# A Real Account of Deep Fakes

Benjamin L.W. Sobel*

*Laws regulating pornographic deepfakes are usually characterized as protecting privacy or preventing defamation. But privacy and defamation laws paradigmatically regulate true or false assertions of fact about persons. Anti-deepfakes laws do not: the typical law bans even media that no reasonable observer would understand as factual. Instead of regulating statements of fact, anti-deepfakes laws ban outrageous depictions as such. This is an unrecognized departure from established privacy and defamation law, and it carries serious constitutional stakes. Because anti-deepfakes laws ban outrageous imagery irrespective of any factual assertions it makes, the rationales for defamation and almost all privacy doctrines do not justify these statutes under the First Amendment. Properly understood, anti-deepfakes laws fall into a distinct, and constitutionally disfavored, category: laws that forbid expression* because it is offensive.

*This Article conducts the first scholarly survey of every anti-deepfakes statute and distills the typical law. It then uses semiotic theory to explain how deepfakes differ from the media they mimic and why those differences matter legally. Photographs and video recordings* record *events. Deepfakes merely* depict *them. Justifications for regulating records do not necessarily justify regulating depictions. Many laws—covering everything from trademark dilution to flag burning to "morphed" child sexual abuse material (CSAM)—have banned offensive depictions as such. Several remain in effect today. Yet when such bans are challenged, courts mischaracterize imagery to sidestep constitutional scrutiny: courts pretend fictional depictions are factual records. Anti-deepfakes laws resist this dodge. The laws will force courts to confront squarely whether a statute may ban offensive expression as such—and thus to determine whether a history and tradition of such regulation can be reconciled with First Amendment jurisprudence that would seem to forbid it.*

---

* Postdoctoral Fellow, Digital Life Initiative, Cornell Tech; Program Affiliate Scholar, Information Law Institute, New York University. Thanks to Amy Adler, Barton Beebe, Dinis Cheian, Bryan Choi, Daniel Francis, Mary Anne Franks, Ruobin Gong, James Grimmelmann, Gautam Hans, Thomas Kadri, Mark Lemley, Amanda Levendowski, Jonathan Masur, Helen Nissenbaum, Ela Leshem, Joseph Miller, Frank Pasquale, Lisa Ramsey, Blake Reid, Ted Sichelman, Joel Sobel, Katherine Strandburg, Rebecca Tushnet, Rebecca Wexler, Kathryn Woolard, Mark Wu, the Fellows and Affiliates of the Cornell DLI and NYU ILI, and the participants in the Cornell Law School Academic Professionals Workshop, the 2024 Legal Scholars Roundtable on Artificial Intelligence at Emory Law School, the 2024 NYU Academic Careers Program Scholarship Clinic, the 2024 Richmond Junior Faculty Forum, the 2024 Privacy Law Scholars Conference, and the 2024 Intellectual Property Scholars Conference. This paper received the Ian Kerr Award for Best Paper at the 2024 Privacy Law Scholars Conference.

**Table of Contents**

## Introduction

*"[R]epresentation* is *reality . . . ."*[1]

Breakthrough technology has made it cheap and easy to synthesize photorealistic images and videos of recognizable individuals. Overwhelmingly, people are using it to generate porn.[2] An AI user needs only a few photographs of his target to generate a pornographic "deepfake" of anyone from an ex-girlfriend to a celebrity.[3] In early 2024, sexually explicit deepfakes of Taylor Swift gathered tens

1. CATHARINE A. MACKINNON, ONLY WORDS 29 (1993) (discussing, with approval, an argument attributed to Susanne Kappeler).

2. HENRY AJDER ET AL., *The State of Deepfakes: Landscape, Threats, and Impact*, 1 (2019), https://regmedia.co.uk/2019/10/08/deepfake_report.pdf (last visited Jan 29, 2024). In keeping with colloquial usage, I use "porn," "pornography," and "pornographic" to denote depictions of nudity or sexual conduct, but it bears noting that jurists have observed that this vocabulary arguably mischaracterizes nonconsensual, sexualized depictions. *See* Mary Anne Franks, *"Revenge Porn" Reform: A View from the Front Lines*, 69 FLA. L. REV. 1251, 1257-58 (2017), *cited in* People v. Austin, 155 N.E.3d 439, 451 (Ill. 2019).

3. To match what appears to be the typical scenario, I use "he/him" pronouns to refer in general to a creator or distributor of nonconsensual, deepfake pornography, and "she/her" pronouns to describe the person depicted. *Id.* at 2; Jess Weatherbed, *Trolls Have Flooded X with Graphic Taylor Swift AI Fakes*, THE VERGE (2024),

of millions of views on the social media site X.[4] Women in politics and journalism are being threatened with deepfake pornography that uses their likenesses.[5] Schoolchildren across the nation are using AI to synthesize naked images of their classmates, and two boys in their early teens have been charged with felonies for allegedly doing so.[6] The White House is alarmed.[7] Every state attorney general is alarmed.[8] Legislators are alarmed.[9] The nation is alarmed.

The alarm is justified. But it is *unjustified* by the legal concepts that are usually invoked, without examination, to explain it. Equally unexamined in any systematic way is the substance of the laws that 26 states have enacted specifically to address nonconsensual, pornographic deepfakes.[10] Congress has considered at least six similar bills, two of which passed the Senate in the months before the 119th Congress was seated in January 2025.[11] This article is the first to survey

---

https://www.theverge.com/2024/1/25/24050334/x-twitter-taylor-swift-ai-fake-images-trending (last visited Jan 29, 2024).

4. Weatherbed, *supra* note 3.

5. Mark Scott, *Deepfake Porn Is Political Violence*, POLITICO, Feb. 8, 2024, https://www.politico.eu/newsletter/digital-bridge/deepfake-porn-is-political-violence/ (last visited Mar 18, 2024); Danielle Keats Citron, *Sexual Privacy*, 128 YALE L. J. 1870, 1922–23 (2019).

6. Caroline Haskins, *Florida Middle Schoolers Arrested for Allegedly Creating Deepfake Nudes of Classmates*, WIRED, https://www.wired.com/story/florida-teens-arrested-deepfake-nudes-classmates/ (last visited Mar 13, 2024). *See also* Jason Koebler & Emanuel Maiberg, *A High School Deepfake Nightmare*, 404 MEDIA, Feb. 15, 2024, https://www.404media.co/email/547fa08a-a486-4590-8bf5-1a038bc1c5a1/.

7. Justin Sink, *White House Urges Action After 'Alarming' Taylor Swift Deepfakes*, BLOOMBERG.COM, Jan. 26, 2024, https://www.bloomberg.com/news/articles/2024-01-26/white-house-urges-action-after-alarming-taylor-swift-deepfakes (last visited Feb 19, 2024).

8. Meg Kinnard, *Prosecutors in All 50 States Urge Congress to Strengthen Tools to Fight AI Child Sexual Abuse Images*, AP NEWS, Sep. 5, 2023, https://apnews.com/article/ai-child-pornography-attorneys-general-bc7f9384d469b061d603d6ba9748f38a (last visited Jan 7, 2024).

9. *See* Part I.A, *infra*.

10. These numbers are accurate as of January 2025. I exclude statutes that regulate only AI-generated CSAM, and I also exclude Tennessee's comprehensive digital replica law, which does not focus on sexual harms. *See* TENN. CODE ANN. § 47-25-1101 *et seq.*

11. *See* Disrupt Explicit Forged Images And Non-Consensual Edits Act of 2024, S. 3696, 118TH CONG., 2D. SESS. (engrossed in Senate Jul. 23, 2024), *available at* https://www.congress.gov/bill/118th-congress/senate-bill/3696/text [hereinafter DEFIANCE Act of 2024]; Nurture Originals, Foster Art, and Keep Entertainment Safe Act of 2024 (introduced in Senate Jul. 31, 2024), *available at*

comprehensively these laws' substantive provisions, as well as the first to analyze thoroughly the legal implications of two glaring details that are almost never duly acknowledged: deepfakes are not photographs or video recordings; and often, they don't even pretend that they are.

Anti-deepfakes statutes differ from conventional privacy and defamation laws in a crucial respect. Privacy and defamation law regulate facts, or assertions of fact, about persons. Anti-deepfakes laws ban outrageous depictions of persons, irrespective of any factual assertions they make. This difference matters for two reasons. First, legislators who approach deepfakes as a defamation, fraud, or forgery problem or as a conventional privacy problem—that is, as a problem of false representations of fact or true disclosures of private facts—end up enacting statutes that may not redress the harms they should redress.

Second, because anti-deepfakes statutes ban certain outrageous imagery *per se*, irrespective of any factual assertions it makes, the rationales that justify defamation and almost all invasion-of-privacy regimes do not justify the constitutionality of these laws. Courts applying the First Amendment regard bans on offensive expressions of opinion with a distinct skepticism not directed at regulations of factual statements.[12] This isn't to say that courts never uphold the

---

https://www.congress.gov/bill/118th-congress/senate-bill/4875/text?s=3&r=2 [hereinafter NO FAKES Act of 2024]; No Artificial Intelligence Fake Replicas And Unauthorized Duplications Act of 2024, H.R. 6943, 118TH CONG., 2D. SESS. (introduced in House Jan. 10, 2024), *available at* https://www.congress.gov/bill/118th-congress/house-bill/6943/ [hereinafter No AI FRAUD Act]; Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act, S. __, 118TH CONG., 2D. SESS. (introduced in Senate June 18, 2024), *available at* https://www.commerce.senate.gov/services/files/747FB142-F25D-4BB2-83D3-50FD5D44553E [hereinafter TAKE IT DOWN Act First Draft]; Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act, S. 4569, 118TH CONG., 2D. SESS. (engrossed in Senate Dec. 3, 2024), *available at* https:// https://www.congress.gov/bill/118th-congress/senate-bill/4569/text [hereinafter TAKE IT DOWN Act Second Draft]; Preventing Deepfakes of Intimate Images Act, H.R. 3106, 118TH CONG., 1ST SESS. (introduced in House May 5, 2023); and the Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2023, H.R. 5586, 118TH CONG., 1ST SESS. (introduced in House Sept. 20, 2023) [hereinafter DEEPFAKES Accountability Act].

12. *See, e.g.*, Matal v. Tam, 582 U.S. 582 U.S. 218, 223 (2017) (stating "a bedrock First Amendment principle: Speech may not be banned on the ground that it expresses ideas that offend."); Snyder v. Phelps, 562 U.S. 443, 458 (2011) ("in public debate we must tolerate insulting, and even outrageous, speech in order to provide adequate breathing space to the freedoms protected by the First Amendment" (quoting Boos v. Barry, 485 U.S. 312, 322 (1988)) (quotation marks and alteration marks omitted)).

constitutionality of *per se* bans on outrageous expression. They do. Along with obscenity law, the most notable example is the federal ban on "morphed" child sexual abuse material (CSAM), a lower-tech predecessor to deepfakes in which non-sexual images of identifiable children are edited to depict sexual conduct.[13] But in upholding such bans, courts do not always admit that they forbid outrageous expression *per se*. Rather, courts sometimes pretend that they forbid expression because it signifies particular historical facts. In other words, when evaluating outrageous imagery, courts pretend to be regulating *records of historical fact* when they are in fact regulating mere *depictions of fictional events*. In the language of semiotics, courts pretend to be analyzing indexical images when they are in fact analyzing iconic images.

Properly drafted and properly understood, anti-deepfakes statutes are content-based restrictions on noncommercial speech that is not necessarily obscene, not necessarily defamatory, and discloses no private facts. They are the most recent manifestation of a longstanding impulse to ban the offensive treatment of icons. This impulse motivates not only historical regulations of flag and effigy burning, but also bans on morphed CSAM and trademark dilution by tarnishment that remain in force today. Anti-deepfakes laws are also the most recent manifestation of the law's impulse to mischaracterize *per se* bans on offensive expression as something they are not. Courts relied on the conflation of indexical and iconic imagery to uphold bans on morphed CSAM, and scholars and legislators are reenacting this maneuver today for anti-deepfakes laws.

But anti-deepfakes laws offer none of the analytical offramps that other bans on offensive expression offer. Courts considering morphed CSAM could expediently classify it as "child pornography," which is categorically unprotected by the First Amendment, even though the rationales for that categorical exclusion apply only to *records* of abuse and not mere *depictions* of abuse. This approach is unlikely to work as well for deepfakes, because there is no categorical First Amendment exclusion for pornographic depictions of adults. Nor can deepfakes be forbidden on the ground that they disclose true, private facts, as revenge pornography does, or on the ground that they necessarily make false statements of fact. What anti-deepfakes laws will do, then, is force courts to consider the degree to which American law can ban outrageous iconography as such—something it has reliably done and continues to do today—*even when jurists have to admit that this is what the law is doing*. Anti-deepfakes laws' departure from the theories underlying the established dignitary torts invites us to reconsider how hostile our constitutional

13. *See* Part III.A.2.c), *infra*.

order ought to be to the regulation of offensive expression *per se*—and, indeed, how hostile it ever really has been.

Part I surveys all 26 enacted anti-deepfakes laws to distill the typical law's essential characteristics. It explains that anti-deepfakes laws differ from defamation law, and all but one recognized privacy doctrine, because they regulate not statements of fact but outrageous expression.

Part II uses semiotic theory to provide a rigorous account of how deepfakes differ from photographs and video recordings. Photographs are *indexical*: they record a visual phenomenon as it appeared through a particular lens at a particular moment in time. Deepfakes are *iconic*: they represent by resemblance. We interpret indexical media as assertions of fact; as a result, accurate photographs can reveal private matters and deceptive photographs can defame. But deepfakes are merely icons. They do not necessarily assert facts in the way that photographs do. As a result, the legal rationales historically invoked to regulate indexical imagery cannot support the full breadth of today's anti-deepfakes laws.

Part III tours trademark dilution and CSAM law, past prohibitions on flag desecration and effigy burning, and the criminalization (*vel non*) of sexual fantasy to show that anti-deepfakes laws address a well-understood harm and have close cousins in past and present American legal doctrines. Part III also, however, shows that courts mischaracterize the semiotic status of this harm in order to sidestep First Amendment scrutiny: to uphold the regulation of icons as constitutional, courts invoke rationales that instead justify the regulation of indices.

Part IV, finally, explains that properly understanding the semiotics of deepfakes is essential to appropriate and effective regulation. Congress and the states are rushing to enact laws that address a coming deluge of photorealistic, AI-generated pornography. Our impulse to outlaw outrageous uses of icons *per se* will collide with our tendency to deny that we are doing so. We may try to pretend that deepfakes are equivalent to photographs, but commandeering the law of photographs and video recordings to regulate AI-generated imagery will produce bizarre outcomes—like classifying any sexually explicit image generated by industry-standard AI as "child pornography" under federal law, even if the generated imagery depicts only adults.[14] Properly regulating deepfakes requires acknowledging that they are icons, not indices, and employing the legal theories that regulate them as such: obscenity law and an extended version of the tort of appropriation.

14. *See* Part IV.A, *infra*.

## I. What the Typical Anti-Deepfakes Law Does

26 states have enacted civil or criminal legislation dealing specifically with nonconsensual, pornographic deepfakes depicting adults.[15] Although the issue of pornographic deepfakes has, quite justifiably, commanded a great deal of attention, the actual substance of these laws has received almost no systematic analysis. This Part isolates seven characteristics that typify an anti-deepfakes law. Deepfakes are usually condemned as invasions of privacy akin to so-called "revenge porn," or as injurious falsehoods.[16] These are harms that established privacy and defamation doctrines, respectively, can explain and redress. But Part I shows that the typical anti-deepfakes law extends well beyond the limits of the dignitary torts, because it encompasses even noncommercial deepfakes that are obviously ahistorical.

### A. The Requirements of the Prototypical Law

#### 1. Sexual Content

The laws I analyze almost always limit their coverage to nonconsensual deepfakes that depict sexual conduct and/or nudity.[17] (Indeed, pornographic deepfakes also appear to represent a great proportion of deepfakes published online.)[18] If I use the word "deepfake" without another modifier, this is the sort of deepfake I am referring to.

#### 2. Identifiability

The typical law also expressly requires that the individual depicted be "identifiable" either from the image itself or from information presented in connection with it.[19] This requirement is easily satisfied in the paradigmatic deepfake, which depicts a victim's face. From here on, if I refer to a deepfake being

---

15. *See* Appendix: State Anti-Deepfakes Laws, *infra*.
16. Mary Anne Franks & Ari Waldman, *Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions*, 78 MD. L. REV. 892, 893–94 (2019); Rebecca A. Delfino, *Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act*, 88 FORDHAM L. REV. 887, 897–98 (2019).
17. *See, e.g.*, IDAHO CODE § 18-6606 (eff. July 1, 2024); MINN. STAT. § 604.32.2(2) ; N.Y. CIV. RIGHTS LAW § 52-c(2)(a) (McKinney); TEX. PENAL CODE § 21.165(b) . *But see* NO FAKES Act of 2024 (covering likeness generally).
18. A report from 2019 asserted that 96% of online deepfake videos were pornographic. AJDER ET AL., *supra* note 2 at 1.
19. *See, e.g.*, ALA. CODE § 13A-6-240(b); HAW. REV. STAT. § 711-1110.9(1)(c) ; 740 ILL. COMP. STAT. 190/10(a); MINN. STAT. § 604.32.2(2)-(3); N.Y. PENAL LAW § 245.15(1)(a) (McKinney).

"of" someone or "depicting" someone, I am referring to a deepfake that satisfies the identifiability requirement.

