AN AI'S PICTURE PAINTS A THOUSAND LIES: DESIGNATING RESPONSIBILITY FOR VISUAL LIBEL

Jon M. Garon *

INTRODUCTION

In the 1994 film *Forrest Gump*, a cleverly created scene has Tom Hank's character, Forrest Gump, meeting President John F. Kennedy.[1] The newsreel voice-over begins: "President Kennedy met with the members of the all-collegiate football team today in the Oval Office." The narration is picked up by Gump: "Now the really good thing about meeting the President of the United States is the food. . . . I must have drunk me about fifteen Doctor Peppers."[2] By the time it is his turn to meet the President, however, the sodas have taken their toll on an increasingly anxious Gump. Kennedy is seen asking most players, "How does it feel to be an All-American?"[3] To Gump, he simply says, "How do you feel," to which Gump

* Professor of Law, Nova Southeastern University Shepard Broad College of Law and Director of the Goodwin Program for Society, Technology, and the Law. Special thanks to Eugene Volokh and other authors of this special symposium collection of articles for their insights and feedback. Additional thanks to Cheryl Booth for her assistance.

[1] FORREST GUMP (Paramount Pictures 1994) (directed by Robert Zemeckis, screenplay by Eric Roth, based on the novel by Winston Groom).

[2] *Id.*

[3] *Id.*

answers honestly, "I gotta pee." Kennedy laughs, commenting to the reporters, "I believe he said he had to go pee."[4] This famous interaction between the fictional character and the long-dead president remains shocking in its apparent—but illusory—authenticity.

Two decades later, the technology to construct such scenes has gone from a feat of amazing cinematographic wizardry to common internet filler.[5] Kendrick Lamar used deepfake technology to morph his image into that of "O.J. Simpson, Kanye West, Jussie Smollett, Will Smith, Kobe Bryant, and Nipsey Hussle."[6] In March 2023, a photograph of "Pope Francis wearing a big white Balenciaga-style puffer jacket" became an internet staple.[7] Unsurprisingly, synthetic media has also been used for military disinformation.[8] In the Russian war against Ukraine, a video depicting Ukrainian President Volodymyr Zelenskyy ordering Ukrainian troops to lay down their arms and surrender appeared both on social media and broadcast briefly on Ukrainian news.[9] Some synthetic content has already found commercial adoptions such as the replacement of South Korean news anchor Kim Joo-Ha with a synthetic look-alike on South Korean television channel MBN,[10] or one company's introduction of internet influencer Lil Miquela, an alleged nineteen-year-old, as their spokesperson. In reality, Miquela is an entirely artificial avatar created

---

[4] *Id.*

[5] *See* Joseph Foley & Abi Le Guilcher, *20 of the Best Deepfake Examples that Terrified and Amused the Internet*, CREATIVE BLOQ (Mar. 10, 2023), https://perma.cc/HHQ4-9NX4.

[6] Elijah C. Watson, *Ranking Kendrick Lamar's "The Heart Part 5" Deepfakes from Least to Most Bizarre*, OKAYPLAYER (Mar. 2022), https://perma.cc/6N9D-U4V6.

[7] Sara Morrison, *How Unbelievably Realistic Fake Images Could Take Over the Internet*, VOX (Mar. 30, 2023), https://perma.cc/2E43-243R.

[8] *See* Bobby Allyn, *Deepfake Video of Zelenskyy Could Be 'Tip of the Iceberg' in Info War, Experts Warn*, NPR (Mar. 16, 2022).

[9] *Id.*

[10] *See* Yoon So-Yeon, *MBN Introduces Korea's First AI News Anchor*, KOREA JOONANG DAILY (Nov. 10, 2020), https://perma.cc/84WB-9VLG ("MBN revealed that an AI anchor based on announcer Kim Ju-ha started reporting news on Nov. 6, on the channel's 'MBN News.' AI Kim was almost identical to the real Kim, both in her looks and the sound of her voice . . . ."); Bernd Debusmann Jr., *Deepfake Is the Future of Content Creation*, BBC NEWS (Mar. 8, 2021) ("Viewers had been informed beforehand that this was going to happen, and South Korean media reported a mixed response after people had seen it. While some people were amazed at how realistic it was, others said they were worried that the real Kim Joo-Ha might lose her job.").

by AI media agency Brud.[11] She has over 3 million Instagram followers and has participated in brand campaigns since 2016.[12] She is expected to earn Brud in excess of $1 million in the coming year for her sponsored posts.[13]

"Over a few short years, technology like AI and deepfaking has advanced to the point where it's becoming really quite difficult to see the flaws in these creations."[14] Nor does it necessarily require artificial intelligence technologies to create false narratives from realistic-looking photographs and videos. "Sharing deceptive photos or misinformation online doesn't actually require a lot of talent. Often, just cropping a photo or video can create confusion on social media."[15] As the FTC has recently noted, "Thanks to AI tools that create 'synthetic media' or otherwise generate content, a growing percentage of what we're looking at is not authentic, and it's getting more difficult to tell the difference. And just as these AI tools are becoming more advanced, they're also becoming easier to access and use."[16]

The release of OpenAI's Dall-E 2, Stability AI's Stable Diffusion, and Midjourney Lab's Midjourney image generator dramatically expanded the universe for synthetic imagery generated entirely by text prompts rather than by feeding the computer system preexisting pictures and videos.[17] In the earlier AI training models, the deepfakes were created primarily by generative adversarial networks (GANs), a form of unsupervised machine learning in which a generator input competes with an "adversary, the discriminator network" to distinguish between real and artificial

---

[11] Meredith Clark, *PacSun Sparks Backlash After Announcing Virtual Influencer Lil Miquela As Its Newest Ambassador*, INDEPENDENT (Aug. 17, 2022).

[12] *Id.*

[13] *Id.*

[14] Foley & Guilcher, *supra* note 5.

[15] *Why the Way an Image Is Cropped Can Change Everything*, FRANCE 24, THE OBSERVERS (Aug. 24, 2018).

[16] Michael Atleson, *Chatbots, Deepfakes, and Voice Clones: AI Deception for Sale*, FTC BUSINESS BLOG (Mar. 20, 2023), https://perma.cc/Z2Q8-3QWK.

[17] *See* Vivek Muppalla & Sean Hendryx, *Diffusion Models: A Practical Guide*, SCALE (Oct. 19, 2022), https://perma.cc/B322-RJ5Z (Dall-E 2 was launched in April 2022; Imagen, from Google, was released in May 2022 but has not yet been made available to the general public; Stable Diffusion was made publicly available in August 2022).

images.[18] In contrast, the more recently adopted diffusion model of training involves the use of adding noise to the images to train the system to identify visual elements from the competing data.[19] The diffusion models are similar to that of large language models used for OpenAI's ChatGPT, Google's Bard, and other text-based AI platforms.[20] The diffusion model and similar systems enable the AI to build original images or video from text-based prompts rather than requiring the user to input a source image.[21] One could even daisy-chain systems so that the text prompts were themselves AI generated in the first instance.

There has been significant scholarship on the threats of deepfakes and synthetic media to political discourse and journalism,[22] as well as the potential for individuals to disseminate libelous material about others and even make terroristic threats

---

[18] Jason Brownlee, *A Gentle Introduction to Generative Adversarial Networks (GANs)*, MACHINE LEARNING MASTERY (June 17, 2019), https://perma.cc/Y9FS-5MQF.

[19] Muppalla & Hendryx, *supra* note 17.

[20] *See* Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 J. FREE SPEECH L. 389, 393 (2023) ("Large language models ('LLMs') are a type of deep learning algorithm used to model statistical relationships between words and phrases in large bodies of text data in order to generate human-like language. (ChatGPT is a type of LLM.)").

[21] *See* Will Douglas Heaven, *The Original Startup Behind Stable Diffusion Has Launched a Generative AI For Video*, MIT TECH. REV. (Feb. 6, 2023).

