communicating with other people is the "primary thing we do on the internet," VLOPs would let people use their platforms if they could be held liable for the content of these conversations).

[283] Johnson & Castro, *supra* note 260.

[284] *How Section 230 Helped Shape Speech on the Internet*, U.S. NEWS & WORLD REP. (Feb. 21, 2023, 6:56 AM), https://www.usnews.com/news/business/articles/2023-02-21/what-is-section-230-the-rule-that-made-the-modern-internet [https://perma.cc/H55Y-HYP9].

[285] JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET (2019).

VII. CONCLUSION

Deepfakes are a major problem in the U.S., and they are likely to become increasingly harmful in the immediate future as deepfake generators improve and become more accessible. The EU's DSA and CPD are likely to be effective at mitigating the harms of this problem, and the co-regulatory mechanisms employed in the Act are generally effective at addressing issues in rapidly evolving fields. Domestically, there is no federal legislation in place dealing with deepfakes, and the small patchwork of existing state regulations deal with only niche types of deepfakes. Removing Section 230 immunity would be a remarkably ill-advised way to attempt solving the deepfake problem in the U.S. Instead, it is necessary for Congress to enact new legislation that is comprehensive and broad.

In light of all of this, the U.S. should adopt a measure modeled off the DSA to counter deepfakes. Like the DSA, the American legislation should create a regulatory scheme that targets VLOPs and develops a system of risk assessments, mitigation measures, and audits. These should all be co-regulatory mechanisms, as they are in the DSA, because such mechanisms have proven to be effective means of quickly implementing meaningful regulations that can adapt to rapidly evolving technologies. Like the EU legislation, the U.S. should allow for mitigation to be demonstrated by compliance with a voluntary code, and that code should be largely modeled on the CPD. That is, the code should require VLOPs to develop and implement detectors and authenticators, allow for user reporting of deepfakes, flag and label deepfake content, warn users of past interactions with later-discovered deepfakes, and communicate with other VLOPs on trends in deepfakes on their own platforms—all in accordance with stringent and up-to-date industry standards.

As AI continues to grow and evolve at exponential rates,[286] it will become increasingly important to have a regulatory structure that is flexible and nimble enough to keep up. The EU's DSA and CPD fit the bill. Instead of trying to reinvent the wheel, the U.S. should consider using these as models when crafting its own legislation.

[286] Brown, *supra* note 64 (finding that "AI computing power has doubled every 3.4 months" since 2012, "dwarf[ing] Moore's Law").