### 3. Dissemination

Overwhelmingly—but not universally—laws regulating nonconsensual pornographic deepfakes require an actual or offered dissemination or exhibition of a deepfake, rather than mere creation or possession. No civil anti-deepfakes law prohibits mere creation or possession of a deepfake. California's, Minnesota's, New York's statutes prohibit intentional disclosure of a pornographic deepfake.[20] Illinois's civil statute encompasses both "intentional dissemination" and "threatened dissemination."[21] The proposed federal AI NO FRAUD Act covers "mak[ing] available to the public."[22] The proposed federal DEFIANCE Act and NO FAKES Act, however, both encompass simple "production."[23] Several state criminal laws prohibit something less than disclosure of a nonconsensual pornographic deepfake.[24] The typical criminal anti-deepfakes law, however, makes dissemination, or at least an offer to disseminate, the criminal act.[25] Statutes may also narrowly exempt certain disseminations—such as in the course of "reporting unlawful activity" or "participating in a . . . legal proceeding"[26]—or establish broader exemptions for "material that constitutes a work of political, public interest, or newsworthy value."[27]

### 4. Intent and Scienter

All criminal anti-deepfakes laws require that the *actus reus* be committed knowingly or intentionally, and many require proof of some additional intent or

---

20. CAL. CIV. CODE § 1708.86(a)(7), (b) ; N.Y. CIV. RIGHTS LAW § 52-c(2)(a) (McKinney) ("A depicted individual shall have a cause of action against a person who, discloses, disseminates or publishes . . ."); MINN. STAT. § 604.32.2(a)(1) ("disseminat[ion]").
21. 740 ILL. COMP. STAT. 190/10(a).
22. No AI FRAUD Act § 3(c)(1)(B).
23. DEFIANCE Act of 2024 § 3(b)(1)(A); NO FAKES Act of 2024 § 2(c)(2)(A).
24. *See, e.g.*, TEX. PENAL CODE § 21.165(b) (creation); HAW. REV. STAT. § 711-1110.9 (creation or threatened disclosure); LA. STAT. § 14:73.13(A)-(B) (possession of deepfake depicting minor); MD. CODE, CRIM. LAW § 11-208 (same).
25. ALA. CODE § 13A-6-240(a); FLA. STAT. § 836.13(2) ; GA. CODE § 16-11-90(b) ; N.Y. PENAL LAW § 245.15(1)(a) (McKinney); S.D. CODIFIED LAWS § 22-21-4(3); UTAH CODE § 76-5b-205 ); VA. CODE § 18.2-386.2(A) .
26. FLA. STAT. § 836.13(6) .
27. LA. STAT. § 14:73.13(C)(1).

knowledge.[28] Georgia, Hawaii, New York, and Virginia's criminal laws require an intent to cause some sort of harm to the victim.[29] Intent to harm satisfies South Dakota's *mens rea* too, but, uniquely, so does "intent to self-gratify."[30] Texas is alone in requiring that a deepfake be "created with the intent to deceive."[31] Criminal laws in Alabama, Florida, Louisiana, and Utah have comparatively lower intent and *scienter* requirements.[32] Civil anti-deepfakes laws, meanwhile, tend not to require proof that a violator intended to harm the depicted individual, and instead require actual or constructive knowledge that a depicted person did not consent to the creation or disclosure of a deepfake.[33]

### 5. Subject Matter

Anti-deepfakes laws cover only specific subject matter. The broadest state laws cover both pictorial and aural representations.[34] Others focus on visual depictions, usually limited to videos and still images.[35] No state law covers written or spoken words, although words alone are of course perfectly capable of falsely portraying an identifiable person engaging in sexual activity.[36]

---

28. ALA. CODE § 13A-6-240(a); FLA. STAT. § 836.13(2) ; GA. CODE § 16-11-90(b) ; HAW. REV. STAT. § 711-1110.9(1)(c) ; LA. STAT. § 14:73.13; N.Y. PENAL LAW § 245.15 (McKinney)(1)(a); S.D. CODIFIED LAWS § 22-21-4(3); TEX. PENAL CODE § 21.165(b) ; UTAH CODE § 76-5b-205(2)(a) ; VA. CODE § 18.2-386.2 ("maliciously"); Essex v. Commonwealth, 322 S.E.2d 216, 220 (Va. 1984) ("malice, . . . require[s] that a wrongful act be done 'wilfully or purposefully.' (quoting Williamson v. Commonwealth, 23 S.E.2d 240, 241 (Va. 1942)).
29. GA. CODE § 16-11-90(a)(1), (b) ; HAW. REV. STAT. § 711-1110.9 ; N.Y. PENAL LAW § 245.15 (McKinney); VA. CODE § 18.2-386.2(A) .
30. S.D. CODIFIED LAWS § 22-21-4(3)(d).
31. TEX. PENAL CODE § 21.165 .
32. FLA. STAT. § 836.13(2) ("know[] or reasonably should have known that [the] visual depiction was an altered sexual depiction."). LA. STAT. § 14:73.13(B)(1) ("knowledge that the material is a deepfake that depicts another person."). UTAH CODE § 76-5b-205(2)(a) ("know[] or should reasonably know would cause a reasonable person to suffer emotional or physical distress or harm"). ALA. CODE § 13A-6-240(a) ("knowing[]" distribution).
33. *See, e.g.*, CAL. CIV. CODE § 1708.86(b) ; 740 ILL. COMP. STAT. 190/10(a); MINN. STAT. § 604.32.2(a)(1) ; N.Y. CIV. RIGHTS LAW § 52-c(2)(a) (McKinney).
34. LA. STAT. § 14:73.13; MINN. STAT. 604.32(1)(b); NO FAKES Act of 2024 § 2(a)(1).
35. CAL. CIV. CODE § 1708.86(3)(A) ; N.Y. CIV. RIGHTS LAW § 52-c(3)(a) (McKinney); N.Y. PENAL LAW § 245.15(1)(a) (McKinney); TEX. PENAL CODE § 21.165 ; VA. CODE § 18.2-386.2 .
36. *Cf., e.g.*, James v. Gannett Co., 353 N.E.2d 834, 837 (N.Y. 1976) ("It is old law that written charges imputing unchaste conduct to a woman are libelous per se . . . .). Amy Adler notes

Some regimes cover media "created by any means whatsoever."[37] Other laws target "digitally altered" media.[38] Minnesota's requires the production of the deepfake to have been "substantially dependent upon technical means, rather than the ability of another individual to physically or verbally impersonate such individual."[39] California's civil law and New York's criminal and civil laws limit recourse to individuals who are depicted in deepfakes through "digitization."[40] Strangely, however, California's and New York's civil laws define "digitization" to include "realistic[] depict[ions]" without limitation, such that even realism achieved without digitization would seem to be covered.[41]

### 6. No Deceptiveness Requirement

One of the most unusual characteristics of the laws that forbid pornographic deepfakes—and the characteristic most dissonant with privacy and defamation theories of harm—is that they frequently provide that disclaimers of falsity are no defense to liability. Anti-deepfakes laws in California, Florida, Illinois, New York, and Washington, as well as three proposed federal statutes, provide expressly that violators may be liable even when the deepfake in question contains a disclaimer that communicates that it was made without the authorization of the person depicted in it, and/or that it does not depict the person's actual behavior.[42]

Not every regulation of pornographic deepfakes treats disclaimers as irrelevant, however. Louisiana's law specifically excludes from its scope "any material . . . that includes content, context, or a clear disclosure visible throughout the duration of the recording that would cause a reasonable person to understand that the audio or visual media is not a record of a real event."[43] Texas's law defines

---

that disfavor for images is a durable feature of obscenity and pornography law. *See generally* Amy Adler, *The First Amendment and the Second Commandment*, *in* LAW, CULTURE AND VISUAL STUDIES 161 (Anne Wagner & Richard K. Sherwin eds., 2014), https://link.springer.com/10.1007/978-90-481-9322-6_8 (last visited Jan 15, 2024).

37. VA. CODE § 18.2-386.2 .
38. 740 ILL. COMP. STAT. 190/10(a), (a)(2), (c). *See also* NO FAKES Act of 2024 § (2)(a)(1) ("computer-generated").
39. MINN. STAT. § 604.32(b)(2).
40. CAL. CIV. CODE § 1708.86(a)(4); N.Y. PENAL LAW § 245.15(1)(a), (2)(d) (McKinney).
41. CAL. CIV. CODE § 1708.86(a)(6) ; N.Y. CIV. RIGHTS LAW § 52-c(1)(b) (McKinney).
42. CAL. CIV. CODE § 1708.86(d) ; FLA. STAT. § 836.13(4) ; 740 ILL. COMP. STAT. 190/10(c); N.Y. CIV. RIGHTS LAW § 52-c(2)(b) (McKinney); Wash. Subst. H.B. 1999, 68th Leg., 2024 Sess. at 14 (2024); NO FAKES Act of 2024 § 2(e)(3); NO AI FRAUD Act § 3(c)(2)(D); DEFIANCE Act of 2024 § 2(a)(4).
43. LA. STAT. § 14:73.13.

actionable deepfakes as having been "created with the intent to deceive"; presumably, a deepfake creator who included an effective disclaimer would not meet this intent requirement.[44] And in contrast to laws regulating pornographic deepfakes, laws regulating deepfakes intended to influence elections typically do extinguish liability when a disclaimer is present.[45]

**7. "Realism"**

Typical anti-deepfakes laws expressly target material that is "realistic"—without defining what "realism" is.[46] "Realistic" must mean something different from "deceptive," because the laws require no proof of deception and often specify that a violation can occur even when the media contains a disclaimer that it is false.[47] In colloquial usage, too, media's realism is distinct from its tendency to deceive. In 2019, Disney released a remake of *The Lion King* that redid the 1994 cartoon film in what press coverage called an "incredibly realistic" computer-generated style, so realistic that the remake was described as "live action."[48] But the photorealistic *Lion King* remake didn't deceive its audience into believing that Disney had turned actual Serengeti beasts into thespians: one commentator wrote that "live action" "can't possibly be an accurate description, because — the last time we checked — lions, hyenas, meerkats and warthogs don't actually talk. Or dance. Or burst into

---

44. TEX. PENAL CODE § 21.165(a)(1) .

45. *See, e.g.*, CAL. ELEC. CODE § 20010(b) ; NY LEGIS 58 (2024), 2024 SESS. LAW NEWS OF N.Y. CH. 58 (A. 8808-C) (McKinney) (amending N.Y. ELEC. LAW § 14-106 to require disclosure, "This (image, video, or audio) has been manipulated," in "any political communication that was produced by or includes materially deceptive media"); WASH. REV. CODE § 42.62.020(4) . One election-related anti-deepfake law has already been preliminarily enjoined as unconstitutional. *See* Order Granting Plaintiff's Motion for Preliminary Injunction, Dkt. No. 14, Kohls v. Bonta, No. 2:24-cv-02527 (E.D. Cal. Oct. 2, 2024).

46. CAL. CIV. CODE § 1708.86(a)(6); WEST'S F.S.A. § 836.13(a); MINN. STAT. § 604.32.1(b)(1) ; N.Y. CIV. RIGHTS LAW § 52-c (McKinney); N.Y. PENAL LAW § 245.15(2)(d) (McKinney); S.D. CODIFIED LAWS § 22-21-4(3)(a)

47. CAL. CIV. CODE § 1708.86(d); WEST'S F.S.A. § 836.13(4); N.Y. CIV. RIGHTS LAW § 52-c(2)(b) (McKinney).

48. Abrar Al-Heeti, *The Lion King Review: Remake Might Be Too Realistic for Its Own Good*, CNET, https://www.cnet.com/culture/entertainment/the-lion-king-review-remake-of-disney-classic-might-be-too-realistic-for-its-own-good/ (last visited Jan 9, 2024); Josh Rottenberg, *"The Lion King": Is It Animated or Live-Action? It's Complicated*, LOS ANGELES TIMES (2019), https://www.latimes.com/entertainment-arts/movies/story/2019-07-19/the-lion-king-remake-animation-live-action-photo-real (last visited Jan 9, 2024).

song."[49] Moreover, because realism is culturally contingent, deceptive media are not necessarily realistic—unless "realistic" just means "believable" and "believable" just means "someone believed it to document true events."[50]

Realism is always contextual and stylized.[51] Some anti-deepfakes statutes are clearer than others about what style of realism they target. Louisiana, Massachusetts, South Dakota, and the DEFIANCE Act all require a sort of realism that would make the media appear "authentic" to a reasonable observer.[52] Louisiana's law refers specifically to media that "record . . . the actual speech or conduct of the individual" depicted.[53] In this context, the word "authentic" denotes documentary media, like photographs and video recordings. Unlike other forms of pictorial representation, which merely *depict* events, documentary media actually *record* real-life events.[54] As I use these words, to depict an event is to represent it pictorially, whereas to record an event is to capture contemporaneous evidence of its occurrence.[55] A record may be, but is not necessarily, a depiction, and vice-versa. In this sense, an "authentic" photograph is an image that is actually a photograph. Although one might also ask whether an oil painting is "authentic," this question typically connotes whether it was painted by a particular artist, not whether it is actually an oil painting. Statutes that use the word "authentic" almost certainly require *photorealism*, "the quality in art . . . of depicting or seeming to depict real people . . . with the exactness of a photograph."[56] A photorealistic image is one rendered in a style of realism that resembles an "authentic" photograph or video recording. Statutes that limit their coverage to "recording[s] of a person"[57] or media

49. Rottenberg, *supra* note 49.
50. *Cf.* Rebecca Tushnet, *Worth a Thousand Words: The Images of Copyright*, 125 HARV. L. REV. 683, 730 (2011).
51. *Id.* at 724; Benjamin L.W. Sobel, *Elements of Style: Copyright, Similarity, and Generative AI*, 38.1 HARV. J.L. & TECH. __, __ (forthcoming 2025).
52. LA. STAT. § 14:73.13(C)(1); 2023 MASS. H.B. 4744, 193RD GENERAL COURT OF THE COMMONWEALTH OF MASS.; S.D. CODIFIED LAWS § 22-21-4(3)(a); DEFIANCE Act of 2024 § (3)(a)(3)(D).
53. LA. STAT. § 14:73.13(C)(1).
54. *Compare* Oxford English Dictionary, s.v. "depict (v.)," July 2023, https://doi.org/10.1093/OED/2187623999 ("to portray, delineate, figure anyhow") *with* Oxford English Dictionary, s.v. "record (v.1)," July 2023, https://doi.org/10.1093/OED/1987418679 ("To convert (sounds, images, a broadcast, etc.) into permanent form . . . chiefly using magnetic tape or digital electronic techniques.")
55. Part II, *infra*, explains this distinction more precisely using semiotic terminology.
56. s.v. "photorealism," MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/photorealism (last visited January 24, 2024).
57. ALA. CODE § 13A-6-240(b).

"substantially derivative" of video recordings and photographs[58] are probably also meant to capture photorealism.

Other anti-deepfakes laws use more ambiguous language. Florida's law, for example, covers "any visual depiction that . . . depicts a realistic version of an identifiable person" nude or engaged in sexual conduct.[59] Texas's statute—which does not specify that covered material must be "realistic"—covers "video . . . that appears to depict a real person performing an action that did not occur in reality."[60] These laws are probably intended to regulate photorealistic, AI-generated media rather than pictorial depictions in general.[61] But because the statutes do not refer specifically to documentary media, and because they cover "depict[ions]" rather than apparent "recordings," their text could cover all sorts of pictorial depictions. Texas's definition, for example, literally encompasses videos of fictionalized theatre or puppetry performances, which can "depict a real person performing an action that did not occur."[62]

58. MINN. STAT. § 604.32.1(b) . Minnesota's statute also includes "electronic image[s]," but because this term is enumerated in a list of otherwise documentary media, it probably does not encompass non-documentary images. *See* Yates v. United States, 574 U.S. 528, 543 (2015) (describing "the principle of *noscitur a sociis*").

59. FLA. STAT. § 836.13(1)(A) .

60. TEX. PENAL CODE § 21.165(1)(a) . *See also, e.g.*, IND. CODE § 35-45-4-8(c)(3) ; UTAH CODE § 76-5b-205(1)(a)(ii) .

61. The "sponsor's statement of intent" accompanying the Texas bill specifically mentions "artificial intelligence." 2023 Texas Senate Bill No. 1361 Texas Eighty-Eighth Legislature, *available at* https://capitol.texas.gov/tlodocs/88R/analysis/html/SB01361F.htm. Similarly, a legislative report on the Florida statute frames the bill as a response to "technology advancing at a rapid rate." Bill Analysis and Fiscal Impact at 2, 2022 Florida Senate Bill No. 1798, Florida One Hundred Twenty-Fourth Regular Session (Jan. 26, 2022).

62. Both the musical *Hamilton* and the marionette movie *Team America: World Police* depict real people—Alexander Hamilton and Kim Jong-Il, respectively—performing musical numbers that, to the best of my knowledge, they never performed.

#### 8. A Distillation of the Typical Anti-Deepfakes Law

The survey above indicates that, while statutes differ, the typical civil or criminal anti-deepfakes law:

- Addresses nonconsensual deepfakes depicting sex or nudity.
- Requires that the victim be identifiable.
- Forbids dissemination of a deepfake, but not mere creation or possession.
- Does not necessarily require intent to harm the depicted person.
- Applies only to images and videos and focuses on digitally altered media.
- Does not require that the deepfake be deceptive, and may expressly foreclose defenses based on a disclaimer that the deepfake is unauthorized and ahistorical.
- Is intended to apply only to photorealistic media.

These requirements track the harms intuitively linked to nonconsensual pornographic deepfakes. They are not, however, a constellation of elements that privacy and defamation doctrine neatly explain. The following sub-Parts explain how privacy and defamation doctrine fall short, and characterize the interest that anti-deepfakes laws actually protect.

### B. The Typical Anti-Deepfakes Law Extends Beyond the Dignitary Torts

Many uses of deepfakes present straightforward harms sounding in defamation or invasion of privacy. But the typical anti-deepfakes law sweeps more broadly. It prohibits certain degrading uses of likeness irrespective of whether they communicate a true or false factual proposition. The breadth of this prohibition tracks our intuitive notions of harm, but it also bypasses the limitations that constrain the established dignitary torts.

Defamation and privacy doctrines in the United States tend, with one notable exception, to focus on actual or purported facts about a plaintiff. Defamatory harm is injury wrought by false statements of fact. Paradigmatic invasion of privacy is injury wrought by improper access to or use of true facts about a person.[63] These doctrines' focus on true or false factual propositions meaningfully

63. *Cf. generally* Danielle Keats Citron & Daniel J. Solove, *Privacy Harms*, 102 B.U. L. REV. 793, 810 (2022) (focusing on "privacy problems involving the collection, use, and disclosure of personal data").

limits them. Statements that do not assert facts about a person cannot defame, and—with one exception—cannot invade that person's privacy, either.

The exception, lumped in with the privacy torts, is the tort of appropriation of likeness. Appropriation is unconcerned with factual representations about persons. As formulated in the Restatement, the tort simply requires a defendant's "appropriat[ion] to his own use or benefit the name or likeness of another."[64] A textbook case of appropriation is the nonconsensual use of a plaintiff's likeness in advertising. But appropriation isn't false endorsement; it doesn't require proof that a use of likeness was actually likely to mislead anyone. Rather, appropriation redresses the dignitary harm wrought by "the use of one's identity or personality for the purposes and goals of another."[65] Appropriation affords certain control over the indicia of identity—faces, names, and so on—irrespective of whether a use of likeness communicates any false statements of fact, or any true, private facts, about the person it identifies. It is appropriation that best describes the harm caused by nonconsensual, pornographic deepfakes, but contemporary appropriation doctrine has limits that anti-deepfakes laws exceed.

The following subsections explain how and why anti-deepfakes laws address more than defamatory harm and more than generic invasion of privacy. When proponents of anti-deepfakes laws invoke "privacy," they are appealing to a specific and distinct theory of privacy: appropriation. Indeed, anti-deepfakes statutes are best understood as an expansive civil and criminal law of appropriation. Like appropriation, anti-deepfakes laws eschew the limits of defamation and paradigmatic invasion of privacy. Anti-deepfakes laws *also*, however, eschew appropriation's limitation to uses that improperly "advantage" a defendant.

Does this mean that anti-deepfakes laws are blanket prohibitions on offensive expressions of opinion about a person, bereft of any limiting principles—and, as such, likely unconstitutional? No. Anti-deepfakes laws *do* have a limiting principle. But all the rhetorical focus on factual falsity and/or invasion of privacy obscures it. Anti-deepfakes laws are an appropriation regime that replaces appropriation's focus on commercial use with a limitation from a different privacy tort, false light: outrageousness.