[22] *E.g.*, Regina Rini & Leah Cohen, *Deepfakes, Deep Harms*, 22 J. ETHICS & SOC. PHIL. 143, 148 (2022) ("Deepfakes do not have to trick anyone in order to be harmful. Even if a deepfake is ultimately debunked, or never believed at all, it can still hurt the person it falsely depicts by changing the discursive context around them."); Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 VA. J.L. & TECH. 1, 12 (2020) ("Even the most ardent supporters of journalism will be forced to question the authenticity of video and audio footage relied on by journalists, without additional evidence that the depicted events occurred. When people cannot tell the difference between what is true and false, it reduces trust in traditional media . . . ."); Jared Schroeder, *Free Expression Rationales and the Problem of Deepfakes Within the E.U. and U.S. Legal Systems*, 70 SYRACUSE L. REV. 1171, 1179 (2020) (addressing two harms: "(1) The creation of widely believable false statements or facts that mislead and distort the search for consensus or shared truth in society and (2) the unique power of video clips to provide credibility to false messages and to enrage and enflame citizens into action"); Holly Kathleen Hall, *Deepfake Videos: When Seeing Isn't Believing,* 27 CATH. U. J.L. & TECH. 51, 52 (2018) ("The extraordinary success of fake news being accepted in the marketplace creates grave concerns for individuals and democracy. This is exacerbated when a video is added to the equation.").

using these images and videos.[23] Given the generative AI's ability to create AI-authored original works, there is a rather new concern that the AI system will itself create works that harm individuals and the public. As with potential risks associated with ChatGPT, images generated by AI systems may have unintended and highly inaccurate content.[24]

This article focuses on responsibility and liability for libelous publication of generative synthetic media. It summarizes the textbook example of a person creating intentionally false depictions of the victim with the purpose of holding that

---

[23] *E.g.*, Cass R. Sunstein, *Falsehoods and the First Amendment*, 33 HARV. J.L. & TECH. 387, 405 (2020) ("To be sure, many falsehoods can be harmful even if they do not fall in the traditional categories. . . . If social welfare is the goal, we would want to measure the benefits against the costs of allowing the false statement in question, or perhaps the category of statements of which it is a part."); Jon M. Garon, *When AI Goes to War: Corporate Accountability for Virtual Mass Disinformation, Algorithmic Atrocities, and Synthetic Propaganda*, 49 N. KY. L. REV. 181, 192 (2022) ("The expansion and reliance on artificial intelligence technologies are among the greatest advances in the twenty-first century and greatest threats for misuse."); Rachel E. VanLandingham, *Jailing the Twitter Bird: Social Media, Material Support to Terrorism, and Muzzling the Modern Press*, 39 CARDOZO L. REV. 1, 13 (2017) ("The link between social media platforms and terrorism competes with privacy concerns as one of the most discussed and most concerning, dynamics emanating from modern society's explosive utilization of these communication technologies.").

[24] *See* Rachel Metz, *AI Made These Stunning Images. Here's Why Experts Are Worried*, CNN BUSINESS (Aug. 2, 2022) ("The technology can perpetuate hurtful biases and stereotypes. They're concerned that . . . their ability to automate image-making means they could automate bias on a massive scale. They also have the potential to be used for nefarious purposes, such as spreading disinformation."); Kevin Jiang, *These AI Images Look Just Like Me. What Does That Mean for the Future of Deepfakes?*, TORONTO STAR (Dec. 1, 2022) ("Google's DreamBooth research has led to a breakthrough in AI art, enabling anyone to create digital replicas of real people. Experts are concerned over its implications for misuse."); Isaac Stanley-Becker & Drew Harwell, *How a Tiny Company With Few Rules Is Making Fake Images Go Mainstream*, WASH. POST (Mar. 30, 2023) ("Midjourney 'has unchecked authority to determine how those powers are used. It allows, for example, users to generate images of President Biden, Vladimir Putin of Russia and other world leaders—but not China's president, Xi Jinping.'"). *See also* Thomas Macaulay, *These Laughable Depictions of AI Can Have Serious Consequences*, NEXT WEB (June 17, 2021), https://perma.cc/T6AU-D5EF (noting systemic racial and gender biases in the images used to illustrate robots and artificial intelligence as white male dominated); Peter Henderson, Tatsunori Hashimoto & Mark Lemley, *Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589, 592 (2023) ("Because language models are good at modeling human writing, they pepper their false reports of crimes with the same things a real report would include—including (made up) quotations from reputable sources (whose articles are also made up).").

individual out for hatred, contempt, or ridicule. The article then compares that example to the situation in which the AI system itself generated the content to identify who among the parties that published the libelous images might face civil liability for that publication. Would an owner of the AI system, the platform on which the AI system was operating, the individual who created the prompts that generated the offensive imagery, or no one be liable? By providing this framework, the article should also identify the steps that can be taken by the parties involved in the AI content production chain to protect individuals from the misuse of these systems.

I. Visual Libel

The legal tests for libel vary slightly from jurisdiction to jurisdiction, but California provides a representative example. In California, "[l]ibel is a false and unprivileged publication by writing, printing, picture, effigy, or other fixed representation to the eye, which exposes any person to hatred, contempt, ridicule, or obloquy, or which causes him to be shunned or avoided, or which has a tendency to injure him in his occupation."[25] To state a claim for libel in California, the plaintiff must "allege (1) a publication that is (2) false, (3) defamatory, (4) unprivileged, and (5) has a natural tendency to injure or causes special damage."[26] California's Anti-SLAPP statute further protects speakers by promoting the early dismissal of unmeritorious claims. "A cause of action against a person arising from any act of that person in furtherance of the person's First Amendment right of petition or free speech in connection with a public issue shall be subject to a special motion to strike, unless the court determines that the plaintiff has established that there is a probability that the plaintiff will prevail on the claim."[27]

The statute itself specifies that a "picture, effigy, or other fixed representation to the eye" falls within the definition of libel.[28] Many of the potentially libelous pictures were created in the form of political cartoons upon which the Supreme Court commented in *Hustler Magazine, Inc. v. Falwell*.[29]

---

[25] Cal. Civ. Code § 45.

[26] Stossel v. Meta Platforms, Inc., 21-CV-07385-VKD, 2022 WL 6791430, at *6 (N.D. Cal. Oct. 11, 2022).

[27] Cal. Code Civ. Proc. § 425.16(b)(1).

[28] Cal. Civ. Code § 45.

[29] Hustler Mag., Inc. v. Falwell, 485 U.S. 46 (1988).

> The appeal of the political cartoon or caricature is often based on exploitation of unfortunate physical traits or politically embarrassing events—an exploitation often calculated to injure the feelings of the subject of the portrayal. The art of the cartoonist is often not reasoned or evenhanded, but slashing and one-sided. One cartoonist expressed the nature of the art in these words:
>
> "The political cartoon is a weapon of attack, of scorn and ridicule and satire; it is least effective when it tries to pat some politician on the back. It is usually as welcome as a bee sting and is always controversial in some quarters."[30]

Despite the Supreme Court's fond characterizations of cartoonist Thomas Nast and others, plaintiffs occasionally scored victories against political cartoonists prior to *New York Times Co. v. Sullivan*.[31] In *Times v. Sullivan*, the Supreme Court instituted constitutional protections against libel actions. "[A] public figure may hold a speaker liable for the damage to reputation caused by publication of a defamatory falsehood, but only if the statement was made 'with knowledge that it was false or with reckless disregard of whether it was false or not.'"[32] The various procedural and substantive protections afforded speakers are summarized by the Supreme Court in *Milkovich v. Lorain Journal Co*.[33] These protections include requiring the plaintiff to establish the fault, eliminating the common-law presumption that defamatory speech is false, limiting claims of libel to assertions of verifiable fact, excluding hyperbole, and promoting a vigorous and robust political debate.[34] *Milkovich* also helps clarify what is meant by an assertion of fact. "The dispositive question . . . becomes whether a reasonable factfinder could conclude that the statements . . . imply an assertion that . . . is sufficiently factual to be susceptible of being proved true or false."[35]

---

[30] *Id*. at 54 (quoting Long, *The Political Cartoon: Journalism's Strongest Weapon,* THE QUILL 56, 57 (Nov. 1962)).

[31] New York Times Co. v. Sullivan, 376 U.S. 254, 285–86 (1964) (actual malice standard must be established by clear and convincing evidence upon an independent examination of the whole record.). Cases involving libel by political cartoons include *Snively v. Record Pub. Co.*, 185 Cal. 565 (1921), and *Newby v. Times-Mirror Co.*, 46 Cal. App. 110 (1920). *See generally* Gregory R. Naron, *With Malice Toward All: The Political Cartoon and Law of Libel*, 15 COLUM.-VLA J.L. & ARTS 93 (1990).

[32] *Hustler*, 485 U.S. at 52 (quoting *New York Times Co. v. Sullivan*, 376 U.S. at 279–80).

[33] Milkovich v. Lorain J. Co., 497 U.S. 1 (1990).

[34] *Id*. at 16–22.

[35] *Id*. at 21.