### 1. Beyond Defamation

Some legislative history justifies anti-deepfakes laws by referring to viewers' inability to differentiate deepfakes from factual records, which suggests concerns

64. RESTATEMENT (SECOND) OF TORTS § 652C (1977).
65. Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477, 546 (2005).

about misleading statements of fact.[66] Similarly, legislation may refer to deepfakes as "digital forger[ies]" or imply that they entail "fraud," words that presuppose an intent to deceive viewers about a factual proposition.[67] But a common characteristic of many anti-deepfake laws reveals that these statutes seek to remedy a wrong distinct from harmful falsehood. The California, Florida, New York, and Washington laws, as well as several draft federal bills, state explicitly that a violator cannot avoid liability by including a disclaimer that states that a deepfake does not record factual conduct by the person depicted.[68] This provision is inconsistent with the theory that the harmfulness of deepfakes comes solely from their ability to defame, deceive, and defraud.

According to the Supreme Court, "statements that cannot 'reasonably be interpreted as stating actual facts' about an individual" are protected by the First Amendment, at least when they pertain to matters of public concern.[69] This requirement reflects that a "false representation of fact" is an essential element of defamation.[70] As the Court has explained, "[u]nder the First Amendment there is no such thing as a false idea."[71] On this understanding, an expressive statement that does not assert or imply a factual proposition cannot be defamatory.[72] The Restatement's definition of defamation corroborates this view, although as a

---

66. *See* 2023 N.Y. S.B. 1042, 246th SESS. ("As this technology improves, . . . it becomes nearly impossible to depict what is a real image and what is doctored."); Public Hearing re: SHB 1999, Wa. Senate Law & Justice Committee (Feb. 16, 2024) at 43:25 https://tvw.org/video/senate-law-justice-2024021275/?eventID=2024021275 (remarks of Rep. Tina Orwall) ("The bill in front of you is really about the fabricated images . . . . [T]hey're just as harmful, right? Someone cannot distinguish."). *See also generally* Marc Jonathan Blitz, *Deepfakes and Other Non-Testimonial Falsehoods: When Is Belief Manipulation (Not) First Amendment Speech?*, 23 YALE J.L. & TECH. 160 (2020) (characterizing the harms of deepfakes as related to falsity).

67. *See generally, e.g.*, DEFIANCE Act of 2024; No AI FRAUD Act. *See also* FRAUD, BLACK'S LAW DICTIONARY (12TH ED. 2024) ("A knowing misrepresentation or knowing concealment of a material fact made to induce another to act to his or her detriment"); FORGERY, BLACK'S LAW DICTIONARY (12TH ED. 2024) ("The act of fraudulently making a false document or altering a real one to be used as if genuine").

68. *See supra*, Part I.A.6.

69. Milkovich v. Lorain J. Co., 497 U.S. 1, 20 (1990) (quoting Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 50 (1988)).

70. Pring v. Penthouse Int'l, Ltd., 695 F.2d 438, 440 (10th Cir. 1982).

71. Gertz v. Robert Welch, Inc., 418 U.S. 323, 339–40 (1974).

72. I am using the word "defamatory" as shorthand for "*actionably* defamatory." *Cf.* Jeffrey S. Helmreich, *True Defamation*, 4 J. FREE SPEECH L. 835, 840–46 (2024) (noting that some definitions of "defamatory" include truthful speech).

constitutional matter it remains an "open issue" whether "statements not provably false about matters of purely private significance" can be actionably defamatory.[73]

If anti-deepfake laws addressed only defamatory harm, it would make no sense for them to provide for liability even when the deepfake is accompanied by a disclaimer that effectively communicates that it is unauthorized and ahistorical. Of course, the presence of a disclaimer does not render a statement non-defamatory *per se*.[74] But some disclaimers surely make it impossible for any reasonable viewer to interpret a deepfake "as stating actual facts about an individual."[75] To declare disclaimers legally irrelevant, as many anti-deepfakes laws do, is to declare that legally cognizable harm occurs even when no reasonable viewer could understand a deepfake as stating actual facts about the individual depicted. In other words, it is to acknowledge that legally cognizable harm occurs even when a deepfake is non-defamatory as a matter of law.

The Senate made precisely this acknowledgement in late July of 2024, when it passed a version of the DEFIANCE Act that had been amended with a finding that "individuals depicted in [sexually explicit] digital forgeries are profoundly harmed when the content is produced, disclosed, or obtained without the consent of those individuals. *These harms are not mitigated through labels or other information that indicates that the depiction is fake*."[76] In a decision published two days later, Meta's Oversight Board made the same observation.[77] Popular attitudes, too, corroborate that the perceived harms of deepfakes are about something more than defamation. One study asked participants to evaluate a hypothetical nonconsensual, pornographic deepfake that was labeled as false and found that "[t]here was no significant effect of labeling on the perceived harmfulness or blameworthiness of the video."[78] Indeed, there is ample reason to think that relatively few primary consumers of deepfakes actually believe the materials are documentary. The

---

73. RESTATEMENT (SECOND) OF TORTS § 566 (1977) (opinion "is actionable *only if* it implies the allegation of undisclosed defamatory facts" (emphasis added)); § 14:2.4 SACK ON DEFAMATION, 4-25—26. *See also* Robert D. Sack, *Protection of Opinion under the First Amendment: Reflections on Alfred Hill, Defamation and Privacy under the First Amendment*, 100 COLUM. L. REV. 294, 329–30 (2000).

74. New Times, Inc. v. Isaacks, 146 S.W.3d 144, 161 (Tex. 2004).

75. *Milkovich*, 497 U.S. at 20 (quotation marks omitted); *see* Stanton v. Metro Corp., 438 F.3d 119, 128 (1st Cir. 2006).

76. DEFIANCE Act of 2024 § 2(3) (emphasis added).

77. OVERSIGHT BD., *Multiple Case Decision IG-JPEE85LL*, https://www.oversightboard.com/decision/bun-7e941o1n/ (Jul. 25, 2024).

78. Matthew B. Kugler & Carly Pace, *Deepfake Privacy: Attitudes and Regulation*, 116 NW. U. L. REV. 611, 639 (2021).

anonymous Reddit user whose work kicked off the mainstream crisis response to deepfakes—and whose alias is the namesake of the word "deepfake"—spent his time inserting the likenesses of young, female celebrities into hardcore pornography. Viewers looking to watch or share these videos would congregate in a forum called "r/deepfakes," which Reddit has since taken offline.[79] It is implausible that patrons of a forum with "fake" in its name, accessing videos depicting A-list celebrities in hardcore pornography, thought that they were viewing recordings of real-world events.[80] Similarly, a recent spate of outraged writeups focused on the AI platform Civitai's "bounty" system, which permits requests for "AI models that generate images of . . . specific real people, and reward[s] the best AI model . . . with a virtual currency."[81] Civitai bounties don't solicit documentary images of real people; they quite intentionally solicit tools for synthesizing images. Of course, material from r/deepfakes proliferated across the Internet and was sometimes presented as historical fact, and this scenario raises familiar defamation concerns.[82] But the focus of the outraged press coverage was that the r/deepfakes forum and deepfake bounties existed at all, not that deepfakes might eventually proliferate in contexts that do not indicate that they are synthetic.

In a refreshing *New Yorker* article from late 2023, Daniel Immerwahr recognizes that the real trouble with deepfakes is not that they deceive their audiences. He writes,

> A.I.-generated videos are not, in general, operating in our media as counterfeited evidence. Their role better resembles that of cartoons, especially smutty ones.
>
> Manipulated media is far from harmless, but its harms have not been epistemic. Rather, they've been demagogic, giving voice to what the historian Sam Lebovic calls "the politics of outrageous expression." At their best, fakes—GIFS, memes, and the like—condense complex thoughts into clarifying, rousing images. But,

---

79. Samantha Cole, *We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now*, VICE (Jan. 24, 2018), https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley (last visited Dec 18, 2023).

80. *See* Quentin J. Ullrich, *Is This Video Real? The Principal Mischief of Deepfakes and How the Lanham Act Can Address It*, 55 COLUM. JL & SOC. PROBS. 1, 8 (2021) (acknowledging that "[v]iewers of videos on sites like MrDeepFakes almost certainly know they are fake").

81. Emanuel Maiberg ·, *Giant AI Platform Introduces 'Bounties' for Deepfakes of Real People*, 404 MEDIA (2023), https://www.404media.co/giant-ai-platform-introduces-bounties-for-nonconsensual-images-of-real-people/ (last visited Dec 18, 2023).

82. Cole, *supra* note 80.

> at their worst, they amplify our views rather than complicate them, facilitate the harassment of women, and help turn politics into a blood sport won by insulting one's opponent in an entertaining way.[83]

Immerwahr corrects prevailing accounts of deepfakes in two important ways. First, instead of parroting the frequent assertion that deepfakes are uniquely powerful tools of deception, Immerwahr observes that deepfakes do not, in fact, seem to deceive people all that much more than lower-tech fraud does. And just as importantly, Immerwahr apprehends what deepfakes' *real* harm is: they facilitate not false expression, but "outrageous expression."[84] Deepfakes manipulate people's effigies in a way that outrages us more than the manipulation of mere symbols ordinarily does.

### 2. Beyond Paradigmatic Invasion of Privacy

Legislators and scholars also frequently characterize deepfakes' harms as privacy harms.[85] Scholars describe the harm of a deepfake as a hijacking of identity: Danielle Citron has commented that deepfakes "mak[e] [a subject] be a sexual object in ways that [she] didn't choose. . . . [I]t takes your sexual identity and

---

83. Daniel Immerwahr, *What the Doomsayers Get Wrong About Deepfakes*, THE NEW YORKER, Nov. 2023, https://www.newyorker.com/magazine/2023/11/20/a-history-of-fake-things-on-the-internet-walter-j-scheirer-book-review (last visited Dec 18, 2023).

84. *Id.*

85. DEFIANCE Act of 2024 § 2(4) ("the privacy of . . . victims is violated"); Natalie Lussier, *Nonconsensual Deepfakes: Detecting and Regulating This Rising Threat to Privacy*, 58 IDAHO L. REV. 353 (2022); Citron, *supra* note 5 at 1924–25 (discussing the harm of nonconsensual, pornographic deepfakes as a "sexual-privacy invasion"); Congressman Joe Morelle Authors Legislation to Make AI-Generated Deepfakes Illegal, U.S. CONGRESSMAN JOSEPH MORELLE (2023), http://morelle.house.gov/media/press-releases/congressman-joe-morelle-authors-legislation-make-ai-generated-deepfakes (last visited Jan 25, 2024) (announcing "legislation to protect the right to privacy online amid a rise of artificial intelligence [AI] and digitally-manipulated content."); SB309 HD2, https://www.capitol.hawaii.gov/sessions/session2021/bills/SB309_HD2_.HTM (last visited Jan 25, 2024) (discussing "privacy issues . . . including . . . deep fake technology."); *Cf. also* Danielle Citron & Mary Anne Franks, *Evaluating New York's "Revenge Porn" Law: A Missed Opportunity to Protect Sexual Privacy*, HARVARD LAW REVIEW BLOG (Mar. 19, 2019), https://harvardlawreview.org/blog/2019/03/evaluating-new-yorks-revenge-porn-law-a-missed-opportunity-to-protect-sexual-privacy/ (last visited Feb 13, 2024) (discussing revenge porn, not deepfakes).

exposes it in ways you didn't choose."[86] Franks and Waldman explain that deepfake pornography "turns individuals into objects of sexual entertainment against their will, causing intense distress, humiliation, and reputational injury."[87] Having the autonomy to determine one's sexual identity wrenched away, Citron argues, interferes with liberty, autonomy, and self-development.[88]

Citron is justified in characterizing the harms of deepfakes as privacy harms. But she is justified by one very specific theory of privacy that differs materially from every other contemporary privacy doctrine recognized in the United States. Privacy law's paradigm case is the collection or disclosure of true information about a person. Deepfakes are different: what makes them objectionable is precisely that they do *not* disclose true information about a person. Deepfakes don't necessarily communicate any private facts, or any falsehoods, about their subject at all. Rather, deepfakes appropriate a person's likeness in an objectionable way.

We can, by process of elimination, expose the mismatch between conventional invasion of privacy and the harms of deepfakes. Daniel Solove sets forth a taxonomy of privacy that contains sixteen subcategories of "socially recognized privacy violations," divided into four main classes.[89] The harms of nonconsensual deepfakes clearly fall outside all subcategories except one: appropriation.[90]

Solove's category of "information collection" focuses entirely on the adverse consequences that flow from the collection of factual information about persons, either through surveillance or interrogation.[91] Deepfakes do not implicate an interest in preventing information collection, nor do they make factual inferences about persons derived from collected information. Rather, deepfakes exploit readily available factual information about people's likenesses in order to confabulate non-factual expression using those likenesses.

Nor do deepfakes implicate the harms Solove associates with "information processing." "Aggregation," "identification," "insecurity," and "exclusion" all, unlike deepfakes, involve the acquisition or derivation of true facts about a person, or

---

86. Brian Feldman, *Danielle Citron on Deepfakes and the Katie Hill Scandal*, INTELLIGENCER (2019), https://nymag.com/intelligencer/2019/10/danielle-citron-on-the-danger-of-deepfakes-and-revenge-porn.html (last visited Jan 10, 2024).

87. Franks and Waldman, *supra* note 16 at 893.

88. Citron, *supra* note 5 at 1884–85. *Cf.* Citron and Franks, *supra* note 67 (discussing revenge porn, not deepfakes).

89. Solove, *supra* note 66 at 483.

90. Thanks to James Grimmelmann for suggesting this approach.

91. Solove, *supra* note 66 at 491–505.

access to and correction of purportedly true but erroneous information.[92] Meanwhile, the primary harm Solove identifies in "secondary use" is that it delegitimizes a person's initial choice to reveal information under the expectation that it will be put to limited use.[93] But the production of deepfakes doesn't abuse an initial choice to reveal information, because deepfakes require only information that subjects voluntarily broadcast.

Equally inapplicable are the harms Solove catalogs as "invasion." Deepfakes can be created and consumed without any "intrusion" upon a person's seclusion. Nor do deepfakes implicate Solove's concept of "decisional interference," which denotes "governmental interference with people's decisions regarding certain matters of their lives" such as the choice to make certain sexual and reproductive decisions.[94] If anything, a policy against decisional interference supports an unrestricted entitlement to create and consume deepfakes in private.[95]

Finally, it's easy to knock out all but one of the harms that Solove places under the umbrella of "information dissemination." "Disclosure," "exposure," "increased accessibility," and "blackmail" are all out, since according to Solove these categories concern *true* information about a person.[96] "Breach of confidentiality" is harmful because it abuses a relationship of trust, but the creation of a deepfake does not require a trusting relationship between the creator and the subject.[97]

Another of Solove's information-dissemination harms, "distortion," seems like a good fit, because it "is the manipulation of the way a person is perceived and judged by others, and involves the victim being inaccurately exposed to the public."[98] According to Solove, distortion corresponds to the tort of false light

92. *Id.* at 508, 512, 517, 525.

93. *Id.* at 522.

94. *Id.* at 558–59, 561.

95. As an example of judicial protection from decisional interference, Solove cites *Stanley v. Georgia*, in which the Supreme Court struck down a ban on the private possession of obscene materials, reasoning that "the Constitution protects the right to receive information and ideas . . . regardless of their social worth" and that this "right takes on an added dimension" in a "prosecution for mere possession of printed or filmed matter in the privacy of a person's own home." *Id.* at 560 (quoting Stanley v. Georgia, 394 U.S. 557, 564 [1969]). *Stanley* would suggest that government regulation of the private creation and consumption of deepfakes abridges the decisional privacy of the deepfake consumer—although that interest might of course be subordinate to other interests. *See, e.g.*, Osborne v. Ohio, 495 U.S. 103, 108 (1990) ("[T]he interests underlying child pornography prohibitions far exceed the interests justifying the Georgia law at issue in *Stanley*.").

96. Solove, *supra* note 89 at 530, 536, 539–40, 543–44.

97. *Id.* at 527.

98. *Id.* at 550.

invasion of privacy.[99] False light actions give recourse to a plaintiff who was "place[d] . . . before the public in a false light" that "would be highly offensive to a reasonable person."[100] Paradigmatic false light claims include a newspaper's false insinuation that a family was living in "dirty and dilapidated conditions" and a tabloid's use of the plaintiff's photograph to illustrate a fictitious story of a centenarian who became pregnant after an extramarital affair.[101]

But current false light doctrine makes the tort a non-starter for non-deceptive deepfakes, because it incorporates the same falsity requirement as defamation.[102] Expression that "cannot 'reasonably [be] interpreted as stating actual facts' about an individual" cannot support a false light claim.[103] A false light claim may be predicated on true statements, but only if the way in which those truths are presented creates a specific false impression of fact.[104]

A recent paper by John Goldberg and Benjamin Zipursky argues for extending false light to address even deepfakes that disclaim their inauthenticity. Although they do not contest the authority requiring false light claims to rest on a specific false statement, Goldberg and Zipursky argue that "[f]or a whole range of non-newsworthy statements that are publicly disseminated and would be found highly offensive by a reasonable person, there *should* be no requirement to sort the true from the false because they should be actionable either way."[105] They observe specifically that pornographic deepfakes "should not become nonactionable just by virtue of a visible disclaimer (of nonauthenticity) on the image itself."[106] False light, they argue, protects a person's interest in controlling the "present[ation of] ordinarily private aspects of their lives to the public"—an interest presumably implicated whether or not an offensive depiction is understood as factual.[107]

---

99. *Cf. id.* at 549–50 ("[W]ith distortion, the information revealed is false and misleading."). *See* § 12:3.1 SACK ON DEFAMATION, 12-20—22.
100. RESTATEMENT (SECOND) OF TORTS § 652E (1977).
101. Cantrell v. Forest City Pub. Co., 419 U.S. 245, 248 (1974); Peoples Bank & Tr. Co. of Mountain Home v. Globe Int'l Pub., Inc., 978 F.2d 1065, 1067 (8th Cir. 1992).
102. *Cf.* Solove, *supra* note 66 at 550.
103. Khodorkovskaya v. Gay, 5 F.4th 80, 84 (D.C. Cir. 2021). *New York Times v. Sullivan*'s "actual malice" standard applies to false light claims involving "matters of public interest." Time, Inc. v. Hill, 385 U.S. 374, 387–88 (1967).
104. § 12:3.1 SACK ON DEFAMATION, 12-20–22.
105. Benjamin C. Zipursky & John C. P. Goldberg, *A Tort for the Digital Age: False Light Invasion of Privacy Reconsidered*, 73 DEPAUL L. REV. __, 18 (forthcoming) (emphasis added).
106. *Id.*
107. *Id.* at 13 (emphasis omitted); *see id.* at 18–19.