For photorealistic pictures and videos, the concept of fact is not self-evident. Still photography and audiovisual films create the impression that they capture what is in front of the lens, but this is rarely the whole truth.[36] "In some cases cropping can radically transform the meaning of a shot."[37] Without any retouching, photographers could always manipulate the message being conveyed using framing choices when taking a picture or cropping choices when developing the photograph. Where a political cartoon would inform the viewer that it was the impression of the artist, a photograph or video asserts that the images are exactly as they appeared.[38] This assertion of accuracy makes a photorealistic image appear self-validating as to its truthfulness.[39] When applying the standard of *Milkovich*, the factual depiction shown in a photorealistic image or video is the assertion of fact "susceptible of being proved true or false."[40]

---

[36] *See* Virginia Seymour, *Exploring Images in (and out of) Context*, JSTOR DAILY (Nov. 9, 2022), https://perma.cc/KG7L-TK8Q ("Conversations around manipulation of information in images focuses largely on content manipulation—where images, often photographs, are purposefully edited—but missing or misleading context, even when it is unintentional, can just as easily skew the interpretation of images.").

[37] John Suler, *Cropping and the Frame*, in JOHN SULER, PHOTOGRAPHIC PSYCHOLOGY: IMAGE AND PSYCHE, https://perma.cc/6BH4-XMYS.

[38] Lisa Fazio, *Out-of-Context Photos Are a Powerful Low-Tech Form of Misinformation*, THE CONVERSATION (Feb. 14, 2020), https://perma.cc/2R9E-H4CA ("images are a powerful tool for swaying popular opinion and promoting false beliefs. Psychological research has shown that people are more likely to believe true and false trivia statements, such as 'turtles are deaf,' when they're presented alongside an image").

[39] *See id.*:

> There are a number of reasons photographs likely increase your belief in statements.
>
> First, you're used to photographs being used for photojournalism and serving as proof that an event happened.
>
> Second, seeing a photograph can help you more quickly retrieve related information from memory. People tend to use this ease of retrieval as a signal that information is true.
>
> Photographs also make it more easy to imagine an event happening, which can make it feel more true.
>
> Finally, pictures simply capture your attention. A 2015 study by Adobe found that posts that included images received more than three times the Facebook interactions than posts with just text.

[40] *Milkovich*, 497 U.S. at 21.

Even though courts and visually literate observers know that photography is not self-validating, it is often very hard to discern truth from deception. "For artistic purposes, . . . alterations of the image may be perfectly acceptable. In the case of photojournalism or other images that should be presenting a factual reality, the crop could be an objectionable act of deception."[41] As a consequence, falsified photographs and videos are more likely to be assumed to be accurate than textual information and are more likely to be taken at face value than other forms of falsehoods.

The Supreme Court has addressed an analogous situation when dealing with falsely attributed quotations. In *Masson v. New Yorker*,[42] the Court explained the unique harm that can come from falsely attributed content.

> A fabricated quotation may injure reputation in at least two senses, either giving rise to a conceivable claim of defamation. First, the quotation might injure because it attributes an untrue factual assertion to the speaker. . . . Second, regardless of the truth or falsity of the factual matters asserted within the quoted statement, the attribution may result in injury to reputation because the manner of expression or even the fact that the statement was made indicates a negative personal trait or an attitude the speaker does not hold.[43]

Photorealistic images that assert their accuracy to the viewer and fail to inform the viewer that the image is not an actual depiction are much like the quote that is falsely attributed to a speaker. Separate from its falsity, it also falsely suggests the person depicted was a voluntary participant in the taking of the photograph (or at least caught on camera doing the action which is depicted). Like the false quote, the engagement of the victim and the attribution to the victim creates a second level of falsity.[44]

---

[41] Suler, *supra* note 37.

[42] Masson v. New Yorker Mag., Inc., 501 U.S. 496 (1991).

[43] *Id.* at 511.

[44] *Id.* at 512 ("A self-condemnatory quotation may carry more force than criticism by another. It is against self-interest to admit one's own criminal liability, arrogance, or lack of integrity, and so all the more easy to credit when it happens.").

In addition, while many high-profile libel cases involve public officials or public figures, not all do so.[45] *Gertz v. Robert Welch, Inc.*[46] required only a negligence standard in the case of private figures. It defined public figures as follows:

> For the most part, those who attain this status have assumed roles of especial prominence in the affairs of society. Some occupy positions of such persuasive power and influence that they are deemed public figures for all purposes. More commonly, those classed as public figures have thrust themselves to the forefront of particular public controversies in order to influence the resolution of the issues involved.[47]

The precise dividing line between private individual and public figure remains elusive. In *Time, Inc. v. Firestone*,[48] however, the court made clear that participation in a lawsuit is insufficient to make one a public figure, thus refusing "to equate 'public controversy' with all controversies of interest to the public."[49]

The Court refused to extend the actual malice standard "to falsehoods defamatory of private persons whenever the statements concern matters of general or public interest."[50] To add to the subjective nature of the public figure standard, some individuals may be limited-purpose public figures, who "'voluntarily inject[ ]' themselves or are 'drawn into a particular public controversy' and thereby become public figures 'for a limited range of issues' defined by their 'participation in the particular controversy giving rise to the defamation.'"[51] The controversy at issue "must predate the alleged defamation."[52]

In *Milkovich*, for example, the plaintiff was a high school wrestling coach.[53] He was only adjudicated a private plaintiff after the Ohio Supreme Court reversed a

---

[45] *See* Curtis Publishing Co. v. Butts, 388 U.S. 130 (1967) (extending the actual malice standard to public figures); Associated Press v. Walker, 388 U.S. 130 (1967) (same).

[46] Gertz v. Robert Welch, Inc., 418 U. S. 323 (1974).

[47] *Id.* at 345. *See also* Time, Inc. v. Firestone, 424 U.S. 448 (1976).

[48] *Time, Inc. v. Firestone*, 424 U.S. 448 (1976).

[49] *Id.* at 454.

[50] *Id.*

[51] Ayyadurai v. Floor64, Inc., 270 F. Supp. 3d 343, 357 (D. Mass. 2017) (quoting Lluberes v. Uncommon Prods., LLC, 663 F.3d 6, 13 (1st Cir. 2011); *Gertz*, 418 U.S. at 351–52).

[52] *Lluberes*, 663 F.3d at 14.

[53] Milkovich v. Lorain J. Co., 497 U.S. 1 (1990).

lower court determination to the contrary.[54] Courts often gravitate to the actual malice standard by using the public figure or limited public figure labels.[55] And in the age of social media and influencers, easy-to-achieve public figure status may well have swallowed the distinctions espoused in *Gertz*. As a consequence, Justice Gorsuch has suggested that the radical changes in the public media landscape have necessitated a reexamination of the public figure standard, and perhaps the actual malice standard as well.[56] AI only increases this concern. The ease with which AI-generated images may be propounded as factual depictions and spread across social media and into mass media may require the courts to rethink the interpretations of

---

[54] *Id*. at 8 (the Ohio Supreme Court "first decided that petitioner was neither a public figure nor a public official under the relevant decisions of this Court").

[55] *Cf*. Wolston v. Reader's Dig. Ass'n, Inc., 443 U.S. 157, 167 (1979) (reversing a lower court finding that the plaintiff/petitioner had become a public figure because he was convicted for contempt after failing to respond to a grand jury subpoena).

[56] *See* Berisha v. Lawson, 141 S. Ct. 2424, 2427–29 (2021) (Gorsuch, J., dissenting) (citations omitted):

> Since 1964, however, our Nation's media landscape has shifted in ways few could have foreseen. . . . [T]hanks to revolutions in technology, today virtually anyone in this country can publish virtually anything for immediate consumption virtually anywhere in the world. The effect of these technological changes on our Nation's media may be hard to overstate. . . .
>
> It's hard not to wonder what these changes mean for the law. . . . The bottom line? It seems that publishing without investigation, fact-checking, or editing has become the optimal legal strategy. . . .
>
> Other developments raise still more questions. In 1964, the Court may have thought the actual malice standard would apply only to a small number of prominent governmental officials whose names were always in the news and whose actions involved the administration of public affairs. . . . Now, private citizens can become "public figures" on social media overnight. Individuals can be deemed "famous" because of their notoriety in certain channels of our now-highly segmented media even as they remain unknown in most.

But see Ballard Spahr LLP & Davis Wright Tremaine LLP, *Chapter 4: The Reality of Contemporary Libel Litigation*, in NEW YORK TIMES V. SULLIVAN: THE CASE FOR PRESERVING AN ESSENTIAL PRECEDENT (Media Law Resource Center 2022), https://perma.cc/7MLQ-4RTF:

> But an analysis of defamation cases over the last twenty-five years demonstrates that these criticisms are unwarranted. *Sullivan* has not dissuaded public officials or public figures from bringing libel suits; to the contrary, in our experience, the last decade has seen a significant number of these kinds of cases. Nor does the actual malice standard act as an absolute (or even near-absolute) bar to these kinds of claims getting before a jury.

the public figure label, the parameters of the public concern constraint espoused in the cases extending *Times v. Sullivan*, and the appropriate contours of the actual malice standard for matters unrelated to robust and wide-open debate.[57]

II. The Culpability Matrix

For online libelous imagery, the question remains which party or parties should be culpable for the harms caused by the publication of the content. These parties may include the persons who created the content; the persons who posted the content to online sites; the sites that hosted the content; and any persons or services that repurposed that content onto other websites or services.