Goldberg and Zipursky correctly apprehend that the harm of pornographic deepfakes—unlike defamation or paradigmatic invasion of privacy—is independent of factual truth or falsity, because deepfakes are "highly offensive" either way.[108] Whether understood as factual or not, pornographic deepfakes interfere with self-determination, as Citron, Franks and Waldman, and others argue. But in my view, Goldberg and Zipursky's proposal—to turn the tort of false light into the tort of outrageous light—is better justified by the rationale underlying the tort of appropriation.

### 3. Deepfakes Are Highly Offensive Appropriations of Likeness

The one category in Solove's taxonomy that tracks the harms of deepfakes is appropriation. Appropriation is a confusing body of law because it has bifurcated into two doctrines that redress distinct injuries.[109] The first, the "right of publicity," is generally understood as a property right in commercial uses of name, image, and likeness, which permits individuals to internalize the commercial value of their identity.[110] Many proposals tout the right of publicity as a tool to combat harmful deepfakes, and the right may in many cases be an excellent fit.[111] But because the right of publicity paradigmatically requires commercial uses of likeness and emphasizes the economic value of a use rather than the dignitary harm it causes,

---

108. *Cf. id.* at 18.

109. Some commentators have suggested that appropriation and the right of publicity are better thought of as a single legal doctrine. *See* Eric E. Johnson, *Disentangling the Right of Publicity*, 111 NW. U. L. REV. 891, 903 (2016).

110. *Cf., e.g.*, Waits v. Frito-Lay, 978 F.2d 1093, 1098 (9th Cir. 1992). It bears noting that Jennifer Rothman, one of the foremost right of publicity scholars, resists a dichotomous "public-private" or "privacy-property" categorization and emphasizes that the right is rooted in privacy interests. *See generally* JENNIFER E. ROTHMAN, THE RIGHT OF PUBLICITY: PRIVACY REIMAGINED FOR A PUBLIC WORLD (2018).

111. Alice Preminger & Matthew B. Kugler, *The Right of Publicity Can Save Actors from Deepfake Armageddon*, 39 BERKELEY TECH. L.J. (forthcoming), https://ssrn.com/abstract=4563774; What's in a Name, Likeness, and Image? The Case for a Federal Right of Publicity Law, https://www.americanbar.org/groups/intellectual_property_law/publications/landslide/2020-21/september-october/what-s-in-a-name-likeness-image-case-for-federal-right-of-publicity-law/ (last visited Dec 31, 2023) ("The right of publicity could be an unexpected vehicle by which to combat th[e] issue" of pornographic deepfakes); Ullrich, *supra* note 81 at 26; Jesse Lempel, *Combatting Deepfakes through the Right of Publicity*, LAWFARE (2018), https://www.lawfaremedia.org/article/combatting-deepfakes-through-right-publicity (last visited Dec 31, 2023).

even its proponents acknowledge that it may not address pornographic deepfakes circulated in a noncommercial context.[112]

The second branch of appropriation jurisprudence, however, better approximates the harm of nonconsensual pornographic deepfakes, and it is this body of law I refer to as the appropriation tort. Like the right of publicity, the appropriation tort focuses on commercial uses of likeness, but unlike the right of publicity, the appropriation tort aims to recompense plaintiffs for dignitary injuries rather than missed licensing revenue. Appropriation is a privacy violation that wreaks a dignitary harm; it "turns a man into a commodity and makes him serve the economic needs and interest of others."[113]

Appropriation was first recognized, both at common law and in statutes, in the early 20th century.[114] In the 1905 case *Pavesich v. New England Life Insurance Company*, the Supreme Court of Georgia permitted a common-law privacy claim against an insurance company that used the plaintiff's portrait in an advertisement without his consent.[115] *Pavesich* identified a harm strikingly similar to the harm scholars associate with deepfakes. The court observed that a nonconsensual use of likeness in advertising can instill in the wronged person "a realization that his liberty has been taken away from him; and, as long as the advertiser uses him for these purposes, he cannot be otherwise than conscious of the fact that he is for the time being under the control of another . . . ."[116] Nonconsensual deepfakes can interfere with liberty in a similar way. As Citron explains, "Being able to reveal one's naked body, gender identity, or sexual orientation at the pace and in the way of one's choosing is crucial to identity formation. When the revelation of people's sexuality or gender is out of their hands at pivotal moments, it can shatter their sense of self."[117]

Indeed, *Pavesich*'s reasoning would seem to apply to commercial and noncommercial uses of likeness in equal measure. The court, scandalized, wrote,

> If one's picture may be used by another for advertising purposes, it may be reproduced and exhibited anywhere. If it may be used in a

112. Preminger and Kugler, *supra* note 112 at 43.
113. Edward J. Bloustein, *Privacy as an Aspect of Human Dignity: An Answer to Dean Prosser*, 39 N.Y.U. L. REV. 962, 988 (1964); *see also id.* at 871 (suggesting that dignitary interests of this sort are privacy interests).
114. For an early statute, see N.Y. L. 1909, Ch. 14. *See also* Lohan v. Take-Two Interactive Software, Inc., 97 N.E.3d 389, 393 (N.Y. 2018).
115. Pavesich v. New England Life Ins. Co., 50 S.E. 68 (Ga. 1905).
116. *Pavesich*, 122 Ga. at 80. *See also* Solove, *supra* note 89 at 548 (quoting *Pavesich*).
117. Citron, *supra* note 5 at 1884.

> newspaper, it may be used on a poster or a placard. It may be posted upon the walls of private dwellings or upon the streets. It may ornament the bar of the saloon keeper or decorate the walls of a brothel. By becoming a member of society, neither man nor woman can be presumed to have consented to such uses of the impression of their faces and features upon paper or upon canvas.[118]

Similarly, the Restatement asserts that the appropriation tort may apply even when "the use is not a commercial one, and even though the benefit sought to be obtained is not a pecuniary one."[119] At its most expansive, then, appropriation would seem to forbid almost any use of a plaintiff's likeness.

Appropriation's broad blackletter definition can't, however, be taken at face value. One scholar calls it "nonsensically overbroad."[120] Proscribing literally all uses of likeness that redound to a defendant's "advantage" or "benefit" would ensnare commonplace activities in the heartland of First Amendment protection, like news reporting, artistic photography and portraiture, fiction and nonfiction writing, and much more.[121] *Pavesich*'s scandalized discussion of portraiture clashes with present-day understandings of First Amendment rights.[122] Contrary to *Pavesich*'s assertion that "[b]y becoming a member of society, neither man nor woman can be presumed to have consented to" public displays of their likenesses, modern consensus is that by becoming a member of society we *have* generally assented to others publicly displaying our countenances.[123] In affirming the dismissal of a statutory appropriation claim by a plaintiff who was photographed on the street and used to illustrate a news story, the New York Court of Appeals observed in 1982 that, "other than in the purely commercial setting covered by [the statute], an inability to vindicate a personal predelection [sic] for greater privacy may be part of the price every person must be prepared to pay for a society in which information and opinion flow freely."[124]

---

118. 50 S.E. at 80.
119. RESTATEMENT (SECOND) OF TORTS § 652C & cmt. c (1977).
120. Johnson, *supra* note 110 at 906.
121. *Id.* at 905–07. *Cf.* Robert C. Post, *Rereading Warren and Brandeis: Privacy, Property, and Appropriation*, 41 CASE W. RES. L. REV. 647, 671–72 (1990).
122. See ROTHMAN, *supra* note 111 at 13. Indeed, *Pavesich* was decided before the Supreme Court held that the Fourteenth Amendment incorporated the First Amendment against state governments. Gitlow v. New York, 268 U.S. 652 (1925).
123. 50 S.E. at 80.
124. Arrington v. New York Times Co., 434 N.E.2d 1319, 1323 (1982). *See also* Nussenzweig v. DiCorcia, 814 N.Y.S.2d 891 (N.Y. Sup. Ct. 2006).

The development of the appropriation tort has been a search for limits consistent with free-speech values. Courts define a *prima facie* appropriation claim in extremely broad terms, then invoke the First Amendment to cabin the tort's practical reach.[125] As a result of these limitations, appropriation has been interpreted primarily as redressing the indignity that results from the "demeaning and humiliating . . . *commercialization* of an aspect of personality."[126] Courts, even as they recite the Restatement's broad definition, routinely reject appropriation claims based on uses of identity in expressive works, including expressive works sold for profit.[127] Treatises characterize appropriation as being undertaken "for a commercial use" or hedge and state that it is "usually for commercial gain."[128] Statutory regimes often expressly limit appropriation to uses in advertising or trade.[129] Scholarly commentary on appropriation often presents the cause of action as *commercial* appropriation of identity.[130] Use in advertising or trade has emerged as a *de facto* guardrail that appropriation's broad definition lacks.

Of course, many deepfakes—such as simulated product endorsements—are cut-and-dry appropriation. But harmful deepfakes often lack the indicia of "advantage" that delimits the appropriation tort. Deepfakes may be disseminated without any prospect of direct or indirect pecuniary gain. While a number of cases permit appropriation or right-of-publicity claims against noncommercial speech, these cases at least tend to involve expressive media sold for profit or situations in which the plaintiff's reputation or standing realizes publicity for a defendant.[131]

---

125. Johnson, *supra* note 110 at 904.

126. Bloustein, *supra* note 112 at 987 (emphasis added). *See also* Post, *supra* note 120 at 670 (quoting Bloustein); Solove, *supra* note 66 at 546 (quoting Bloustein).

127. *See, e.g.*, De Havilland v. FX Networks, LLC, 21 Cal. App. 5th 845, 860-62 (Ct. App. 2018); Benavidez v. Anheuser Busch, Inc., 873 F.2d 102, 104 (5th Cir. 1989); Neff v. Time, Inc., 406 F. Supp. 858, 861 (W.D. Pa. 1976).

128. William L. Prosser, *Miscellaneous Chapter 21*, 1 WILLIAM L. PROSSER, HANDBOOK OF THE LAW OF TORTS 1050, 1056 (1941). 2 LAW OF DEFAMATION § 10:2 (2d ed.), *cited in* Moore v. Sun Pub. Corp., 881 P.2d 735, 743 (N.M. Ct. App. 1994).

129. *See, e.g.*, N.Y. CIV. RIGHTS LAW § 50 (McKinney); CAL. CIV. CODE § 3344 ; WASH. REV. CODE § 63.60.050 .

130. Johnson, *supra* note 110 at 839; Samantha Barbas, *From Privacy to Publicity: The Tort of Appropriation in the Age of Mass Consumption*, 61 BUFF. L. REV. 1119, 1120 (2013); Mark P. McKenna, *The Right of Publicity and Autonomous Self-Definition*, 67 U. PITT. L. REV. 225, 225 n.2 (2005); Bloustein, *supra* note 114 at 985–86.

131. *See, e.g.*, Cardtoons, L.C. v. Major League Baseball Players Ass'n, 95 F.3d 959, 968, 970 (10th Cir. 1996) (use of baseball players' likenesses on parody trading cards violated Oklahoma right-of-publicity statute *prima facie*, but First Amendment precluded liability);

Meanwhile, noncommercial deepfakes do not trade on the plaintiff's identity in quite the manner that appropriation doctrine contemplates. Disseminating a deepfake in an online chat between schoolmates, for example, would violate anti-deepfakes statutes but probably wouldn't support a claim of appropriation.[132] An advertisement would qualify as appropriation if it used a plaintiff's name or her picture; whether the invocation is written or pictorial, the same identity is being employed for the same benefit. Meanwhile, sharing a written sexual fantasy that refers to another party by name would not violate an anti-deepfakes statute—or, absent unusual circumstances, any other law—but depicting that fantasy in a deepfake would. The written fantasy and the deepfake both employ the depicted person's identity for the same end. What distinguishes the deepfake is that it employs a more outrageous mode of expression.

What anti-deepfakes laws are, then, is a hybrid of false light and appropriation. Today's anti-deepfakes statutes redress the injury that appropriation redresses, subject not to appropriation's "advantage" requirement, but rather to the offensiveness limitation that appears in the false light tort. By focusing on the most offensive uses of identity—those that are (a) pornographic and (b) involve the manipulation of persons' realistic visual likenesses rather than merely the invocation of their names—anti-deepfakes laws incorporate something akin to false light's "highly offensive to a reasonable person" requirement, which is a limitation the appropriation tort lacks.[133] In turn, anti-deepfakes statutes eschew false light's requirement of a false statement and appropriation's focus on commercial use.

The "highly offensive" nature of deepfakes relates to a specific victim, not the general public. This is an important distinction: while it is unconstitutional to ban "express[ing] ideas that offend" in general,[134] offense *inflicted upon a specific person* is an element of numerous legal claims.[135] Nonconsensual deepfakes are highly offensive in the relevant sense because of (a) what they depict and (b) the

---

Doe v. TCI Cablevision, 110 S.W.3d 363, 371 (Mo. 2003); Browne v. McCain, 611 F. Supp. 2d 1062, 1070 (C.D. Cal. 2009) (denying motion to strike right-of-publicity claim arising out of use of a song in a political commercial). For an example closer to the paradigmatic deepfake scenario of private, noncommercial use, see Jarrett v. Butts, 379 S.E.2d 583, 585 (Ga. App. 1989) ("no wrongful appropriation occurred" where defendant photographed plaintiff and the "photographs were never sold, published, or *publicly* displayed" (emphasis added)).

132. *Cf. Jarrett*, 379 S.E.2d at 585.

133. *Compare* RESTATEMENT (SECOND) OF TORTS § 652E (1977) *with id.* § 652C.

134. *Tam*, 582 U.S. at 223

135. *See, e.g.*, RESTATEMENT (SECOND) OF TORTS § 19 (1965) (battery); *id.* § 652E (1977) (false light). *See also* JOEL FEINBERG, OFFENSE TO OTHERS 10-22 (1987).

circumstances of their production. Both ingredients matter. A depiction of a nude body is not highly offensive in itself.[136] Neither is a nonconsensual use of an identifiable person's likeness as such. What makes nonconsensual deepfakes offensive is that they are sexual depictions, rendered in a particular representational style, without the consent of the person depicted. The conjunction of these properties distinguishes deepfakes from sexual depictions generally and nonconsensual depictions generally. Unlike the harm to the general public that obscenity law addresses, the offensiveness of deepfakes is specific and relational.[137]

Thus, Citron, Goldberg, and Zipursky are all correct, because they're saying essentially the same thing.[138] Goldberg and Zipursky propose that "false light" drop the requirement of a false statement. Meanwhile, Citron proposes extending appropriation to cover outrageous, noncommercial speech.[139] Their arguments share a premise: certain offensive uses of likeness interfere with one's interest in controlling how intimate aspects of one's identity area presented to the public—*even if those uses of likeness are noncommercial and do not assert facts about the person depicted*. It is this premise that differentiates deepfakes' harms from those redressed by any established privacy or defamation doctrine.

---

136. Of course, attitudes vary. *See, e.g.*, *Justice Department Ends Nude Cover-Ups*, NBC NEWS, Jun. 26, 2005, https://www.nbcnews.com/id/wbna8360632 (last visited Oct 8, 2024) (describing nude statues in the Justice Department's atrium being restored to full view after being obscured by drapes during the tenure of Attorney General John Ashcroft).

137. Feminist legal scholarship differentiates between the alleged harms of obscenity and the harms of pornography. *See* Catharine A. MacKinnon, *Not a Moral Issue*, 2 YALE L. & POL'Y REV. 321, 329 (1983) (quoting Louis *Henkin, Morals and the Constitution: The Sin of Obscenity*, 63 COL. L. REV. 391, 395 (1963)); *id.* at 332.

138. Citron does not parse the different privacy torts as finely as Goldberg and Zipursky do, and Goldberg and Zipursky acknowledge that false light "shares with . . . appropriation . . . a focus on the wrong of interfering unduly with the entitlement of a person to exercise a degree of control over how they present ordinarily private aspects of their lives to the public." Zipursky and Goldberg, *supra* note 106 at 13 (emphasis omitted). Similarly, in an article that does not discuss pornographic deepfakes, Robert Post and Jennifer Rothman propose a "right of dignity" to "fill a gap in the existing dignitary torts" by restricting "highly offensive" appropriations of likeness irrespective of whether they defame, disclose private facts, or are intended to cause emotional distress. Robert C. Post & Jennifer E. Rothman, *The First Amendment and the Right(s) of Publicity*, 130 YALE L. J. 86, 122, 124-25 (2020).

139. DANIELLE KEATS CITRON, THE FIGHT FOR PRIVACY: PROTECTING DIGNITY, IDENTITY, AND LOVE IN THE DIGITAL AGE 137 (2022).

### 4. Why Getting the Theory Right Matters

Identifying anti-deepfakes laws' theoretical pedigree is important because it informs both how these statutes should be written and how they should be read. Recognizing that harmful deepfakes are highly offensive appropriations of likeness reveals respects in which some statutes are too narrow. Laws that extinguish liability in the presence of a disclaimer—as is the case in Arizona, Louisiana, the proposed federal Deepfakes Accountability Act, and probably Texas—misdiagnose an appropriative harm as a defamatory harm.[140] By contrast, laws that ban "depiction[s]" without limitation to highly offensive representational styles are unduly broad restrictions on fictional expression.[141]

Realizing that deepfakes' harms aren't really about assertions of fact also helps us reconstruct what ambiguous legislation is probably intended to mean, and how it might be redrafted to effectuate that intent better. As an illustration, consider the first draft of the federal TAKE IT DOWN Act, introduced by Sen. Ted Cruz on June 18, 2024. The bill passed the Senate in December 2024 in a significantly modified form and was reintroduced in January 2025, but its first draft exemplifies problems that appear in enacted state legislation.[142] The first draft of the TAKE IT DOWN Act provided that to be covered, a deepfake must "falsely depict an individual's appearance or conduct."[143] It further specified, "an individual appears in an intimate visual depiction if the individual is actually the individual identified in the intimate visual depiction; or a deepfake of the individual is used to realistically depict the individual such that a reasonable person would believe the individual is actually depicted in the intimate visual depiction."[144]

Looking closely at these provisions reveals that they make no sense. Let's start with "falsely." Is a "false[] depict[ion]" simply *any* ahistorical depiction? Or is it only a depiction that can reasonably be interpreted as asserting historical facts about a person? Many expressive works can't be placed in a true/false binary. Are Monet's haystack paintings true or false? The question is nonsensical: as Rebecca

140. ARIZ. REV. STAT. § 16-1023(A)(2); LA. STAT. § 14:73.13(C)(1); DEEPFAKES Accountability Act (c)-(e); TEX. PENAL CODE § 21.165(a)(1) .

141. *See, e.g.*, FLA. STAT. § 836.13.

142. Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act, S. __, 119TH CONG., 1ST SESS. (introduced in Senate Jan. 16, 2025), *available at* https://www.commerce.senate.gov/services/files/A42A827D-03B5-4377-9863-3B1263A7E3B2.

143. TAKE IT DOWN Act First Draft § (h)(1)(B).

144. *Id.* § (h)(1)(C)(ii) (numbering omitted).

Tushnet observes, "[visual] styles are neither true nor false."[145] *Star Wars* is not a documentary—but it's fiction rather than falsehood. Falsity denotes being "*[c]ontrary* to what is true, erroneous," or outright "mendacious."[146] If we don't interpret media to be reporting historical facts, then it isn't false—even if it isn't true, either.[147] Given that the core harm of deepfakes is that they appropriate likeness in an offensive manner, and not necessarily that they assert falsehoods, there is reason to suspect that the TAKE IT DOWN Act's unclear language was meant to address all *photorealistic* depictions (which may not be false) rather than all *false* depictions (which may not be photorealistic). Otherwise, the bill would just establish a takedown regime for defamation: it would cover images of some libelous oil paintings—which are indeed "false[] depict[ions]"—but not photorealistic deepfakes whose content and/or context clearly communicate that they are fictional.

What about the proposed provision that an individual "appears in an intimate visual depiction if the individual is actually the individual identified" in it?[148] Read literally, this text is almost meaningless. Being "actually the individual identified in [an] intimate visual depiction" was probably meant to cover situations like revenge porn, in which the victim "is actually the individual" *recorded* in the media. But being *recorded* is distinct from being *identified*. To "actually . . . [be] identified" is simply to be identified. A person identified in surveillance footage of a bank robbery and a person identified in an oil painting of a bank robbery is, in both cases, "actually" the person identified in the material in question. But only the surveillance footage directly evidences the individual's participation in the robbery, because only it is a record rather than a depiction.

Equally confusing was the bill's provision that an individual "appears in an intimate visual depiction" if she "is . . . realistically depict[ed] . . . such that a reasonable person would believe [she] is actually depicted."[149] This requirement can be read in at least three ways. On the narrowest reading, the definition requires that

---

145. Tushnet, *supra* note 51 at 724.

146. *Oxford English Dictionary*, s.v. "false (*adj.*, *adv.*, & *n.*)," June 2024, https://doi.org/10.1093/OED/8506717465 (emphasis added).

147. *See* Marc A. Franklin, Fiction, Libel, and the First Amendment, 51 BROOK. L. REV. 269, 273 (1984) ("[V]irtually the entire range of poetic and prose fiction would occupy this middle ground between truth and falsity. Language that does not purport to be reportorial is not automatically to be deprecated as 'false.'"). *Cf.* RICHARD A. POSNER, LAW AND LITERATURE 515 (2009) ("one of the adjustments we make in reading a work as literature rather than as history or sociology is generally to ignore issues of factuality").

148. TAKE IT DOWN Act First Draft § (h)(1)(C)(ii).

149. TAKE IT DOWN Act First Draft § (h)(1)(C)(ii).

a reasonable person actually would believe that the deepfake is an authentic photograph or video recording of the person it depicts. Reading the bill in this way makes the deepfake's content and context especially significant: if it contains a disclaimer that it is a deepfake, or if it depicts fantastical conduct that could never take place in real life, then it may not be reasonable for an observer to "believe it depicts speech or conduct of an individual," just as it isn't reasonable for a moviegoer to think that the photorealistic *Lion King* remake is an authentic record of real-life song-and-dance performances by anthropomorphic Serengeti animals.[150]

A broader reading of the first-draft TAKE IT DOWN Act would be that it required a deepfake to be photorealistic but not necessarily misleading. So construed, the law would have covered even deepfakes that are so obviously fake that no reasonable observer could mistake them for authentic photographs or videos, as long as they are rendered in a photorealistic style. This is the best reading of the numerous statutes that require "realism" while simultaneously providing that disclaimers of falsity are no defense to liability.[151] Unhelpfully, however, the first draft of the TAKE IT DOWN Act did not specify whether a disclaimer of falsity affects liability.

Finally, the broadest reading of this language is that it required only that the depicted individual be *identifiable* from the deepfake, and not that the deepfake be photorealistic at all. The first-draft bill required only "that a reasonable person would believe the individual is actually *depicted*," not that a reasonable observer would believe that the individual is actually *recorded*. This is an important distinction. Both the surveillance footage of the bank robbery and the oil painting of the bank robbery *depict* the robber. But only the footage *records* the robber. Other statutes acknowledge this difference: Louisiana's, for example, defines deepfakes as "falsely appear[ing] to a reasonable observer to be an authentic *record* of the actual speech or conduct of the individual."[152] The broadest reading of the first-draft TAKE IT DOWN Act, meanwhile, would require only that a reasonable observer would recognize that a specific individual is depicted in the deepfake—which means the deepfake could be a cartoon, a flipbook, or any other non-photorealistic pictorial medium capable of presentation in a still or video image.

Perhaps in recognition of the serious flaws in the proposed legislation, the TAKE IT DOWN Act passed the Senate only after substantial modifications

150. *See* Part I.A.7, *supra*.
151. *See id.*
152. LA. STAT. § 14:73.13(c)(1) (emphasis added).

addressing the issues cataloged above, and in its modified form it was reintroduced in the current Congress. The modified bill removes the incoherent language about "actual[] depict[ion]" as well as all mentions of the ambiguous word "false."[153] Instead, the revision defines a covered "digital forgery" as "any intimate visual depiction of an identifiable individual . . . that, *when viewed as a whole by a reasonable person, is indistinguishable from an authentic visual depiction of the individual*."[154] Much more clearly than the draft, the revised bill's coverage is limited to deepfakes that, "viewed as a whole," would deceive a reasonable observer into regarding them as factual records. The revision also addresses offenses involving "digital forgeries" in a separate section from offenses involving documentary revenge porn, while the first draft grouped them together.[155] Yet the first draft's shortcomings remain illustrative, because enacted state legislation recapitulates these errors. Delaware's statute, for example, contains the problematic "appears to depict" language, and it defines covered deepfakes as "realistic but false" media.[156] Georgia's law repeats the incoherence of the first-draft TAKE IT DOWN Act by covering "falsely created" images.[157] And many state laws attempt to treat deepfakes and revenge porn identically, despite fundamental ontological differences in these media.[158]

The conceptual incoherence of some anti-deepfakes statutes results from an imprecise identification of the harm to which they respond. Because deepfakes are appropriative, they are objectionable irrespective of whether they assert a falsehood or disclose a private fact. However, because the harms of deepfakes are so frequently mischaracterized as standard invasion-of-privacy or defamation injuries, legislators mistakenly attempt to target the problem by regulating "false" imagery. Statutes that regulate pornographic deepfakes by targeting assertions of fact will never address what's most objectionable about them, which is neither falsity nor truth, but their appropriative use of an unconsenting person's photorealistic likeness.

There is a more fundamental reason for the confusion that surrounds anti-deepfakes laws. Anti-deepfakes laws aren't just regulating a different harm from

153. TAKE IT DOWN Act Second Draft.
154. *Id.* § 2(h)(1)(B).
155. *Compare id.* §§ 2(h)(1), (2) *with* TAKE IT DOWN Act First Draft § 2(h)(2).
156. DEL. CODE tit. 10, § 7802 . *See also* IND. CODE § 35-45-4-8(c)(3) ; UTAH CODE § 76-5b-205(1)(a)(ii) .
157. GA. CODE § 16-11-90 .
158. CAL. CIV. CODE § 1708.86 ; COLO. REV. STAT. § 18-7-107 ; GA. CODE § 16-11-90 ; HAW. REV. STAT. § 711-1110.9 ; N.C. Gen. Stat. § 14-190.5A; VA. CODE § 18.2-386.2 ; VT. STAT. tit. 13, § 2606 .

defamation or paradigmatic invasion of privacy; they're also regulating an entirely different subject matter. Defamation and privacy laws regulate materials that assert facts about persons. In other words, they regulate materials that purport to be *records* about persons. But anti-deepfakes laws regulate offensive *depictions* of persons, pure and simple. Whether these depictions purport to disclose facts is beside the point. Regulating images because of *how they depict someone* is dramatically different from regulating images because of *what facts they assert*. Anti-deepfakes laws do the former, while defamation and revenge-porn laws do the latter. This difference places anti-deepfakes laws in a distinct, and constitutionally disfavored, category of regulation from fact-based regimes like defamation and generic invasion of privacy. The discipline of semiotics explains this difference with a precision that has eluded scholars and legislators to date.

## II. Semiotics Helps Us Understand Anti-Deepfakes Laws

Semiotics is the study of "signs."[159] A sign, in turn, "is an object which stands for another to some mind": it is something that communicates meaning.[160] The words in this article are signs; an air raid siren is a sign; a photograph is a sign; a deepfake is a sign. These examples, like all signs, are representations that communicate some meaning to a beholder. The semiotic theory of the 19th-century scholar C.S. Peirce provides an analytical framework for probing the differences between deepfakes and other media and for understanding the subject matter and the harm that the typical anti-deepfakes law regulates.

### A. Peircian Icons, Indices, and Symbols

For Peirce, each sign involves a signifying element (e.g., the written word "book"); an object (e.g., the book that the signifying element represents); and an interpretant ("the understanding that we have of the sign/object relation").[161] Most relevant for present purposes are the trichotomous categories that Peirce propounded to taxonomize signs in terms of their relationship to their objects: icons, indices, and symbols.

---

159. Philipp Strazny, *Semiotics*, ENCYCLOPEDIA OF LINGUISTICS (2005); Barton Beebe, *The Semiotic Analysis of Trademark Law*, 51 UCLA L. REV. 621, 626–29 (2003).

160. JAMES HOOPES, PEIRCE ON SIGNS: WRITINGS ON SEMIOTIC BY CHARLES SANDERS PEIRCE 141 (2014), https://muse.jhu.edu/book/41103 (last visited Jun 3, 2023).

161. Albert Atkin, *Peirce's Theory of Signs*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2013), https://plato.stanford.edu/archives/sum2013/entries/peirce-semiotics/ (last visited Aug 13, 2018).

- An *icon* is a sign that relates to its object through resemblance.[162] This thumbs-up wingding,  , is an iconic representation of a fist with a thumb raised.
- An *index* is a sign that relates to its object through physical or temporal contiguity.[163] A footprint in sand is an indexical sign.[164] Some force—very probably a foot—made contact with the sand and left the foot-shaped depression. We interpret the footprint as a sign that someone walked on the sand.
- A *symbol* is a sign that relates to its object by convention.[165] Both the phrase "thumbs-up!" and the image  are symbols that signify approval. This meaning is arbitrary. In a different linguistic or cultural context, the written words "thumbs-up!", or the sounds of those words spoken, or the thumbs-up gesture, may not be meaningful. Or their arbitrarily assigned meaning may differ: in some cultures, for example, the thumbs-up gesture is obscene.[166]

A sign can function in multiple ways at once. Photographs, for example, signify both iconically and indexically. Photographs record how their subject "really appeared" in the sense that they record photons that bounced off an object and onto a photosensitive surface at a particular moment (and thus are indexical); they also generally "look like" their subjects (and thus are iconic).[167]

---

162. RICHARD J. PARMENTIER, SIGNS IN SOCIETY: STUDIES IN SEMIOTIC ANTHROPOLOGY 4, 17 (1994); *see also* Albert Atkin, *Peirce's Theory of Signs*, *in* THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed., 2013), https://plato.stanford.edu/archives/sum2013/entries/peirce-semiotics/ (last visited Aug 13, 2018).

163. PARMENTIER, *supra* note 162 at 4, 17; *see also* Atkin, *supra* note 2.

164. *See* PARMENTIER, *supra* note 162 at 8.

165. *Id.* at 6.

166. David Anderson et al., *5 Everyday Hand Gestures That Can Get You in Serious Trouble Outside the US*, BUSINESS INSIDER, https://www.businessinsider.com/hand-gestures-offensive-different-countries-2018-6 (last visited Nov 27, 2023).

167. *See, e.g.*, Göran Sonesson, *Visual Signs in the Age of Digital Reproduction*, 19 *in* ENSAYOS SEMIÓTICOS, DOMINIOS, MODELOS Y MIRADAS DESDE EL CRUCE DE LA NATURALEZA Y LA CULTURA. PROCEEDINGS OF THE 6TH INTERNATIONAL CONGRESS OF THE IASS, GUADALAJARA, MEXICO, JULY 13 TO 1073, 3–4 (2000),

Although Peirce contributed much more to semiotics than just his icon-index-symbol trichotomy, it is all we need to clarify sufficiently our understanding of deepfakes and the laws that regulate them. Up until this point, my article has used words like "*documentary* photograph" or "video *recording*" in an effort to differentiate media that *record* perceptual information from media that merely *depict* things pictorially. But I used these words as approximations for concepts that Peircian semiotics allows us to discuss precisely. When I differentiated between "documentary" "records" like photographs on the one hand and "non-documentary" "depictions" like drawings and paintings on the other, what I was really doing was differentiating between indices and icons.

Peirce's theory lets us describe exactly how deepfakes differ from the photographs and videos they resemble, and why these differences may be legally significant. The next two subsections, respectively, explain the semiotic relationship between a photograph or video recording and what it represents, and the distinct semiotic relationship between a deepfake and what it represents.

### B. The Semiotics of Photographs and Video Recordings

Photographs bear an indexical relation to their objects. When film is developed after being exposed to light, the image that results is determined by the manner in which light made contact with the film.[168] In this way, there is a physical relationship between what the camera lens "saw" and what viewers of the photograph "see." A similar relationship holds for digital photographs and videos.[169]

Our understanding that photographs are indexical helps explain why we often ascribe "truth claims" to photographs.[170] We understand photographs to depict reality mechanically, and thus to fulfill a documentary function that drawings and paintings do not.[171] We refer to video evidence as a "smoking gun."[172] A

---

https://www.academia.edu/download/32552334/PhotoPost.pdf (last visited Nov 27, 2023). *See also* Part II.B, *infra*.

168. Kris Paulsen, *The Index and the Interface*, 122 REPRESENTATIONS 83, 86–87 (2013).

169. Although the indexicality of *digital* photographs is a more contested proposition, this article treats both digital and analog photographs as indices. For an overview of the debate, see *id.* at 87-89.

170. Tom Gunning, *What's the Point of an Index? or, Faking Photographs*, 25 NORDICOM REVIEW 39, 42 (2004).

171. *Id.* at 41–42.

172. *See, e.g.*, David Wickert, *'Smoking Gun' Video of Georgia Vote Count is Now Evidence Against Trump*, ATLANTA JOURNAL-CONSTITUTION, Aug. 5, 2023,

smoking gun is an indexical, and thus incriminating, sign that a gun was fired recently.[173] Like a smoking gun, we understand video evidence to be an indexical—indeed, essentially irrefutable—showing that something happened.

However, photographs don't just signify indexically. What makes photographs especially powerful is that they also signify iconically. Photographs don't merely record a phenomenon; they also resemble a contemporaneous visual experience of that phenomenon. In contrast to, say, a seismograph—which is an indexical record of an earthquake, but which doesn't "look" or "feel" like an earthquake in any phenomenologically relevant sense—a photograph actually looks like the pattern of light that it records indexically.

Photographs' indexical properties don't make them immutably truthful. Photographs can be doctored or deceptively composed to misrepresent reality.[174] Mischief results when photographs invite us to misinterpret them (or, in Peircian terms, when photographs produce interpretants that diverge from the photographs' true relationships with their objects). That a photograph communicates a misrepresentation does not mean that the photograph is not an index. Rather, it means that *the photograph is being misinterpreted*. For example, Civil War photographers would move corpses from the locations in which they had fallen and pose them for photographs. The resulting photographs are still indexical signs. They just don't indexically signify the visual appearance of the circumstances of a soldier's death, which is what naïve viewers might expect. Rather, these photographs indexically signify the visual appearance of a scene that reflects the photographer's compositional alterations. Historical methods help us assign the correct interpretant to a photograph that might otherwise mislead us. A historian studying a series of posed Civil War photographs that purported to depict a dead "sharpshooter" concluded, "the type of weapon seen in these photographs was not used by sharpshooters. This particular firearm is seen in a number of [the

---

https://www.ajc.com/politics/smoking-gun-video-now-evidence-against-trump/J6ORVROLMRBPZHK2DYALIZJ624/; Kyle Schnitzer & Georgett Roberts, *Video from Ed Sheeran Concert Is "smoking Gun" in Marvin Gaye Copyright Case*, N.Y. POST, https://nypost.com/2023/04/25/video-from-ed-sheeran-concert-is-smoking-gun-in-marvin-gaye-copyright-case/ (last visited Nov 27, 2023).

173. Smoking gun, WIKIPEDIA (2023), https://en.wikipedia.org/w/index.php?title=Smoking_gun&oldid=1161207239 (last visited Nov 27, 2023); Definition of SMOKING GUN, (2023), https://www.merriam-webster.com/dictionary/smoking+gun (last visited Nov 27, 2023).

174. *See generally* Gunning, *supra* note 170; Paulsen, *supra* note 165; Katrina Geddes, *Ocularcentrism and Deepfakes: Should Seeing Be Believing?*, 31 FORDHAM INTELL. PROP. MEDIA & ENT. LJ 1042 (2020).

photographer's] scenes at Gettysburg and probably was the photographer's prop."[175] The takeaway is that we interpret photographs in light of our knowledge and assumptions about what photographs communicate. Photographs, though indexical, can mislead: the veracity of our interpretations depends on the information we have available to assist us.

What makes photographs special, then, is not so much that they are indices, but that they are indexical icons. After all, every sign is an index of whatever caused it—we just may not find much of interest in what the sign signifies indexically. A paperback book indexically signifies that paper underwent the bookbinding process, but readers are generally much more interested in the *symbolic* signification of the book's text. A doctored photograph isn't an indexical sign of how photons passed through a lens, but it *is* an indexical sign of post-hoc manipulation.[176] This tees up a crucial insight: as important as a sign's actual relationship with its object is the relationship we *think* it has.

Photographs' and video recordings' status as indexical icons—and our corresponding convention of interpreting them as such—explains why the law often affords them singular treatment. Sometimes, indexical status alone makes media special. While police sketches are inadmissible hearsay, authenticated surveillance video is admissible evidence.[177] Indexical evidence of some event, like surveillance footage or a murder weapon, can be admitted in court as substantive "real" evidence, while merely iconic evidence, like a wholly computer-generated recreation of an event, is "demonstrative" evidence.[178] In other contexts, what gives photographs a special legal position is that they are both indexical *and* iconic. Consider child pornography laws. There are many indexical signs of the sexual

175. The Case of the Moved Body, LIBRARY OF CONGRESS, https://www.loc.gov/collections/civil-war-glass-negatives/articles-and-essays/does-the-camera-ever-lie/the-case-of-the-moved-body/ (last visited Apr 16, 2024).

176. Indeed, analysts interpret photographs of Kim Jong-Un not to learn what the dictator "truly" looks like, but rather to ascertain how North Korean propaganda entities choose to photoshop him. Sarah Emerson, *Why Does North Korea Keep Photoshopping Kim Jong-Un's Ears?*, VICE (2017), https://www.vice.com/en/article/ywzjj5/why-does-north-korea-keep-photoshopping-kim-jong-uns-ears (last visited Apr 16, 2024).

177. *Compare, e.g.*, People v. Coffey, 182 N.E.2d 92, 94 (N.Y. 1962) (composite sketch by police artist is "[o]rdinarily" inadmissible hearsay) *with* People v. Patterson, 710 N.E.2d 665, 667 (N.Y. 1999) ("[R]elevant videotapes and technologically generated documentation are ordinarily admissible under standard evidentiary rubrics.").