Nonconsensual pornography,[58] including "revenge porn,"[59] provides a paradigmatic example of the problem.[60] Revenge porn is commonly defined as unauthorized publication of consensually created content, which distinguishes it from the broader category of nonconsensual pornography.[61] A smaller category of nonconsensual pornography refers to content that was surreptitiously captured by the

---

[57] *See* Michael Norwick, *Chapter 3: The Empirical Reality of Contemporary Libel Litigation* in New York Times v. Sullivan: The Case for Preserving an Essential Precedent (Media Law Resource Center 2022), https://perma.cc/H6WZ-BPBD:

> The fact is that there are a wide range of speech-protective elements and defenses to a libel cause of action unrelated to actual malice, that are commonly utilized to defend cases brought against the media. Some of the most common libel defenses are the fair report privilege, opinion, rhetorical hyperbole, defamatory meaning, the "of and concerning" requirement, and substantial truth.

[58] *See* Chance Carter, *An Update on the Legal Landscape of Revenge Porn*, Nat. Ass'n of Att'ys Gen. (Nov. 16, 2021), https://perma.cc/ZC4S-TMSH.

[59] *Id.* ("Eighty percent of nonconsensual porn is revenge porn, meaning it was originally sent between two consenting individuals in the context of an intimate relationship. Often, this content is posted alongside the victim's name and other identifying information such as their phone numbers, emails, or social media links.").

[60] *See* Danielle Keats Citron, *Sexual Privacy*, 128 Yale L.J. 1870, 1874 (2019) ("Sexual privacy sits at the apex of privacy values because of its importance to sexual agency, intimacy, and equality.").

[61] *See* Carter, *supra* note 58.

perpetrator.[62] Existing laws now prohibit the publication of revenge porn.[63] For example, the crime (and tort) of revenge porn is now prohibited by specific statute in 48 states plus the District of Columbia.[64] The text of these laws varies considerably from one state to the next, and the specific statute of a particular state may or may not apply to or include AI-generated content. Indiana, for example, requires the image be "taken, captured, or recorded," making it unlikely that an image rendered by a computer to simulate the offending image would fall within the statute.[65]

As a tort, revenge porn falls into the category of invasion of privacy[66] while AI-generated look-alike revenge porn should be understood as libelous.[67] Unlike the unauthorized distribution of a sex tape that is accurate but highly invasive of one's

---

[62] *See id*. *See also* Jessica A. Magaldi, Jonathan S. Sales, & John Paul, *Revenge Porn: The Name Doesn't Do Nonconsensual Pornography Justice and the Remedies Don't Offer the Victims Enough Justice*, 98 OR. L. REV. 197, 215 (2020) ("[The] privacy or intrusion upon seclusion . . . tort is easily applicable to a hacker who invaded a victim's computer or a perpetrator who surreptitiously captured images of a victim. But establishing such a cause of action is more challenging to a victim that provided a selfie to a current or former romantic partner.").

[63] *See*, *e.g.*, People v. Bollaert, 203 Cal. Rptr. 3d 814, 819 (Ct. App. 2016) ("A jury convicted Kevin Christopher Bollaert of extortion . . . and the unlawful use of personal identifying information . . . stemming from his operation of Web sites, 'UGotPosted.com,' . . . and 'ChangeMyReputation.com,' through which victims could pay to have the information removed."); Rebecca Grant, *Accused Revenge Porn Site Owner Arrested and Charged with 31 Felony Counts*, VENTURE BEAT (Dec. 11, 2013) ("The state's attorneys office accused Bollaert of receiving and reviewing 2,000 emails from individuals, requesting their information and images be removed from the site [ugotposted.com], between June 20, 2103 and August 26 2013.").

[64] *See State Revenge Porn Laws*, FINDLAW (Dec. 8, 2022), https://perma.cc/3AMN-P8MX ("All states, excluding Massachusetts and South Carolina, have separate statutes that are specifically related to revenge porn. It's important to note, however, that a person may still be prosecuted for revenge porn under other statutes in those two states.").

[65] *See* IND. CODE § 35-45-4-8.

[66] State v. VanBuren, 210 Vt. 293, 322 (2019) (upholding Vermont's criminal statute as meeting the strict scrutiny standard due to the compelling privacy interests):

> In the constellation of privacy interests, it is difficult to imagine something more private than images depicting an individual engaging in sexual conduct, or of a person's genitals, anus, or pubic area, that the person has not consented to sharing publicly. The personal consequences of such profound personal violation and humiliation generally include, at a minimum, extreme emotional distress.

[67] *See generally* Eugene Volokh, *One-to-One Speech vs. One-to-Many Speech, Criminal Harassment Laws, and "Cyberstalking,"* 107 NW. U. L. REV. 731, 752–53 (2013).

privacy, a deepfake is a false depiction that factually asserts the participant engaged in videotaped or photographed sexual activity—perhaps for distribution on the internet. That will likely constitute a libelous statement for the vast majority of victims.[68]

When a person generates revenge porn using a GAN generated image, the process requires the person to upload an image of the victim and either synthesize the person's likeness into the pornographic video or to use a service that does so. The person utilizing the system, then, is directly responsible for the resulting content if that output is libelous. A person creating that content and publishing it to any website viewable to a third party would meet the requirement to be liable for defamation, provided it was reasonably understood by the viewer to be a true depiction.[69]

As noted in *Hustler Magazine, Inc. v. Falwell*, such liability will not attach when a "parody could not reasonably be understood as describing actual facts about respondent or actual events in which he participated."[70] Where the revenge porn is clearly labeled as parody or noted for depicting false images, those disclaimers may take it outside the realm of falsehood. Where there is no indication of known falsity, however, reasonable viewers may well believe their eyes. Moreover, the *Hustler* standard for intentional infliction of emotional distress was decided in the context of public figures involving matters of public concern.[71] In the context of a private individual involving matters that are not matters of public concern, some state courts have found the requirements of the intentional infliction of emotional distress claim create a sufficiently high bar that there is no need to preclude such cases

---

[68] *See* Citron, *supra* note 60 at 1937 ("Public officials and public figures like Gal Gadot could sue for defamation if there is clear and convincing evidence of actual malice (that is, the defendant knew the deep-fake sex videos were false or recklessly disregarded the possibility that they were false)."). *See also* Bobby Chesney & Danielle Citron, *Deep Fakes*, 107 CAL. L. REV. 1753 (2019).

[69] *See generally*, Brown, *supra* note 20, at 392 ("The default assumption may be that someone who is defamed by an AI chatbot would have a case for defamation. But there are hurdles in applying defamation law to speech generated by chatbot, particularly because defamation law requires assessing mens rea that will be difficult to assign to a chatbot (or its developers)").

[70] 485 U.S. at 57 (internal brackets and quotations omitted).

[71] *Id*. at 50 ("We must decide whether a public figure may recover damages for emotional harm caused by the publication of an ad parody offensive to him.").

by imposing the actual malice standard.[72] The limitations on revenge porn, which will not involve matters of public concern in most cases, therefore may remain outside of First Amendment protection.[73]

Turning to the website operators, most of them will not be liable for the resulting defamation. Under the broad immunity that interactive computer services (a/k/a web platforms) have from liability for third party content, the sites on which the images appear would not be liable for the content posted by third parties under the limits of liability provided by Section 230 of the Communications Decency Act.[74]

When a user creates pornography using deepfakes, the user is providing the AI both the images of the pornography and the face of the person being added into the pornographic or offensive image. In other methods of generative AI, the system may be pulling reference images from its database without the specific direction of

---

[72] *See e.g.*, State v. Carpenter, 171 P.3d 41, 58 (Alaska 2007); Esposito-Hilder v. SFX Broad., Inc., 665 N.Y.S.2d 697 (App. Div. 1997).

> To recover for intentional infliction of emotional distress, an IIED claimant must prove that there was "extreme and outrageous conduct" that intentionally or recklessly inflicted severe emotional distress. . . . Mere insults, indignities, threats, annoyances, petty oppressions or other trivialities" cannot form the basis of an IIED claim. . . . [E]ven harmful conduct characterized by 'malice' is insufficient to make out an IIED claim if the conduct is not "extreme and outrageous." An IIED claim is therefore arguably no easier to prove than a defamation claim, even a defamation claim that must satisfy the "actual malice" standard. In addition, because IIED requires proof of an intentional or reckless mental state, an IIED plaintiff must show that "the defendant acted in deliberate disregard of a high degree of probability that the emotional distress will follow."

Carpenter, 171 P.3d at 58 (internal quotations and citations omitted).

[73] *Cf.* Snyder v. Phelps, 562 U.S. 443, 458 (2011) ("Given that Westboro's speech was at a public place on a matter of public concern, that speech is entitled to 'special protection' under the First Amendment. Such speech cannot be restricted simply because it is upsetting or arouses contempt.")

[74] *See*, *e.g.*, Johnson v. Arden, 614 F.3d 785, 790–91 (8th Cir. 2010) ("The CDA states that '[n]o provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider,' 47 U.S.C. § 230(c)(1), and expressly preempts any state law to the contrary, *id.* § 230(e)(3)"); Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157, 1162–64 (9th Cir. 2008). *But cf.* Beatriz Botero Arcila, *Is It a Platform? Is It a Search engine? It's ChatGPT!*, 3 J. FREE SPEECH L. 455, 460 (2023) ("Unlike in the United States where there is at least disagreement on whether tools like ChatGPT are covered by section 230 or not, scholars in the EU don't think that LLM-chatbots fall naturally under the EU safe harbor for intermediary liability.").

the user. As a result, if instead of a deepfake, the false sexual depiction is created through generative AI using text prompts, the question of responsibility becomes less clear. For example, the website Mage, which launched in September 2022, harnesses the tools of Stable Diffusion to generate pornographic images based upon text prompts.[75] These use cases are increasing.[76] There is an entire community operating under the name "Unstable Diffusion" "that explores and experiments with [not suitable for work] AI-generated content using Stable Diffusion."[77] The group explains that "[w]e believe erotic art needs a place to flourish and be cultivated in a space without judgement or censorship."[78]

Stable Diffusion has been sued for copyright infringement for scraping millions of Getty Images photographs from the internet.[79] The litigation alleges that the training data for Stable Diffusion has been culled and scraped from billions of internet images without the permission of the owners of those images.[80] The AI is not, itself, an infringer under copyright law,[81] but the parties exploiting those tools will be. As a consequence, the Mage imagery and similar services may be using images to generate the pornographic content from those who never agreed to appear in such material. There is no reason to think that the participants in Unstable

---

[75] *See* Jim Clyde Monge, *This Website Can Generate NSFW Images with Stable Diffusion AI*, MEDIUM (Sept. 25, 2022), https://perma.cc/7DV6-FJEC.

[76] *See* Matt Burgess, *The Fightback Against AI-Generated Fake Pornography Has Begun*, WIRED (Aug. 2, 2018) ("Reddit's move to prohibit 'involuntary porn' follows moves by other sites and services to ban the sharing and hosting of such material. . . . Before Reddit deleted the community more than 90,000 people had subscribed to it.").

[77] *See* Monge, *supra* note 75.

[78] *Id.*

[79] *See* Getty Images (US) Inc. v. Stability AI, Inc., No. 1:23-cv-00135-UNA (D. Del. filed Feb. 3, 2023). *See also Getty Images Statement*, GETTY IMAGES (Jan. 17, 2023), https://perma.cc/FV3P-7BVZ (announcing litigation against Stability AI in the United Kingdom).

[80] *See id. See also* Gloria Levine, *Exploring the Images Used to Train Stable Diffusion's AI*, 80.LV (Sept. 7, 2022), https://perma.cc/5GNW-ZUVL ("Stable Diffusion was trained on 2.3 billion images.").

[81] *See* Robert A. Heverly, *AI, Creativity, and Liability for Direct Copyright Infringement* (working paper) ("AI systems do not have 'agency.' They do not make choices as we understand that word. Instead, they make predictions based on algorithmic equations and training data sets. This is the underlying reason that AI technologies cannot—and should not—be considered authors for copyright purposes.").

Diffusion are seeking to depict particular individuals captured in the Stable Diffusion training set, but since the data may potentially capture any adult or child ever depicted on a website, the potential to depict real people is quite high.[82] Moreover, "early on in the Stable Diffusion beta on its Discord server, testers found that almost every request for a 'beautiful woman' involved unintentional nudity of some kind, which reflects how Western society often depicts women on the Internet."[83] All the biases in the data set will naturally inform the AI-generated content. In addition, celebrity depictions can be generated by including the person's name in the prompt.[84]

The example of pornography is just one instance of potential visual libel. Another is the broad topic of civil unrest. If generative AI is used to depict images of violent demonstrations, then any actual persons falsely depicted in those images could also claim libel, if those images were published as factually accurate depictions. Consider this non-AI hypothetical. A photograph shows a university instructor standing amidst a crowd of yelling students. If that photograph were identified as coming from a controversial anti-administration event[85] or from an event where speakers were being silenced by a heckling crowd,[86] that instructor could face serious reprisal from the university. If the actual photograph had been taken at a pep

---

[82] *See* Levine, *supra* note 80 ("[R]esearch showed that almost half of the images were sourced from only 100 domains, with the largest number of images coming from Pinterest. Other sources include WordPress-hosted blogs, Smugmug, Blogspot, Flickr, DeviantArt, Wikimedia, 500px, and Tumblr. Shopping sites were also well-represented.").

[83] Benj Edwards, *With Stable Diffusion, You May Never Believe What You See Online Again*, ART TECHNICA (Sept. 6, 2022), https://perma.cc/WR9F-9G6E.

[84] *See*, *e.g.*, Damir Yalalov, *Best 100+ Stable Diffusion Prompts: The Most Beautiful AI Text-to-Image Prompts*, METAVERSE POST (Sept. 16, 2022), https://perma.cc/3P47-QYEK (depicting Keanu Reeves, Christina Hendricks, James Dean, Sadie Sink).

[85] *Cf.* Tyler Kingkade, *A Texas Teacher Faces Losing Her Job After Fighting for Gay Pride Symbols in School*, NBC News (Apr. 7, 2022) ("The school year at MacArthur High in Irving, Texas, began last fall with the administration scraping off rainbow stickers . . . . One faculty sponsor of the school's Gay-Straight Alliance is facing having her contract terminated, another is preparing to resign, and a third has been removed from the classroom.").

[86] *Cf.* Madison Alder, *Stanford Mandates Free Speech Training After Protest of US Judge*, BLOOMBERG (Mar. 22, 2023) ("Students protested Judge Kyle Duncan's talk on campus. Administrator who spoke at event put on leave, dean says. Stanford Law School is requiring all students to attend educational programming on free speech after protesters interrupted a speech by a conservative federal judge earlier this month.").

rally and mis-labeled because it was merely a good crowd shot, such misidentification could be highly injurious and potentially libelous. If instead, a blogger asked Dall-E or Stable Diffusion for a picture to illustrate the blogger's story about the event, the image generator could potentially generate such an image on its own. The image is no less injurious. But is the image less libelous?

Since Section 230 of the Communications Decency Act provides broad immunity from civil liability for defamatory content provided by third parties, the host site will be immune from action in most cases. The blogger who posted the offending image will be liable only if negligent in selecting the illustration as to those depicted individuals who are neither public officials nor public figures.[87] The blogger would only be liable under the actual malice standard for public officials and public figures. Certain facts might inform the determination of the applicable level of culpability of that blogger. One could certainly conclude that it is at least negligent to use any photograph to illustrate a story about a protest unless that image was actually from the protest itself. If a reasonable reader would be mistaken as to the authenticity of the image, then potentially, the action of the blogger could meet the "actual malice standard," "a term of art denoting deliberate or reckless falsification."[88] To meet the actual malice standard, a "plaintiff must demonstrate that the author 'in fact entertained serious doubts as to the truth of his publication,' or acted with a 'high degree of awareness of . . . probable falsity.'"[89] The use of the AI-generated image to accompany the news story arguably meets this awareness of

---

[87] *See, e.g.*, Levinsky's, Inc. v. Wal-Mart Stores, Inc., 127 F.3d 122, 128 (1st Cir. 1997) (applying negligence standard under state law and noting "[i]t is unclear whether the First Amendment prohibits a state from imposing strict liability in a defamation case brought by a private plaintiff concerning statements that implicate a matter of private concern"); Snead v. Redland Aggregates, Ltd., 998 F.2d 1325, 1334 (5th Cir. 1993). Even if the instructor was employed at a public university, employment alone would not require that the individual be treated as a public official. *See* Rosenblatt v. Baer, 383 U.S. 75, 85 (1966) ("the 'public official' designation applies at the very least to those among the hierarchy of government employees who have, or appear to the public to have, substantial responsibility for or control over the conduct of governmental affairs"); Hutchinson v. Proxmire, 443 U.S. 111, 135 (1979) ("Hutchinson [a research behavioral scientist] did not thrust himself or his views into public controversy to influence others. Respondents have not identified such a particular controversy; at most, they point to concern about general public expenditures; it is not sufficient to make Hutchinson a public figure.").