178. *See, e.g.*, United States v. Rembert, 863 F.2d 1023, 1028 (D.C. Cir. 1988); People v. Jennings, 324 N.W.2d 625, 627 (Mich. App. 1982). I thank Rebecca Wexler for suggesting this point to me.

abuse of minors that these laws do not encompass. It is a traditional practice in a variety of cultures to inspect newlyweds' bedsheets for bloodstains, which supposedly signify a bride's virginity.[179] Despite being an indexical sign of sexual activity, an underage bride's bloody bedsheets are not covered by child pornography laws. Neither are other indices of the sexual abuse of a minor, like the results of a pregnancy or paternity test. And some strictly iconic depictions of minors—like, say, an oil painting or a sculpture—are expressly exempted by the federal child pornography statute.[180] What makes audiovisual child sexual abuse material uniquely sensitive is that typically it both records real-life abuse *and* resembles that activity in the way we deem most morally salient.[181]

### C. The Semiotics of Deepfakes

Deepfakes do not indexically record visual phenomena as photographs do.[182] Sure, deepfakes indexically signify *something*, just like a painting indexically signifies that paint made contact with canvas. What distinguishes deepfakes from photographs is not that deepfakes "aren't indexical," it's that deepfakes don't indexically signify visual phenomena in the way photographs do. Instead, what deepfakes signify indexically is the outcome of complicated statistical analyses in an AI model.[183] If image-generating AI produces a photorealistic picture of a white man in a lab coat when prompted to depict a "doctor," that generated image may indexically signify certain patterns in the data that trained the AI. What that image doesn't signify indexically, however, is that particular photons passed through a particular lens at a particular moment in time to create that particular image. Unlike photographs, which are indexical icons, deepfakes are *icons* of indexical icons.[184] Anti-deepfakes laws' photorealism requirement is simply a requirement that a covered deepfake must be a convincing icon of a photograph.[185]

Because deepfakes are icons, not indexical records of a discrete, observable event—which is what we interpret photographs and video recordings to be—most

---

179. Nicola Heath, *The Historic Tradition of Wedding Night-Virginity Testing*, SBS VOICES (Jan. 16, 2018), https://www.sbs.com.au/voices/article/the-historic-tradition-of-wedding-night-virginity-testing/cknarl44w (last visited Jan 28, 2024).

180. *See* 18 U.S.C. § 2256(8)(B), (11).

181. *See* Part III.A.2.a), *infra*.

182. *See* Rebecca Uliasz, *On the Truth Claims of Deepfakes: Indexing Images and Semantic Forensics*, THE JOURNAL OF MEDIA ART STUDY AND THEORY, 69 (2021), https://www.mast-journal.org/2022vol3no14 (last visited Jun 5, 2023).

183. *See id.*

184. *Cf. id.* at 68.

185. *See* Part I.A.7, *supra*.

of the rationales that justify regulating sexually explicit photographs do not apply to deepfakes. Photo- or videographic revenge porn presents a conventional privacy issue because it indexically documents a victim's private actions and the actual appearance of private body parts. Similarly, photo- or videographic CSAM can be regulated on the ground that its creation is "intrinsically related to the sexual abuse of children": indexical records of abuse require that abuse take place.[186] Correspondingly, deceptive deepfakes present a conventional defamation issue because we interpret them to be indexical records of perceptual facts, even though they are merely icons.[187] Both revenge porn and deceptive deepfakes assert factual propositions because we understand both to be indices and we understand photographic indices to make truth claims.

By contrast, many deepfakes are, in context, obviously fake. They can't reasonably be understood to make a truth claim about how photons struck a lens at a particular moment in time and space. The many grounds that justify regulating indexical images, or deceptive simulacra, are unavailable to regulate obvious deepfakes, which are mere icons. To regulate non-deceptive deepfakes, we must acknowledge that we are not regulating the dissemination of true or false factual propositions, as defamation and privacy law do. Nor are we regulating the real-life conduct that the imagery appears to depict, as we are when we regulate indexical CSAM. Instead, we are regulating something closer to flag burning or blasphemous drawings: we are regulating outrageous uses of icons *per se*.

One counterargument might be that no matter how conspicuously disclaimed, viewers simply cannot regard deepfakes as anything other than assertions of fact.[188] This argument treats deepfake viewers like the apocryphal silent-film audience that fled the theater in fear, believing a train on screen was

---

186. *See Ferber*, 458 U.S. at 747.

187. The term "dicentization" describes the interpretation of an icon as an index. Christopher Ball, *On Dicentization*, 24 J. LINGUISTIC ANTHROPOL. 151, 152 (2014).

188. *See* Abigail George, *Defamation in the Time of Deepfakes*, 37–38 & n.239 (2024), https://papers.ssrn.com/abstract=4719803 (last visited May 28, 2024). Distinct from the argument that photorealistic depictions are always *understood* as assertions of fact, George also argues that "hyperrealistic" styles are always *intended* as statements of fact because they "aim at truth and assert knowledge" even when they carry disclaimers of falsity. *See id.* at 37. This, I believe, is mistaken. Representational styles are neither true nor false; media like the *Lion King* remake can be both fictional and photorealistic without being false. *See supra*, text accompanying notes 48-50, 145. To quote Searle, the novelist's "speech act . . . does not commit her to the possession of evidence" and involves "no commitment to the truth of the proposition." John R. Searle, *The Logical Status of Fictional Discourse*, 6 NEW LITERARY HISTORY 319, 323 (1975).

actually barreling into their seats.[189] Instead of positing that we believe anything that looks like a train barreling at us to be a train barreling at us, this argument posits that we reflexively assume that anything that looks like a photograph is a photograph.

But this counterargument collapses when we consider the differing regulation of disclaimers in pornographic deepfakes and electioneering deepfakes. States that treat disclaimers as irrelevant in pornographic deepfakes have also passed laws regulating election-related deepfakes, and these laws expressly *exempt* electioneering deepfakes that contain disclaimers.[190] That state laws treat disclaimers as effective in electioneering media but ignore them in sexual media shows that laws prohibiting pornographic deepfakes are regulating something beyond deception. Moreover, a reasonable observer in 2024 *doesn't* interpret all photorealistic media as photographic. The photorealistic *Lion King* remake didn't convince the world that Disney had taught Serengeti beasts to sing and dance.[191] Deepfakes' critics acknowledge that "as the public becomes more educated about the threats posed by deep fakes," the public will quite sensibly doubt that photorealistic media is authentically photographic.[192] If we read anti-deepfakes laws as establishing an irrebuttable presumption that it is always reasonable to interpret a photorealistic image as a photograph, then we are reading these laws to codify media illiteracy unmoored from present-day interpretive practice.[193]

So no, the premise of anti-deepfakes laws isn't, and shouldn't be, that deepfakes always assert facts in the same manner that indexical images do. Rather, the premise is that deepfakes are harmful even when they cannot reasonably be understood as documentary fact. Anti-deepfakes laws self-consciously regulate icons *qua* icons, and they target pornographic uses precisely because of what those uses communicate. This means that anti-deepfakes laws are in the business of regulating outrageous expressions of opinion, rather than true or false assertions of fact. And this, in turn, might suggest that these laws are unconstitutional. It turns

189. Eric Grundhauser, *Did a Silent Film About a Train Really Cause Audiences to Stampede?*, SLATE, Jan. 2017, https://slate.com/human-interest/2017/01/the-silent-film-that-caused-audiences-to-stampede-from-the-theater.html (last visited Dec 12, 2023).

190. *See, e.g.*, CAL. ELEC. CODE § 20010(b) ; WASH. REV. CODE § 42.62.020(4) . *See also* note 45, *supra*.

191. *See* Part I.A.7, *supra*.

192. Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1785–86 (2019).

193. Cf. Geddes, *supra* note 171 at 1073 (questioning "the social utility of ocularcentrism").

out, however, that American law has a long and diverse history of content-based regulation of outrageous iconography, which Part III recounts.

## III. Iconic Signs in American Law

The subsections of this Part examine disparate American legal doctrines, each of which illuminates a distinct aspect of the typical anti-deepfakes law. The law of trademark dilution tells us that it has not been held unconstitutional to regulate non-deceptive uses of images because of their putative tendency to distort emotional attitudes towards the referents of those images. CSAM caselaw tells us both that courts often conflate indices and icons and that, at least in the context of "morphed" images, criminal prohibitions on non-deceptive icons have been upheld as constitutional. Attitudes towards flag and effigy burning confirm both that desecrating a symbol is a distinct, and more offensive, communication than disparaging its referent in words, and that this distinction has been legally dispositive. The legal status of written sexual fantasies tells us that fantastical images can be more "real" than fantastical words, and thus illuminates anti-deepfakes laws' realism requirements.

These doctrines show both that anti-deepfakes laws address time-honored concerns and that bans on non-deceptive, disparaging uses of icons are precedented. But they also reveal the constitutional challenges that anti-deepfakes laws present. Anti-deepfakes laws will force courts to consider head-on whether certain uses of icons can be proscribed simply because they are outrageous. History suggests a deep human impulse to enact such proscriptions, and some still appear in American law. But courts have rebuffed prior constitutional challenges by mischaracterizing icons as indices and/or by relying on categorical First Amendment exclusions that do not encompass deepfakes. These offramps will be unavailable for anti-deepfakes laws. Courts will have to choose between, on one hand, our longstanding and legally enshrined impulses to regulate outrageous iconography; and, on the other, a line of First Amendment caselaw that would seem to disfavor precisely such regulation.

### A. Contemporary Bans on Outrageous Iconography

#### 1. Trademark Dilution

The doctrine of trademark dilution by "tarnishment" illuminates why anti-deepfakes laws focus on pornographic media.[194] Semiotically speaking, trademarks

194. *See* Part I.A.1, *supra*.

function as *symbols* rather than icons: they are arbitrary signs that designate a provider of goods or services.[195] But just as an iconic depiction of a person represents that person, a trademark symbol represents the brand to which it corresponds. And just as anti-deepfakes laws prohibit degrading uses of images irrespective of their deceptiveness, dilution law allows brands to prevent uses of their marks that associate the marks with negative connotations—even if those uses do not convey false information or confuse consumers.[196]

Trademark tarnishment occurs when a mark is used in an unsavory context, like when a pornographic film features clothing that resembles a plaintiff's cheerleading uniform.[197] Although run-of-the-mill trademark infringement requires a plaintiff to show a likelihood that consumers will be confused as to the source or origin of goods, dilution is actionable even with no likelihood of confusion.[198] Dilution law lets brands prohibit uses of symbols that might imbue those symbols with undesirable emotional connotations.

As Rebecca Tushnet notes, dilution law goes beyond defamation because it gives brands control over the social and emotional atmospherics of their marks, not just assertions of fact.[199] So do anti-deepfakes laws.[200] Just as dilution law finds cognizable harm in the absence of confusion, the law of nonconsensual, simulated pornography can find cognizable harm even when the person who altered the images is the only person who has seen them or when the imagery is so obviously fake that no reasonable observer could regard it as documenting fact.[201] And while federal dilution law covers only commercial uses, anti-deepfakes laws cover noncommercial uses, too.[202] The Supreme Court has never held dilution law to be constitutional, and in recent years it has struck down bans on registering "disparag[ing]" and "immoral[] or scandalous" trademarks as impermissible

---

195. *See* Beebe, *supra* note 156 at 637 & n.85.

196. *See* Rebecca Tushnet, *More than a Feeling: Emotion and the First Amendment*, 127 HARV. L. REV. 2392, 2398 (2013).

197. *See* 4 MCCARTHY ON TRADEMARKS AND UNFAIR COMPETITION § 24:67 (5th ed.); Dallas Cowboys Cheerleaders, Inc. v. Pussycat Cinema, Ltd., 467 F. Supp. 366, 371, 377 (S.D.N.Y.), *aff'd*, 604 F.2d 200 (2d Cir. 1979).

198. *See, e.g.*, 15 U.S.C.A. § 1125(c)(1); NY GEN. BUS. L. § 360-L (2022).

199. Tushnet, *supra* note 183 at 2392; *see also* Jessica Litman, *Breakfast with Batman: The Public Interest in the Advertising Age*, 108 YALE L.J. 1717, 1728 (1999).

200. Indeed, seemingly by coincidence, a recent paper uses the exact language of "tarnish[ment]" to describe how deepfakes harm the persons they depict. George, *supra* note 185 at 33.

201. *See* DEFIANCE Act of 2024 § 3(b)(1)(A) (permitting relief for production of a deepfake); Part I.A.6, *supra*.

202. 15 U.S.C. § 1125(c)(3).

viewpoint-based discrimination.[203] However, even after decades of litigation, dilution statutes have not been struck down as unconstitutional.[204]

### 2. Child Sexual Abuse Material (CSAM)

The law of CSAM is a tour of all the same semiotic issues and imprecise reasoning that deepfakes have elicited. CSAM jurisprudence reveals that courts have upheld the constitutionality of regimes that regulate non-deceptive icons *qua* icons. Peirce's semiotics maps seamlessly onto the law of CSAM. The federal definition of "child pornography" encompasses (1) images whose production "involve[d] the use of a minor engaging in sexually explicit conduct," (2) "digital . . . or computer-generated image[s] that . . . [are] indistinguishable from, that of a minor engaging in sexually explicit conduct," and (3) images that were "created, adapted, or modified to appear that an identifiable minor is engaging in sexually explicit conduct."[205] In turn, jurisprudence identifies three distinct categories of CSAM. The first is indexical imagery whose manufacture necessitates abusing children or photographing intimate parts of their bodies. The second is purely iconic "virtual" CSAM, which is constitutionally protected speech precisely because it is not indexical.[206] The third category comprises non-sexual photographs of children that have been "morphed" to depict sexual conduct.

#### *a) Indexical Images of Child Sexual Abuse*

The rationale for restricting the first category of images—those whose production necessarily involves children's participation—is for obvious reasons the

---

203. Iancu v. Brunetti, 588 U.S. 388, 390 (2019); *Tam*, 582 U.S. at 223.

204. The constitutionality of dilution by tarnishment is currently under attack. In May 2024, the federal government agreed to intervene in a district-court proceeding "for the limited purpose of defending the constitutionality" of dilution by tarnishment. *See* Notice of Intervention by the United States of America and Unopposed Motion to Set Briefing Schedule at 1, Dkt. No. 361, VIP Prods. LLC v. Jack Daniel's Props., Inc., No. 2:14-cv-02057 (D. Ariz. May 8, 2024), *remanded by* Jack Daniel's Properties, Inc. v. VIP Prods. LLC, 599 U.S. 140 (2023). For a skeptical view of trademark dilution's constitutionality, see Mark A. Lemley & Rebecca Tushnet, *First Amendment Neglect in Supreme Court Intellectual Property Cases*, 2023 S. CT. REV. 85, 97–109 (2024).

205. 18 U.S.C. § 2256(8).

206. The Supreme Court has not yet considered "virtual" child pornography falling under the "indistinguishable from" prong of the statute. In a 2002 case, the Court held that earlier statutory language—which encompassed media that "appears to be[] of a minor engaging in sexually explicit conduct"—was unconstitutionally overbroad. Ashcroft v. Free Speech Coal., 535 U.S. 234, 241, 258 (2002). *See also* Part III.A.2.b), *infra*.

least controversial. In *New York v. Ferber*, the Supreme Court held that the First Amendment did not bar a bookstore owner's conviction under a New York statute that prohibited "promoting a sexual performance by a child."[207] The defendant, Ferber, sold films "depicting young boys masturbating" to an undercover police officer, and the jury had found that the images in question were not obscene.[208] There was apparently no dispute that the materials Ferber sold were indexical, videographic depictions.[209] The Court held that material of this sort is unprotected by the First Amendment, even if not obscene.[210] It noted the "surpassing importance" of protecting children from sexual abuse, and it reasoned that the distribution of media recording minors engaged in sexual activity "is intrinsically related to the sexual abuse of children in at least two ways": first, the images "are a permanent record of the children's participation and the harm to the child is exacerbated by their circulation;" and second, controlling the production of child pornography requires controlling its distribution network.[211]

The harms that *Ferber* identified depend on CSAM being an indexical sign. Most importantly, *Ferber* identified a harm "intrinsic[]" to the creation of child pornography—it requires that children engage in real-life sexual conduct—and it concluded that regulating its distribution was a proper means of targeting that harm.[212] Second, the Court called the images "a permanent *record* of the child's participation," the circulation of which "may haunt [a victim] in future years."[213]

---

207. New York v. Ferber, 458 U.S. 747, 751-52 (1982); N.Y. PENAL LAW § 263.15 (1977) (McKinney).

208. *See Ferber*, 458 U.S. at 752, 759, 764-65. The New York statute, read literally, might have been broad enough to encompass altered images that depict sexual conduct in which no child actually participated, because its definition of "sexual conduct" included "simulated" depictions that "create[] the appearance of such conduct." N.Y. PENAL LAW § 263.00(1) (McKinney) (1977); *id.* §§ (3), (5). But the filings leave little doubt that the films in *Ferber* were actual records of children's real-life conduct, and subsequent interpretations of the statute imply that a conviction could not be based solely on images altered to depict sexual conduct in which children did not actually participate. *See* People v. Foley, 692 N.Y.S.2d 248, 256 (App. Div. 4th Dep't 1999), *aff'd*, 731 N.E.2d 123 (N.Y. 2000).

209. *Cf. Ferber*, 458 U.S. at 764-65 ("[D]istribution of . . . depictions of sexual conduct, not otherwise obscene, which do not involve live performance or photographic or other visual reproduction of live performances, retains First Amendment protection.").

210. *Id.* at 752, 765.

211. *Id.* at 757-60.

212. *Id.* at 759-60.

213. *Id.* at 759 & n.10 (quoting Shouvlin, *Preventing the Sexual Exploitation of Children: A Model Act*, 17 WAKE FOREST L. REV. 535, 545 (1981)) (emphasis added).

Because the images record abuse that children suffered before the camera, their circulation "violates 'the individual interest in avoiding disclosure of personal matters.'"[214] This framing depends on the images' status as indexical records of real-life events. In a subsequent case, *Osborne v. Ohio*, the Court relied on these rationales to uphold bans on private possession of child pornography.[215] *Ferber* and *Osborne* teach that indexical CSAM's documentary properties warrant treatment as a distinct category of unprotected speech under the First Amendment.