[88] Masson v. New Yorker Mag., Inc., 501 U.S. 496, 499 (1991).

[89] *Id.* at 510 (quoting St. Amant v. Thompson, 390 U. S. 727, 731 (1968); Garrison v. Louisiana, 379 U. S. 64, 74 (1964)).

falsity if used to demonstrate what occurred at the event. In contrast, it is unlikely to be reckless—or even negligent—if the blogger makes clear that the image is an illustration rather than a depiction of actual events.

In this way, the two different use examples of AI-generated content are very different. In the example of the protest, it is the context of the image that makes the image defamatory. Without that context, the instructor's face among a crowd of students is unlikely to create any questions of injurious content. In the example of generative AI nonconsensual pornography, the context within the image makes it inherently injurious. For photorealistic AI content that depicts images that are libelous irrespective of external context, then, the selection and publication on the internet will likely result in liability. Works that do not present themselves as factual depictions would not meet this test.[90]

This leaves an open question whether the company that created and trained the AI system can be liable.[91] When the AI system publisher provides the generated image to the individual submitting the text-based inquiry, it is publishing content to a third party. If that content otherwise meets all the other tests for defamation, it stands to reason that the publisher could be culpable for its content. In the case of pornography, for example, jurisdictions will likely hold AI system operators liable for production of child pornography if the AI generates simulated child pornography using depictions of actual children pulled from its training set.[92] Such

---

[90] *See, e.g.*, Scott Indrisek, *Can Art Legally Threaten the President?*, ARTSY (May 3, 2017), https://perma.cc/2995-Y35Z ("You might argue that exhibiting a nude Captain America flaunting the President's disembodied cranium is flirting with 'risk to others.' But Chung's painting, he says, is more nuanced than it might seem.").

[91] *See* Brown, *supra* note 20, at 400 ("Chatbots cannot act carelessly or recklessly. They likely cannot 'know' information is false. They are algorithms, and algorithms that behave by following a list of instructions. Considering this, it might seem prudent to ask whether the individuals responsible for programming the chatbot had the requisite mental state.").

[92] *Compare* Ashcroft v. Free Speech Coalition, 535 U.S. 234, 241 (2002) (finding unconstitutional a law banning "'virtual child pornography,' which include[s] computer-generated images, as well as images produced by more traditional means"), *with* United States v. Mecham, 950 F.3d 257, 260 (5th Cir. 2020) ("morphed child pornography does not enjoy First Amendment protection"); Doe v. Boland, 698 F.3d 877 (6th Cir. 2012) (holding that morphed images of actual children on images involving sex acts gave rise to liability for federal torts).

"morphed" child pornography is outside the protection of the First Amendment.[93] The circuit courts have followed dicta from the Supreme Court[94] and held that there is no requirement that the creator or other purveyors of the child pornography have any *mens rea* as to age of the victim.[95]

If a jurisdiction can reach morphed child pornography, it also stands to reason that it could reach other content created that runs afoul of the actual malice standard, such as photorealistic AI-generated pornography depicting real persons. Meeting this fact-based standard, however, will be extremely difficult, and the AI service will attempt to put the blame on the person offering the prompt rather than the response by the AI. Others will argue that the AI itself is the sole author of its output images, independent of the corporate ownership and the AI cannot be liable under any mens rea standard since "measuring mental state is impossible when there is no mind."[96]

---

[93] *See* People v. McKown, 2022 IL 127683, ¶ 36:

> *United States v. Mecham*, 950 F.3d 257, 260 (5th Cir. 2020) (agreeing "with the majority view that morphed child pornography does not enjoy first amendment protection"); *Doe v. Boland*, 698 F.3d 877, 883 (6th Cir. 2012) (rejecting defendant's first amendment challenge to morphed child pornography where "Jane Doe and Jane Roe are real children" whose "likenesses are identifiable in [defendant's] images"); *United States v. Hotaling*, 634 F.3d 725, 730 (2d Cir. 2011) ("Sexually explicit images that use the faces of actual minors are not protected expressive speech under the First Amendment."); *McFadden v. State*, 67 So. 3d 169, 184 (Ala. Crim. App. 2010) (ruling that statutes "which criminalize the possession *** of collage or montage images of child pornography *** created without *** photographing actual sexual conduct on the part of an identifiable minor, but edited to appear as though the children are engaged in sexual conduct, do not violate the First Amendment"); *Tooley*, 2007-Ohio-3698, ¶ 24, 872 N.E.2d 894 (declining to extend Free Speech Coalition "to cover morphed child pornography when the United States Supreme Court did not do so").

[94] *See* United States v. X-Citement Video, Inc., 513 U.S. 64, 77 n.6 (1994) ("The legislative history of § 2251(c) does address the scienter requirement: 'The government must prove that the defendant knew the character of the visual depictions as depicting a minor engaging in sexually explicit conduct, *but need not prove that the defendant actually knew the person depicted was in fact under 18 years of age* or that the depictions violated Federal law.'").

[95] United States v. Tyson, 947 F.3d 139, 146–47 (3d Cir. 2020) ("In the wake of the *X-Citement Video* decision, all of the federal courts of appeals that have considered the issue of scienter under § 2251(a) have held that a defendant's knowledge of the minor's age is not an element of the offense.").

[96] Derek E. Bambauer & Mihai Surdeanu, *Authorbots*, 3 J. FREE SPEECH L. 375, 382 (2023).

Given the significant difficulty in achieving ultimate success in libel actions, attempts to hold AI companies liable will likely be difficult. Unless a system's designers trained it specifically to generate child pornography, revenge pornography, or some other category of unprotected speech, prosecutors and civil litigants will find it very hard to hold system designers liable for the content generated by those systems.

III. THE ALTERNATIVE TO LIBEL LITIGATION: TAKEDOWN REGIMES FOR LIBELOUS CONTENT

While libel actions may provide occasional relief, the common law has simply not kept pace with the modern development of social media, mass media, the metaverse, and generative AI.[97] Not only have the efforts of the common law failed to adequately evolve, neither has federal statutory law. With the proliferation of synthetic content, the time has finally come for §230 to be modified to incorporate a notice and takedown system.[98] The synthetic media market bears little relation to

---

[97] *See* David A. Logan, *Rescuing Our Democracy by Rethinking* New York Times v. Sullivan, 81 OHIO ST. L. J. 759, 808 (2020) ("the most current data suggest a very different landscape, one in which the pendulum has swung so far toward defendants that defamation law gives little redress to the victims of falsehoods and provides virtually no deterrence of falsehoods"); Eugene Volokh, *Anti-Libel Injunctions*, 168 U. PENN. L. REV. 73, 76 (2019) ("The Internet lets speakers publish libels to a potentially broad audience at little cost, and these libels can cause enduring damage. . . . Moreover, [§230] generally immunizes intermediaries . . . . In any practical sense, damages awards do not leave plaintiffs in such cases with an 'adequate remedy at law.'").

[98] *See*, *e.g.*, Michael L. Rustad & Thomas H. Koenig, *Creating a Public Health Disinformation Exception to CDA Section 230*, 71 SYRACUSE L. REV. 1251, 1252–53 (2021) ("Congress should amend . . . Section 230 to recognize a notice-and-takedown exception for deceitful scientific and health information causing specific harm. Arming the direct victims of false health information with a takedown procedure and money damages will allow them to receive compensation for specific harm they suffer from dangerous content."); Gregory M. Dickinson, *Rebooting Internet Immunity*, 89 GEO. WASH. L. REV. 347, 351 (2021) ("Section 230 started off humbly, and suffers from a humble problem: The Congress of 1996 did not foresee the internet of 2020, and the statute is now outdated."); Jon M. Garon, *Constitutional Limits on Administrative Agencies in Cyberspace*, 8 BELMONT L. REV. 499 (2021); Amanda Bennis, *Realism About Remedies and the Need for a CDA Takedown: A Comparative Analysis of § 230 of the CDA and the U.K. Defamation Act 2013*, 27 FLA. J. INT'L L. 297 (2015). Of course, there are equally strong voices in favor of Section 230 immunity. *See, e.g.*, JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET (2019); Eric Goldman, *Why Section 230 Is Better Than the First Amendment*, 95 NOTRE DAME L. REV. REFLECTION 33, 34 (2019) ("The First Amendment and Section 230 are not substitutes for each other. . . . Because the First

the mass media of the 1960s and the power of the political elite to harness that influence. Nor do the enterprises trying to harness the power of artificial intelligence bear much resemblance to the start-ups that were struggling to develop in 1996 when §230 was enacted.

The public official conceived by Justice Brennan in *New York Times v. Sullivan* was a powerful institutional leader who weaponized the law to punish political opponents and stifle dissent. The Court began by holding that Alabama's libel law was "constitutionally deficient for failure to provide the safeguards for freedom of speech and of the press that are required by the First and Fourteenth Amendments in a libel action brought by a public official against critics of his official conduct."[99] It then emphasized the importance of political dissent:

> It is a prized American privilege to speak one's mind, although not always with perfect good taste, on all public institutions, and this opportunity is to be afforded for vigorous advocacy no less than abstract discussion. The First Amendment, said Judge Learned Hand, presupposes that right conclusions are more likely to be gathered out of a multitude of tongues, than through any kind of authoritative selection. To many this is, and always will be, folly; but we have staked upon it our all.[100]

The Court concluded by emphasizing the importance of robust, tumultuous criticism of government and public officials by explaining that any form of

---

Amendment does not backfill these benefits, reducing Section 230's immunity poses major risks to online free speech and the associated benefits to society."); Christian Sarceño Robles, *Section 230 Is Not Broken: Why Most Proposed Section 230 Reforms Will Do More Harm Than Good, and How the Ninth Circuit Got It Right*, 16 FIU L. REV. 213, 225 (2021) ("Unlike targeted proposals, an across-the-board reasonableness requirement would have far-reaching consequences. It would immediately deprive businesses of Section 230's biggest procedural benefits, as more cases would have to go to trial for factual determinations rather than be dismissed, increasing the costs of the litigation.").

[99] *New York Times Co. v. Sullivan*, 376 U.S. at 264.

[100] *Id.* at 269–70 (internal quotations omitted). The opinion continued with a quote from Justice Brandeis' famous concurrence in *Whitney v. California*, 274 U.S. 357, 375–76 (1927):

> Those who won our independence believed . . . that public discussion is a political duty; and that this should be a fundamental principle of the American government. They recognized the risks to which all human institutions are subject. But they knew that order cannot be secured merely through fear of punishment for its infraction; that it is hazardous to discourage thought, hope and imagination; that fear breeds repression; that repression breeds hate; that hate menaces stable government; that the path of safety lies in the opportunity to discuss freely supposed grievances and proposed remedies; and that the fitting remedy for evil counsels is good ones.

seditious libel "has disquieting implications for criticism of governmental conduct. For good reason, 'no court of last resort in this country has ever held, or even suggested, that prosecutions for libel on government have any place in the American system of jurisprudence.'"[101]

Nothing suggests these propositions have changed. What has changed, however, is the expansion of publication out of the hands of mass media and into the hands of a seemingly infinite array of personal machines, devices, websites, and other purveyors.[102] Shortly after Congress passed Section 230 to provide immunity for tort liability it also passed Section 512 of the Digital Millennium Copyright Act[103] which provided a mechanism for copyright holders to protect their copyrighted works from online infringement. While the web hosts were again provided immunity for civil liability, §512 provided a safe harbor only if infringing works were removed expeditiously.[104] Congress recognized that copyright interests were

---

[101] *New York Times Co. v. Sullivan*, 376 U.S. at 291–92.

[102] *See* Logan, *supra* note 97, at 795 ("[In 1964], there was relative economic stability in the media marketplace. Newspapers were often owned by wealthy families with deep ties to their home communities, so they could be relatively impervious to pressures to return a robust profit for shareholders.").

[103] Digital Millennium Copyright Act, Pub. L. 105-304, Title II, § 202(a), 112 Stat. 2877.

[104] 17 U.S.C.A. § 512(c):

> Information residing on systems or networks at direction of users.—
>
> (1) In general.—A service provider shall not be liable for monetary relief, or, except as provided in subsection (j), for injunctive or other equitable relief, for infringement of copyright by reason of the storage at the direction of a user of material that resides on a system or network controlled or operated by or for the service provider, if the service provider—
>
> (A)(i) does not have actual knowledge that the material or an activity using the material on the system or network is infringing;
>
> (ii) in the absence of such actual knowledge, is not aware of facts or circumstances from which infringing activity is apparent; or
>
> (iii) upon obtaining such knowledge or awareness, acts expeditiously to remove, or disable access to, the material;
>
> (B) does not receive a financial benefit directly attributable to the infringing activity, in a case in which the service provider has the right and ability to control such activity; and
>
> (C) upon notification of claimed infringement as described in paragraph (3), responds expeditiously to remove, or disable access to, the material that is claimed to be infringing or to be the subject of infringing activity.

of a different nature than generalized tort claims for defamation and provided a different scheme for content removal.[105]

While there has been widespread criticism of the § 512 takedown regime,[106] because "an estimated 500 hours of video are uploaded to YouTube per minute, service providers simply could not exist in their current form without Section 230 of the CDA and Section 512 of the DMCA to protect them from liability arising from such content."[107] The sheer volume of content processing through the system strongly indicates that for all its limitations § 512 provides a workable compromise to reduce infringement. A somewhat improved version of the takedown regime that incorporates the recent suggestions of the Copyright Office[108] could go even further to provide relief for the victims on nonconsensual pornography and other libelous content without the need for protracted litigation that fails to remedy the underlying harm.

A minimal additional step in the efforts to avoid the harms caused by photorealistic but false depictions of content might be to extend §512 beyond copyright to rights of privacy and defamation when a person is depicted in a work and that person objects to the representation. As with §512, the party posting the content can claim that the use is a proper use and get the content restored using a counter-

---

[105] *See* Neda Shaheen & Jacob Canter, *The CDA and DMCA—Recent Developments and How They Work Together to Regulate Online Services*, CROWELL (Feb. 6, 2023), https://perma.cc/RS99-J5QU ("While the CDA and DMCA are separate statutes, they work together to regulate online services.").

[106] *See* Wesley D. Lewis, *Trends in ISP and Platform Liability: CDA Section 230 and DMCA Safe Harbors*, HAYNES BOONE 1, 4 (Aug. 18, 2020), https://perma.cc/EY7V-WG6Q ("The DMCA's safe-harbor provisions also face mounting criticism from the United States Copyright Office itself."); U.S. Copyright Office, *Section 512 of Title 17: A Report of the Register of Copyrights*, at 1 (May 2020) ("Since its establishment in 1998, as part of the Digital Millennium Copyright Act ('DMCA'), section 512 of title 17 has both provided critical guideposts for the expansion of the internet and produced widespread disagreement over its operation.").

[107] Lewis, *supra* note 106, at 1 (citing *Hours of Video Uploaded to YouTube Every Minute as of May 2019*, STATISTA (Aug. 9, 2019)).

[108] *Section 512 of Title 17: A Report of the Register of Copyrights*, *supra* note 106, at 2–3. The areas of improvement relevant to defamation are the same as those for copyright: clarity on which operators are eligible for the safe harbor provisions, a need to improve obligations around repeat infringer policies, an increase for the duties of operators to remove offending content based on the knowledge standard, and a need to simplify the information needed to trigger the takedown, given the changes in the nature of the ISPs operating systems.

notification.[109] Whether such a model is appropriate for all generative AI content is beyond the scope of this discussion of generative AI imagery.

The use of the §512 system provides a certain amount of friction to the internet's content distribution, but that friction has not proven to negatively impact the copyright industries or unduly limit fair use. More generally, there may be a high degree of social utility in at least some marginal resistance to the immediacy of broadcasting every image that pops out of an AI engine. A system that provides for both take-down and put-back of content at least puts the purveyor of the content on notice that there may be consequences for publishing harmful content.

A takedown regime is likely required under various state privacy laws. As noted earlier, factually accurate nonconsensual pornography is generally a violation of the victim's privacy rights and the falsity of the information raises at least a false light claim as to that privacy as well. As a practical matter, there should be no reason that false harmful information is given more legal protection than truthful information that invades one's privacy. New state laws such as the California Consumer Privacy Act of 2018 (CCPA) as amended in 2020 by the California Privacy Rights Act (CPRA),[110] require a business to grant a person "the right to correct inaccurate

---

[109] *See* 17 U.S.C.A. § 512(g)(2):

(2) Exception.—Paragraph (1) shall not apply with respect to material residing at the direction of a subscriber of the service provider on a system or network controlled or operated by or for the service provider that is removed, or to which access is disabled by the service provider, pursuant to a notice provided under subsection (c)(1)(C), unless the service provider—

(A) takes reasonable steps promptly to notify the subscriber that it has removed or disabled access to the material;

(B) upon receipt of a counter notification described in paragraph (3), promptly provides the person who provided the notification under subsection (c)(1)(C) with a copy of the counter notification, and informs that person that it will replace the removed material or cease disabling access to it in 10 business days; and

(C) replaces the removed material and ceases disabling access to it not less than 10, nor more than 14, business days following receipt of the counter notice, unless its designated agent first receives notice from the person who submitted the notification under subsection (c)(1)(C) that such person has filed an action seeking a court order to restrain the subscriber from engaging in infringing activity relating to the material on the service provider's system or network.