*b) Iconic Images of Child Sexual Abuse*

Just as *Ferber* establishes that indexical images of child abuse fall outside the First Amendment, it is equally clear that images with *no* indexical relationship to an actual child cannot be criminalized categorically as child pornography. In 1996, Congress expanded the definition of child pornography to include "any visual depiction . . . [that] appears to be[] of a minor engaging in sexually explicit conduct."[216] In *Ashcroft v. Free Speech Coalition*, the Supreme Court held that this prohibition of so-called "virtual" child pornography was unconstitutionally overbroad because it covered First Amendment-protected expression.[217] The lack of an indexical relationship between virtual child pornography and the content it depicts was central to the Court's reasoning:

> In contrast to the speech in *Ferber,* speech that itself is the record of sexual abuse, the [challenged law] prohibits speech that records no crime and creates no victims by its production. Virtual child pornography is not "intrinsically related" to the sexual abuse of children, as were the materials in *Ferber*.[218]

Another way of expressing the Court's conclusion that "[v]irtual child pornography is not 'intrinsically related' to the sexual abuse of children" is to say that virtual child pornography is merely an iconic depiction of child abuse, and not an indexical record. It is for this reason that the Court could state that virtual child pornography "creates no victims by its production."[219] After *Free Speech Coalition*, Congress amended the "appears to be" language to instead ban visual depictions

214. *Id.* at 759 n.10 (quoting Whalen v. Roe, 429 U.S. 589, 599 (1977)).
215. 495 U.S. at 110.
216. 18 U.S.C. § 2256(8)(B) (1996).
217. Ashcroft v. Free Speech Coal., 535 U.S. 234, 258 (2002).
218. *Id.* at 250 (citations omitted).
219. *See id.*

"indistinguishable from[] that of a minor engaging in sexually explicit conduct."[220] Although the constitutionality of this revised language remains untested, commentators tend to believe that it is unconstitutional for the same reasons as the "appears to be" language that *Free Speech Coalition* struck down.[221]

*c) Manipulated Indices: "Morphed" Images*

*Ashcroft v. Free Speech Coalition* left an important lacuna. The lawsuit did not challenge § 2256(8)(C) of the federal statute, which included in the definition of child pornography a "visual depiction . . . created, adapted, or modified to appear that an identifiable minor is engaging in sexually explicit conduct."[222] About these so-called "morphed" images, the Court said only the following:

> Although morphed images may fall within the definition of virtual child pornography, they implicate the interests of real children and are in that sense closer to the images in *Ferber*. Respondents do not challenge this provision, and we do not consider it.[223]

The Court's brief discussion of § 2256(8)(C) seemed to acknowledge that the ontology of morphed images is more complicated than that of indexical records of abuse or strictly virtual icons. On one hand, unlike the videos in *Ferber*, morphed images are not records of physical abuse. On the other hand, they are images "of" "real children" in a sense that purely virtual images are not, because the starting point for a morphed image is an indexical image of a real-life child—albeit typically a non-sexual one.[224]

Morphed images are criminalized in a way that reflects this ambivalent status. Even the Department of Justice has stated that "the production of a morphed image of child pornography is not as serious a crime as the production of genuine child pornography."[225] Unlike indexical CSAM, federal law does not

---

220. PUB. L. 108–21, 117 STAT. 650, 678 (Apr. 30, 2003); 18 U.S.C. § 2256(8)(B).

221. RIANA PFEFFERKORN, *Addressing Computer-Generated Child Sex Abuse Imagery: Legal Framework and Policy Implications*, 6–7 (2024), https://www.lawfaremedia.org/article/addressing-computer-generated-child-sex-abuse-imagery-legal-framework-and-policy-implications.

222. 18 U.S.C. § 2256(8)(C); *Free Speech Coalition*, 535 U.S. at 242.

223. *Free Speech Coalition*, 535 U.S. at 252.

224. *See id.*

225. UNITED STATES SENTENCING COMMISSION, *The History of the Child Pornography Guidelines*, 50–51 (2009), https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-projects-and-surveys/sex-offenses/20091030_History_Child_Pornography_Guidelines.pdf.

prohibit simple possession of morphed images or creation without intent to distribute.[226] Rather, the statute prohibits receipt, distribution, and production with intent to distribute.[227] However, several state anti-deepfakes laws prohibit simple possession of pornographic deepfakes depicting a minor.[228]

Although the Supreme Court has not considered the constitutionality of the federal criminalization of morphed images, federal courts of appeals have universally upheld it.[229] The federal courts have, however, split as to their rationale. Three courts of appeals have held morphed images to be categorically unprotected speech, but the Eighth Circuit has assumed that morphed images, unlike indexical CSAM, are protected speech.[230] The earliest and narrowest decision came from the Eighth Circuit in 2005 in *United States v. Bach*. *Bach* affirmed a conviction based on § 2256(8)(C) where the defendant had received an image of a minor exhibiting his genitals, onto which the face of a different minor had been superimposed.[231] The Eighth Circuit reasoned that, because the nude figure in the photo was a minor, "[u]nlike the virtual pornography protected by the Supreme Court in *Free Speech Coalition*, the picture . . . implicates the interests of a real child and does record a crime."[232] The court was careful to note, however, that "[t]his is not the typical morphing case," because the photo Bach received did not involve merely an "innocent picture of a child" but instead incorporated a photograph of the "lasciviously posed body . . . of" a second child.[233] For this reason, *Bach* held that "[a]lthough there may well be instances in which the application of § 2256(8)(C) violates the First Amendment, this is not such a case."[234] The court stated that the image "involves the type of harm which can constitutionally be prosecuted under *Free Speech Coalition* and *Ferber*," thereby suggesting that it was holding it categorically excluded from First Amendment protection.[235]

---

226. 18 U.S.C. § 2252A(a)(7).
227. 18 U.S.C. §§ 2252A(a)(2), (a)(7).
228. *See* note 24, *supra*.
229. The New Hampshire Supreme Court, however, held a state statute unconstitutional as applied to the private possession of images that superimposed minors' faces onto adult bodies engaged in sex acts. State v. Zidel, 940 A.2d 255, 256, 265 (N.H. 2008).
230. *Compare* United States v. Mecham, 950 F.3d 257, 267 (5th Cir. 2020); Doe v. Boland, 698 F.3d 877, 884 (6th Cir. 2012); United States v. Hotaling, 634 F.3d 725, 730 (2d Cir. 2011) *with* United States v. Anderson, 759 F.3d 891, 895 (8th Cir. 2014).
231. United States v. Bach, 400 F.3d 622, 625, 631-32 (8th Cir. 2005).
232. *Id.* at 632.
233. *Id.* at 632.
234. *Id.*
235. *Id.*

The Eight Circuit revisited *Bach* nine years later in *United States v. Anderson*, in which a defendant had been convicted for superimposing a minor's face onto a photograph of *adults* engaged in sexual conduct.[236] Distinguishing *Bach*, the court in *Anderson* observed, "[n]o minor was sexually abused in the production of Anderson's image. . . . [T]his difference is significant enough to distinguish Anderson's image from the unprotected speech in *Bach*."[237] Instead of holding the morphed image to be categorically unprotected by the First Amendment, *Anderson* treated the image as protected speech and held instead that the prohibition on morphed images satisfied strict scrutiny.

For all the care *Anderson* took in distinguishing *Bach*—and all the care *Bach* took in distinguishing *Free Speech Coalition*—*Anderson*'s strict scrutiny analysis was slapdash. The court described the government's compelling interest thusly: "[M]orphed images are like traditional child pornography in that they are records of the harmful sexual exploitation of children. The children, who are identifiable in the images, are violated by being falsely portrayed as engaging in sexual activity."[238] As for narrow tailoring, the court wrote, "the harm a child suffers from appearing as the purported subject of pornography in a digital image that is distributed via the Internet can implicate a compelling government interest regardless of the image's verisimilitude or the initial size of its audience."[239]

*Anderson*'s compelling-interest analysis is incoherent. In stating that "morphed images are like traditional child pornography in that they are records of the harmful sexual exploitation of children," *Anderson* gets the ontology of morphed images exactly wrong. Morphed images are *unlike* traditional child pornography precisely because they are *not* records of the harmful sexual exploitation of children.[240] To call an image a "record" is to suggest that it indexically documents an event. This is how *Bach* used the term when it observed that the picture at issue in that case, an indexical image of a child's genitalia, "*record[s]* a crime."[241] Moreover, in the next sentence, *Anderson* acknowledges that morphed images aren't records at all: rather than document an event that occurred, they "*falsely* portray[]"

---

236. United States v. Anderson, 759 F.3d 891, 893, 895 (8th Cir. 2014).
237. *Id.* at 859.
238. *Anderson*, 759 F.3d at 896 (quoting Shoemaker v. Taylor, 730 F.3d 778, 786 (9th Cir. 2013)).
239. *Id.*
240. Of course, this is only true of what *Bach* called "typical morphing case[s]," which involve non-sexual photographs of children. 400 F.3d at 632.
241. *Id.* (emphasis added).

identifiable minors "as engaging in sexual activity."[242] The harm that *Anderson* is describing is essentially a grievous form of libel.[243] If spreading such libel about a child is "sexual exploitation," then morphed images perhaps *constitute* the "sexual exploitation of children," but in no events do they *record* the sexual exploitation of children.

Moreover, even if *Anderson*'s compelling-interest analysis were coherent, it would still be irreconcilable with the narrow-tailoring analysis that follows it. The harm that the court identified to support the government's compelling interest is the "false[] portray[al] [of victims] as engaging in sexual activity."[244] But just sentences later, in its narrow-tailoring analysis, the court stated that "the harm a child suffers from appearing as the purported subject of pornography in a digital image . . . can implicate a compelling government interest *regardless of the image's verisimilitude*."[245] This assertion belies the court's earlier characterization of the harm of morphed images. If the harm is truly that the images "falsely portray" minors as having engaged in sexual activity, then verisimilitude matters. If expression doesn't purport to assert facts, then it might be fictional, but it isn't false.[246] In the defamation context, if speech cannot "reasonably be understood as describing actual facts," then it cannot be actionably "false."[247] But the charge on which Anderson was convicted did not require a finding that the morphed images would have been perceived as documentations of fact, and nothing in the statute suggests that a defendant could avoid liability simply by adding a disclaimer stating that an image has been morphed. The harm *Anderson* identifies is not a *false* portrayal of an identifiable child engaged in sexual activity, but a portrayal of any sort. The problem is not indices, or even icons that deceptively resemble indices, but icons pure and simple.

Courts hold morphed images categorically unprotected by ignoring the differences between indices and icons. In *United States v. Hotaling*, the Second Circuit affirmed a defendant's conviction under §§ 2252A(a)(5)(B) and 2256(8)(C) for superimposing non-pornographic photographs of minors over the heads of adults photographed in sexually explicit circumstances. Citing *Ferber* and

---

242. *Anderson*, 759 F.3d at 896 (emphasis added).
243. *See* Amy Adler, *Inverting the First Amendment*, 149 U. PA. L. REV. 921, 990 & n.312 (2001).
244. *Anderson*, 759 F.3d at 896 (quoting Shoemaker v. Taylor, 730 F.3d 778, 786 (9th Cir. 2013)).
245. *Id.* (emphasis added).
246. *See* Part I.B.4, *supra*.
247. *See* Pring v. Penthouse Int'l, Ltd., 695 F.2d 438, 440 (10th Cir. 1982).

quoting *Osborne*—cases that both involved indexical records of minors in sexually explicit positions[248]—the Second Circuit asserted,

> the Supreme Court recognized that minors are harmed not only during the creation of child pornography, but are also haunted for years by the knowledge of its continued circulation. These emotional and reputational harms are severe enough to render laws criminalizing the possession of child pornography constitutional in the interest of "stamping out this vice at all levels in the distribution chain."[249]

Then, citing *Free Speech Coalition*'s discussion of *indexical* child pornography, *Hotaling* stated that "the Supreme Court has made it clear that the harm begins when the images are created."[250] *Hotaling* thus held that morphed images "[are] not protected expressive speech under the First Amendment."[251]

*Hotaling* made three significant analytical moves. First, it extended *Ferber* and *Osborne* by reasoning that the "emotional and reputational harms" of the circulation of child pornography justify criminal prohibitions, even when the media are not indexical depictions of the real-life sexual abuse of children or of children's actual bodies. Neither *Ferber* nor *Osborne* suggested that the "haunting" rationale applied to media that did not indexically record minors nude or engaged in sexual conduct. And even if the "haunting" rationale does apply to iconic media, neither *Ferber* nor *Osborne* held that "haunting" alone was a sufficient basis for holding child pornography laws constitutional.[252]

Second, *Hotaling* formulated the constitutional test to match *Free Speech Coalition*'s *dicta* about morphed images. *Hotaling* cited *Free Speech Coalition*'s holding that the virtual child pornography ban was unconstitutional because "the child-protection rationale for speech restriction does not apply to materials produced without children."[253] With this citation as support, *Hotaling* stated that to evaluate whether imagery is protected speech, "[t]he underlying inquiry is whether an image of child pornography implicates the interests of an actual

---

248. *See Osborne*, 495 U.S. at 107 n.1.
249. United States v. Hotaling, 634 F.3d 725, 728-29 (2d Cir. 2011) (citations and alteration marks omitted) (quoting *Osborne*, 495 U.S. 109-11).
250. *Id.* at 730 (citing *Free Speech Coalition*, 535 U.S. at 254).
251. *Id.* at 726-27.
252. *Ferber*, 458 U.S. at 759; *Osborne*, 495 U.S. at 109.
253. *Hotaling*, 634 F.3d at 729 (quoting United States v. Williams, 553 U.S. 285, 289 (2008); *Free Speech Coalition*, 535 U.S. at 258).

minor."[254] It then quoted *Free Speech Coalition*'s *dictum* that "morphed images . . . implicate the interests of real children and are in that sense closer to the images in *Ferber*" than purely virtual images are.[255] *Hotaling* concluded that "the interests of actual minors are implicated when their faces are used in creating morphed images that make it appear that they are performing sexually explicit acts" and thus that such images "are not protected expressive speech under the First Amendment."[256]

*Hotaling* reframed the constitutional inquiry to fit its purposes. In stating that the constitutional test is "whether an image . . . implicates the interests of an actual minor," *Hotaling* did not rely on a Supreme Court case that actually so held. Rather, it relied on *Free Speech Coalition*'s holding that media that did *not* implicate the interests of an actual child *was* protected speech.[257] *Hotaling*'s formulation of the test is the inverse of *Free Speech Coalition*'s: if material *does* implicate the interests of a child, the First Amendment *doesn't* protect it. This holding extends *Free Speech Coalition* because, as a matter of logic, a proposition does not entail its inverse.

More importantly, *Hotaling* exploited the ambiguity that it had baked into its own formulation of the First Amendment test. *Free Speech Coalition* did not hold that "implicating" *any* interest of a minor necessarily relegates an image to the constitutionally unprotected category of child pornography. Rather, its focus was on distinguishing iconic imagery from indexical imagery—or, in its words, distinguishing virtual pornography from material that "caused [harm] to its child *participants*."[258] The images in *Ferber* implicated the interests of actual minor "participants": they were documentary records of sexual conduct by minors, and their circulation would republish private matters and trigger haunting memories. The images in *Hotaling*, on the other hand, were simulacra of abuse; their production did not involve minor "participants" in the sense that *Ferber* used the word. The circulation of the images in *Hotaling* would expose children to the "risk of reputational harm and . . . the psychological harm of knowing that their images were exploited," but their creation did not necessitate the abuse of a minor "participant."[259]

254. *Hotaling*, 634 F.3d at 729.
255. 535 U.S. at 242.
256. *Hotaling*, 634 F.3d at 729-30.
257. *See Free Speech Coalition*, 535 U.S. at 254-56.
258. *See id.* at 249 (emphasis added).
259. *Hotaling*, 634 F.3d at 730.

*Hotaling*'s third move was to equate the harms of morphed images with the harms of indexical records of child abuse. Recall that *Hotaling* relied on "haunting," which is premised upon the harms of images' circulation, rather than of their creation.[260] *Hotaling* also asserted, "the Supreme Court has made it clear that the harm begins when the images are created," citing a page of *Free Speech Coalition* that discussed indexical records of child abuse, not morphed images.[261] *Free Speech Coalition* did not "make clear" that the harm of *morphed* images was inflicted at the moment of their creation; the Court expressly "d[id] not consider" the ban on morphed images.[262] Even federal law does not outlaw morphed images from the moment of creation.[263] The "haunting" rationale "implies a return of a previous experience" and thus, while it may be grounds to regulate indices of child abuse, it is a flimsy basis for regulating icons.[264] Nonetheless, morphed-images caselaw shows that it is constitutional to prohibit some non-deceptive, noncommercial, outrageous uses of icons *per se*.

### 3. Morphed CSAM, Deepfakes, and the First Amendment

Judicial reasoning about morphed images is so tortured because Supreme Court precedents put the appellate courts in an awkward position. *Free Speech Coalition* avoided adjudicating whether morphed images were protected speech. Eight years later, in *United States v. Stevens*, the Court held that depictions of animal cruelty were not categorically unprotected speech. *Stevens* acknowledged that "the First Amendment has permitted restrictions upon the content of speech in a few limited areas . . . . including obscenity, defamation, fraud, incitement, and speech integral to criminal conduct," but it rejected "a freewheeling authority to declare new categories of speech outside the scope of the First Amendment."[265] Soon after, the Court relied on *Stevens* to refuse categorical First Amendment exemptions for violent video games and false statements.[266] Although *Stevens* hedged that there may remain "some categories of speech that have been historically

260. Carissa Hessick, *The Limits of Child Pornography*, 89 IND. L. J. 1437, 1477–78 (2014).
261. *Hotaling*, 634 F.3d at 730 (citing *Free Speech Coalition*, 535 U.S. at 254).
262. *Free Speech Coalition*, 535 U.S. at 242.
263. *See* notes 225-227, *supra*.
264. Adler, *supra* note 240 at 990. *See also Osborne*, 495 U.S. at 145 n.18 (Brennan, J., dissenting).
265. United States v. Stevens, 559 U.S. 460, 468-69, 72 (2010) (citations, quotation marks, and alteration marks omitted).
266. Brown v. Ent. Merchants Ass'n, 564 U.S. 786, 792 (2011); United States v. Alvarez, 567 U.S. 709, 722 (2012).

unprotected, but have not yet been specifically identified or discussed as such in our case law," lower courts read it to suggest an end to novel categorical exclusions from the First Amendment.[267] For example, every state supreme court to consider the constitutionality of a revenge porn statute has, citing *Stevens*, declined to hold revenge porn categorically unprotected speech.[268]

*Free Speech Coalition* and *Stevens* posed a dilemma for lower courts adjudicating constitutional challenges to morphed-images laws. The expedient option was to classify morphed images as unprotected "child pornography" and ignore that they lack the indexical link to real-life sexual abuse that *Ferber*, *Free Speech Coalition*, and *Stevens* all presented as the defining justification for child pornography's categorical exclusion from First Amendment protection.[269] The alternative was to subject the bans to strict scrutiny.

Ultimately, three circuits chose the expediency of categorical exclusion, and the Eighth Circuit contorted itself to hold that morphed-images bans satisfied strict scrutiny.[270] No matter their approach, these courts relied on justifications for banning indexical images without acknowledging that morphed images are icons. The courts thereby managed to sustain the practice of banning outrageous iconography *per se* while purporting to adhere to *Free Speech Coalition* and *Stevens*.