[110] CAL. CIV. CODE §§ 1798.105(a), (c)(3), 1798.130(a)(3)(A).

personal information [and] the right to limit use and disclosure of sensitive personal information."[111] Other states have and will likely continue to follow this effort. A notice-and-takedown system will inevitably be required to implement the CPPA as members of the public look for mechanisms to inform the businesses of their demand to remove offending inaccurate or intrusive content. While the broader discussion of the interplay between defamation and privacy under state law is beyond this article, the need for compliance with state law will likely continue to drive evolution of systems to improve content management by business.[112]

In addition to the obligation on the individual internet service and regulated business enterprise, there is a broader need to create a national clearinghouse mechanism to assure that unauthorized content is identified, digitally fingerprinted, and indexed to help stop the content from being immediately reposted. There needs to be a single point of reporting to do so. A successful tort plaintiff should be able to rely on an indexing service for this purpose. It will need to do more than merely block perfect reproductions of the identified images and instead utilize facial recognition software and other AI-enabled tools to keep substantially similar variation from being posted as well. This should be technologically feasible since this is precisely what GAN image training does best.[113]

Such a system could be modeled on the Child Victim Identification Program (CVIP) of the National Center for Missing and Exploited Children.[114] CVIP ensures

---

[111] *CCPA vs CPRA: What's the Difference?*, BLOOMBERG LAW (Jan. 23, 2023).

[112] *See* Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. FREE SPEECH L. 489, 515 (2023) (noting the possibility that various matter could be blocked using "'post-processing' content filtering code, where the output of the underlying Large Language Model algorithm would be checked, and certain material deleted").

[113] Image hashing is just one of many sources for image identification. Others include metadata and personal identifiers linking each image to the camera used to take it. *See* Jerone Andrews, *The Hidden Fingerprint Inside Your Photos*, BBC (Mar. 24, 2021); Anindya Sarkar, Pratim Ghosh, Emily Moxley & B. S. Manjunath, *Video Fingerprinting: Features for Duplicate and Similar Video Detection and Query-Based Video Retrieval*, PROCEEDINGS OF SPIE - THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING (Jan. 2020), https://perma.cc/552Q-VM62. *See also* Erdogan Taskesen, *Detection of Duplicate Images Using Image Hash Functions,* TOWARDS DATA SCI. (Jan 28, 2022), https://perma.cc/7NC8-M52C.

[114] *See* Adrienne L. Fernandes-Alcantara & Emily J. Hanson, *The Missing and Exploited Children's (MEC) Program: Background and Policies*, RL34050, CONG. RSCH. SERV. at 11 (July 15, 2021)

that previously identified child pornography is kept from being reposted.[115] Content that falls into the category of being adjudicated as unpublishable by a court proceeding following appropriate First Amendment requirements should not still be on the internet simply because there is no system to track and remove the content. The very AI systems that are generating new content can also be used to ensure that content adjudicated as unpublishable can be kept off host systems.

Beyond the removal of adjudicated libelous content, a hash system of digital fingerprinting could also be used to identify and remove other offending content that has been taken down, provided it was not reposted through the statutory counter-notice or "put-back notice" provisions.[116] To avoid the challenge of policing the seemingly infinite internet, it is time for the multinational internet service providers to come together and provide a voluntary system to identify, digitally fingerprint, and automatically remove works that have been identified by victims of nonconsensual pornography from the repeated republication of identified images.

("The Child Victim Identification Program (CVIP) began in 2002 . . . CVIP analysts use computer software and visual analysis to determine whether any of the images contain identified child victims. Additionally, CVIP provides training and educational assistance to law enforcement and attorneys . . . ."); *Our Impact*, NATIONAL CENTER FOR MISSING & EXPLOITED CHILDREN, https://perma.cc/L5WM-RZLC (CVIP "serves as the nation's clearinghouse on identified child victims of CSAM. Files containing unidentified children are reviewed and analyzed for any information as to their potential location or who is responsible for their abuse. When this information can be determined, CVIP provides the analysis to the appropriate law enforcement. . . .").

[115] *CyberTipline 2021 Report*, NATIONAL CENTER FOR MISSING & EXPLOITED CHILDREN, https://perma.cc/7GDQ-WKVX ("Hash values are unique digital fingerprints assigned to pieces of data like images and videos. When an image or video is identified as containing known [Child sexual abuse material (CSAM)], NCMEC adds the hash value to a list that is shared with technology companies."). *See also* Nicola Henry & Alice Witt, *Governing Image-Based Sexual Abuse: Digital Platform Policies, Tools, and Practices*, in THE EMERALD INTERNATIONAL HANDBOOK OF TECHNOLOGY-FACILITATED VIOLENCE AND ABUSE 758 (2021), https://perma.cc/6U88-KSJJ:

> In 2009, Microsoft and Professor Hany Farid from Dartmouth College developed PhotoDNA, a technology that creates a unique digital signature (also known as a "hash" or "digital fingerprint") of child sexual abuse images, which can then be compared against known images stored in a database curated by the National Center for Missing and Exploited Children in the United States.

[116] *See* 17 U.S.C. § 512(g) (providing for a counter-notice. In the event of a counter-notice, the ISP may only maintain safe harbor protection if it reposts the content in a timely manner or the party posting the original notice files suit against the party that posted the offending content.).

Such systems already exist. "[C]ompanies have adopted similar methods. Pornhub uses a third-party automated audio-visual identification system (called MediaWise®) which first identifies the content using 'digital fingerprinting,' and then blocks it from being uploaded again in the future."[117] Although these systems can be fooled,[118] they provide an important start. By centralizing the repository, a third-party operated system reduces the problem that a victim must give even more personal information to the very site that is hosting the harmful content.

The possibility now exists that AI systems can produce an essentially infinite amount of pornographic content. Steps should be taken to ensure that the versions of this content depicting real individuals without their authorization are kept off the internet. There are many First Amendment concerns regarding such a system being enforced through criminal laws. But there are strong business reasons for large host enterprises to avoid the potential liability that content could engender. As a result, such a solution should be an industry imperative.

Conclusion

Generative AI may not have been invented to invade privacy or commit libel, but there is no doubt that one of the consequences of the technology will be to do so, and to create the potential that harmful content can be mass-produced at an alarming scale. By putting safeguards into place, there will be less incentive to create such harmful content. More importantly, the public will benefit by having simpler tools available to overcome the harms caused by libelous photorealistic synthetic content.

Under the common law, libel served to protect the public from harmful speech. But the transformation of civil discourse in the 1960s led to a vastly different model of libel than that which existed at common law. The First Amendment continues to be an essential protection for public comment and for criticism of public officials, and a bedrock of the United States' ethos as a free society. But outside of matters of public concern, the procedural hurdles facing a victim of defamation make the legal system a difficult and unappealing solution to the private harm caused by unwanted and false publications.

---

[117] Henry & Witt, *supra* note 115, at 758.

[118] *See id.* ("These technological solutions have received significant criticism. The Facebook pilot was widely condemned for asking vulnerable individuals to trust Facebook . . . [A] Motherboard investigation found that Pornhub's fingerprinting system 'can be easily and quickly circumvented with minor editing[.]'").

Rather than fix everything that might be wrong with the law of libel and risk undermining all that is right with the law, society is much better served by simplifying the way in which a victim of a falsehood can have that harmful content expunged from the internet. This will not work in matters of public interest or for controversial topics. That should not, however, end such efforts. Far too many examples involve libelous publications of images and videos that falsely depict their contents as fact, that involve private matters and private individuals. Falsity has little value as speech; when it imposes clear harms on private individuals outside the scope of public concern, more can and should be done.

At its heart, this is a modest proposal. The First Amendment is too valuable to suggest more sweeping changes. But The First Amendment is also too important to leave alone jurisprudence from the 1960s and 1970s that bears little or no relationship to the society in which we find ourselves. Just as libel law and social institutions evolved in prior eras, the introduction of generative AI will foster a new era of social and institutional change. The only question, then, is whether that change is for the good. Hopefully, these modest suggestions will nudge us in that direction. Time will tell.