A morphed-images case from the Fifth Circuit, *United States v. Mecham*, contains an instructive First Amendment analysis. *Mecham* justified holding morphed images categorically unprotected by observing that the federal definition of "child pornography" includes not just imagery that records criminal abuse, but also imagery that shows a "lascivious exhibition" of a minor's genitals, like images taken surreptitiously with a hidden camera.[271] True enough. But *Mecham* overlooked a more persuasive distinction between morphed images and indexical child pornography, which is that the latter authentically records private facts about a person and the former does not. Even if it does not document physical abuse, indexical child pornography is objectionable for the same reason that revenge porn

267. 559 U.S. at 472; *see* State v. Katz, 179 N.E.3d 431, 453 (Ind. 2022).

268. *Katz*, 179 N.E.3d at 453; *Casillas*, 952 N.W.2d at 637-38; *Austin*, 155 N.E.3d at 455; *VanBuren*, 214 A.3d at 807. *See also* Ex parte Fairchild-Porche, 638 S.W.3d 770, 782-83 (Tex. App. 2021).

269. *Free Speech Coalition*, 535 U.S. at 250; *Stevens*, 559 U.S. at 471 ("*Ferber* presented a special case: The market for child pornography was 'intrinsically related' to the underlying abuse . . . .'" (quoting *Ferber*, 458 U.S. at 759, 761)).

270. *Mecham*, 950 F.3d at 267; *Boland*, 698 F.3d at 884; *Hotaling*, 634 F.3d at 730; *Anderson*, 759 F.3d at 895.

271. *Mecham*, 950 F.3d at 266. For a critique of classifying images not produced through sexual abuse as "child pornography," see Adler, *supra* note 243, at 946-957.

is: its distribution, as *Ferber* put it, "violates 'the individual interest in avoiding disclosure of personal matters.'" [272] Conspicuously fake morphed images, like deepfakes, do not.

Deepfakes represent a more troublesome extension of the same First Amendment problem that the morphed-images cases presented. Morphed images proved relatively easy for courts to slot into an unprotected category of speech, albeit for questionable reasons. Unlike morphed images, deepfakes depicting adults do not fall into any recognized category of unprotected speech—at least, not if courts' refusals to find revenge porn categorically unprotected are any indication. Assuming *Stevens* is correctly read as hostile towards new categories of unprotected speech that might encompass deepfakes, anti-deepfakes laws will have to run the gauntlet of strict scrutiny.

### B. Impermissible Bans on Outrageous Expression

Although dilution and morphed CSAM show that *per se* bans on outrageous iconography have avoided First Amendment invalidation, such bans are not uniformly constitutional. Attempts to outlaw flag burning, the desecration of effigies, and written sexual fantasies reaffirm that anti-deepfakes laws respond to longstanding interests. And judicial curtailments of these attempted bans help us understand the circumstances that have failed to justify *per se* bans on outrageous symbolic expression.

#### 1. Flag Burning

Historical regulation of flag desecration helps explain anti-deepfakes laws' focus on altered images. Like trademarks, flags operate as Peircian symbols when they signify a particular group or nation. For some, the American flag is a symbol not just of the country, but also "of freedom, of equal opportunity, of religious tolerance, and of good will for other peoples who share our aspirations." [273] Desecrating the flag can be a grave affront; 48 states criminalized it.[274]

In 1989 a 5-4 Supreme Court struck down a Texas ban on flag burning as unconstitutional.[275] The Court held that Texas's asserted "interest in preserving the flag as a symbol of nationhood and national unity" could not justify criminalizing political expression that took the form of flag burning.[276] In dissent, Justice Stevens

272. 458 U.S. at 759 n.10 (quoting *Whalen*, 429 U.S. at 599).
273. Texas v. Johnson, 491 U.S. 397, 437 (1989) (Stevens, J., dissenting).
274. *Id.* at 434 (Rehnquist, C.J., dissenting).
275. *Id.* at 420 (majority opinion).
276. *Id.*

argued that the Texas statute did not proscribe the communication of "disagreeable ideas" so much as a "disagreeable" manner of communicating.[277] "The concept of 'desecration,'" he wrote, "does not turn on the substance of the message the actor intends to convey, but rather on whether those who view the act will take serious offense."[278] "[E]ven if the actor knows that all possible witnesses will understand that he intends to send a message of respect, he might still be guilty of desecration if he also knows that this understanding does not lessen the offense taken by some of those witnesses."[279]

Justice Stevens's argument about flag burning tracks intuitive attitudes about deepfakes. On his telling, what's wrong with flag burning is not so much the proposition it expresses as the manner of expression.[280] The same is true of deepfakes, which elicit special disgust not because they express any proposition, but because they employ a particular manner of expression. For example, one pornographic deepfakes site contained the disclaimer, "We respect each and every celebrity featured. The OBVIOUS fake face swap porn is in no way meant to be demeaning. It's art that celebrates the human body and sexuality."[281] I can't imagine this disclaimer placated anyone. Like Justice Stevens's example of flag burning "intend[ed] to send a message of respect," the "respect[ful]" creation of pornographic deepfakes to "celebrate[] the human body and sexuality" is still objectionable, because it is still desecration.

There remains tremendous political appetite to ban flag desecration. After *Johnson*, Congress passed the Flag Protection Act of 1989, which the Court swiftly held unconstitutional in another 5-4 decision.[282] Within the last 25 years, joint resolutions proposing a constitutional amendment granting Congress authority to ban flag burning have achieved supermajorities in the House and a near-supermajority in the Senate.[283] About half of Americans surveyed in 2020 believed

277. *Id.* at 438 (Stevens, J., dissenting).
278. *Id.*
279. *Id.*
280. *Id.*
281. Megan Farokhmanesh, *Is It Legal to Swap Someone's Face into Porn without Consent?*, THE VERGE (2018), https://www.theverge.com/2018/1/30/16945494/deepfakes-porn-face-swap-legal (last visited Feb 7, 2024).
282. U.S. v. Eichman, 496 U.S. 310, 314, 319 (1990).
283. Final Vote Results for Roll Call 296, https://clerk.house.gov/evs/2005/roll296.xml (last visited Feb 11, 2024); U.S. Senate: U.S. Senate Roll Call Votes 109th Congress - 2nd Session, https://www.senate.gov/legislative/LIS/roll_call_votes/vote1092/vote_109_2_00189.htm

that desecrating the flag should be illegal.[284] At a rally in June of 2020, President Trump proposed "legislation that if somebody wants to burn the American flag . . . they go to jail for one year."[285] He explained, "I'm a big believer in freedom of speech, but that's desecration. That's a terrible thing . . . ."[286]

Nonconsensual deepfakes and flag burning are both offensive not because of the propositions they communicate, but because of the outrageousness of their communicative method. But they do not cause identical sorts of offense. Flag burning concerns only the use of a single, "unique" symbol.[287] By contrast, to proscribe nonconsensual deepfakes is to proscribe infinitely many possible photorealistic depictions of every individual's unique face. Flag burning disrespects a collective, but a deepfake desecrates an individual likeness (although commentators accurately observe that the institution of pornographic deepfakes evinces disrespect for women collectively).[288] But anti-deepfakes laws share the same fundamental stance as Justice Stevens in *Texas v. Johnson*: what must be regulated is not the expression of any particular proposition or idea, but rather a mode of expression that is simply too outrageous to tolerate.[289]

### 2. Libel, Emotional Distress, and Effigies

The legal and cultural significance of effigies helps us understand why anti-deepfakes laws paradigmatically regulate uses of the face of another.[290] In 1992, the late singer Sinéad O'Connor concluded her performance on Saturday Night Live by holding a photograph of Pope John Paul II up to the camera and tearing it.[291]

---

(last visited Feb 11, 2024); Final Vote Results for Roll Call 234, 234, https://clerk.house.gov/evs/2003/roll234.xml (last visited Feb 11, 2024).

284. Jamie Ballard, *Half of Americans Say It Should Be Illegal to Burn the US Flag*, YOUGOV (2020), https://today.yougov.com/politics/articles/30491-flag-burning-legal-illegal-poll-data (last visited Feb 11, 2024).

285. President Trump Campaign Rally in Tulsa, Oklahoma, C-SPAN (2020), https://www.c-span.org/video/?473015-1/president-trump-campaign-rally-tulsa-oklahoma (last visited Feb 11, 2024).

286. *Id.*

287. *See generally Johnson*, 491 U.S. 397 (describing flag as "unique" throughout).

288. *See* Jesselyn Cook, *Here's What It's Like To See Yourself In A Deepfake Porn Video*, HUFFPOST (2019), https://www.huffpost.com/entry/deepfake-porn-heres-what-its-like-to-see-yourself_n_5d0d0faee4b0a3941861fced (last visited Jul 5, 2019) (quoting Mary Anne Franks).

289. *See Texas v. Johnson*, 491 U.S. at 437 (Stevens, J., dissenting).

290. *See* Part I.A.2, *supra*.

291. Jon Caramanica, *Nobody Forgets A Daring Moment On Television*, N.Y. TIMES, Jul. 28, 2023, at 3.

Her act is now praised as a protest against sexual abuse within the Roman Catholic Church.[292] At the time, however, O'Connor's demonstration led to death threats, a literal steamrolling of her records in Times Square, and significant damage to her musical career.[293] Her protest probably would not have elicited the same outcry if, instead of destroying a photograph, O'Connor had simply stared into the camera and stated, "I protest sexual abuse in the Roman Catholic Church." There's something especially outrageous about defacing a realistic likeness of a person.[294] Abrahamic religions acknowledge images' power by forbidding visual depictions of the godhead.[295] Islam forbids visual likenesses of prophets, and the past two decades have seen multiple outbreaks of deadly violence following Western political cartoonists' publication of disparaging caricatures of the Prophet Muhammad.[296]

Perhaps in recognition of the special power of effigies, historical definitions of libel have acknowledged that the defacement of images can provoke legally actionable outrage. Blackstone explained with respect to criminal libel that "it is immaterial with respect to the essence of a libel, whether the matter of it be true or false; since the provocation, and not the falsity, is the thing to be punished criminally."[297] Some historical sources suggest that civil libel could be shown from the defacement of images, such as burning a plaintiff in effigy.[298] And to this day,

---

292. *Id.*

293. Ethan Alter, *Why Sinéad O'Connor's 1992 "Saturday Night Live" Appearance Was "Like a Canceling,"* Yahoo Entertainment (Jul. 26, 2023), https://www.yahoo.com/entertainment/why-sin-ad-oconnors-1992-185050791.html (last visited Dec. 19, 2023); Simon Hattenstone, *Sinéad O'Connor: 'I'll Always Be a Bit Crazy, but That's OK,'* THE GUARDIAN, May 29, 2021, https://www.theguardian.com/music/2021/may/29/sinead-oconnor-ill-always-be-a-bit-crazy-but-thats-ok-rememberings (last visited Dec. 19, 2023).

294. See Tushnet, *supra* note 51 at 705, 709.

295. Adler, *supra* note 36 at 43–44.

296. Why Does Depicting the Prophet Muhammad Cause Offence?, BBC NEWS, Oct. 4, 2021, https://www.bbc.com/news/world-europe-30813742 (last visited Jan 25, 2024); Dan Bilefsky, *Denmark Is Unlikely Front in Islam-West Culture War*, N.Y. TIMES, Jan. 8, 2006, at 3; Nicki Peter Petrikowski, *Charlie Hebdo Shooting*, BRITANNICA (2024), https://www.britannica.com/event/Charlie-Hebdo-shooting (last visited Jan 25, 2024).

297. BLACKSTONE, WILLIAM, COMMENTARIES ON THE LAWS OF ENGLAND: A FACSIMILE OF THE FIRST EDITION OF 1765-1769, 4:150-53 (1979), *available at* https://press-pubs.uchicago.edu/founders/documents/amendI_speechs4.html.

298. Eugene Volokh, *Symbolic Expression and the Original Meaning of the First Amendment*, 97 GEO. L.J. 1057, 1065–66. *See also* Eyre v. Garlick, 42 J.P. 68 (1878) (considering appeal of civil libel action based on the defendants "burning [the plaintiff] in effigy" and declining

courts permit claims for intentional infliction of emotional distress when a plaintiff can show that a defendant intentionally or recklessly caused severe emotional distress by defacing an effigy.[299] (Anti-deepfakes statutes often omit an "intent to harm" requirement, which suggests a purpose of protecting a broader dignitary interest rather than simply preventing harassment.)[300]

In the mid-20th century, the Supreme Court interpreted the First Amendment to require defamation plaintiffs to prove the falsity of the defendant's statement, at least where the subject of the statement was a matter of public concern.[301] The First Amendment similarly limits public figures' claims for intentional infliction of emotional distress.[302] The Court also explained that "[u]nder the First Amendment there is no such thing as a false idea."[303] Accordingly, insofar as a defacement of an effigy merely expresses outrageous disrespect and does not insinuate any factual proposition, it cannot be defamatory today.[304] But the venerable history of legal and theological regulation of the creation of effigies suggests an understanding that images can provoke outrage that text may not.

As with pornographic deepfakes and flag burning, the chief harm of disparaging effigies derives from the mode of expression rather than the proposition expressed. A provocateur who wears a sandwich board displaying a written message of contempt for Muslims is in some abstract sense communicating the same

---

to grant new trial after jury delivered verdict favorable to all defendants but one who pleaded guilty, but observing, "whether this was libellous or not was a question for the jury"); Brown v. Paramount Publix Corp., 270 N.Y.S. 544, 548 (N.Y. App. Div. 1934) (referring to "the ancient libel committed by the burning of the plaintiff in effigy")

299. *See* RESTATEMENT (SECOND) OF TORTS § 46 (1965); Bowman v. Heller, 420 Mass. 517, 527 (Mass. 1995) (imposing liability for intentional infliction of emotional distress on an employee who distributed images of his supervisor's face superimposed on explicit photographs of nude women); Muratore v. M/S Scotia Prince, 656 F. Supp. 471, 482 (D. Me. 1987), *aff'd in relevant part*, 845 F.2d 347 (1st Cir. 1988).

300. *See* Part I.A.4, *supra*; Citron and Franks, *supra* note 86 (critizing "intent to harm" requirements in revenge porn laws as "convert[ing] what should be a sexual privacy law into a harassment law").

301. Philadelphia Newspapers, Inc. v. Hepps, 475 U.S. 767, 776 (1986); Milkovich v. Lorain J. Co., 497 U.S. 1, 19 & n.6 (1990).

302. Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 56 (1988).

303. Gertz v. Robert Welch, Inc., 418 U.S. 323, 339 (1974).

304. Criminal libel—to the extent it remains a viable doctrine at all—also incorporates constitutional limitations on liability, such as a truth defense. *See, e.g.*, State v. Turner, 864 N.W.2d 204, 209 (Minn. Ct. App. 2015); Tollett v. United States, 485 F.2d 1087, 1094 (8th Cir. 1973).

message as a provocateur who burns an effigy of the Prophet Muhammad. But the provocateurs' messages differ meaningfully.[305] Disrespecting an effigy is not a true or false statement, but an outrageous form of disrespect. Anti-deepfakes statutes don't regulate propositional statements; they regulate uses of effigies.

### 3. Written Sexual Fantasies

Legal analysis of fantasies disseminated in writing illuminates anti-deepfakes laws' "realism" requirements.[306] In 2013, a jury convicted former New York City Police Department officer Gilberto Valle of conspiracy to commit kidnapping, based on messages from the sexual fetish website darkfetishnet.com "in which Valle and three alleged co-conspirators discuss in graphic detail kidnapping, torturing, raping, murdering, and cannibalizing women."[307] Valle—who came to be known as the "Cannibal Cop"—had shared "Facebook photographs of women he knew" and exchanged detailed messages with other forum users about kidnapping, torturing, raping, and cannibalizing those women.[308] Valle disclosed the women's true first names, but never their surnames, and he consistently lied about details relevant to a potential kidnapping, such as where he lived, whether he had surveilled his putative targets, and where those targets lived.[309] Valle maintained that these supposedly incriminating communications were in fact consistent with fantasy, and that the government failed to establish that he had the necessary criminal intent or that he entered into an actual agreement to commit kidnapping.[310] More than a year after Valle's conviction, the district court entered a judgment of acquittal, which the Second Circuit affirmed.[311]

The government had argued that, while some of Valle's chats were "fantasy," others demonstrated "real" criminal intent.[312] An FBI agent who reviewed Valle's chats explained,

> In the ones that I believe[d] were fantasy, the individuals said they were fantasy. In the ones that I thought were real, people were sharing, the two people were sharing real details of women, names,

---

305. "The medium is the message." *See generally* MARSHALL MCLUHAN, UNDERSTANDING MEDIA: THE EXTENSIONS OF MAN (1994).

306. *See* Part I.A.7, *supra*.

307. United States v. Valle, 301 F.R.D. 53, 59 (S.D.N.Y. 2014), *aff'd in part, rev'd in part*, 807 F.3d 508 (2d Cir. 2015).

308. *Id.* at 59, 66-77.

309. *Id.* at 61, 85.

310. *Id.* at 83.

311. *Valle*, 301 F.R.D. at 115.

312. *Valle*, 807 F.3d at 516.

> what appeared to be photographs of the women, details of past crimes and they also said they were for real. . . .
>
> [In the "real" chats, the participants] described dates, names and activities that you would use to conduct a real crime . . . .[313]

Both the district court and the Second Circuit held that the "real" and "fantasy" chats were "indistinguishable"—both contained fabricated and fantastical elements, and even Valle's "real" chats were unaccompanied by any effort to meet putative co-conspirators in person—and thus could not prove the necessary criminal intent.[314]

In a similar case, *United States v. Alkhabaz*, the defendant, Jake Baker, posted to a public Internet newsgroup a story that "graphically described the torture, rape, and murder of a woman who was given the name of" one of his university classmates.[315] However, the defendant also transmitted written sexual fantasies in private emails to a single correspondent, and on the basis of these emails the government charged him with violating the interstate threat statute, 18 U.S.C. § 875(c).[316] The emails, among other things, express a desire "to do it to a really young girl" and describe how one might abduct an unspecified woman in Baker's dormitory.[317] The district court quashed the indictment on the ground that the emails did not express an intent to act and were thus First Amendment-protected speech. The court wrote, "Discussion of desires, alone, is not tantamount to threatening to act on those desires. Absent such a threat to act, a statement is protected by the First Amendment."[318] The Sixth Circuit affirmed on statutory rather than constitutional grounds. It interpreted the statute to require "that a reasonable person . . . would perceive" the ostensible threat "as being communicated to effect some change or achieve some goal through intimidation."[319] And, the court concluded, "no reasonable person would perceive such communications as being conveyed to effect some change or achieve some goal through intimidation. Quite the opposite, Baker and [Baker's correspondent]

313. *Valle*, 301 F.R.D. at 65.
314. *Valle*, 807 F.3d at 516-17, 523; *Valle*, 301 F.R.D. at 86.
315. United States v. Baker, 890 F. Supp. 1375, 1379 (E.D. Mich. 1995), *aff'd sub nom.* United States v. Alkhabaz, 104 F.3d 1492 (6th Cir. 1997).
316. *Id.* at 1378, 1386.
317. *Id.* at 1377, 1388.
318. *Id.* at 1388; *see id.* at 1389-90.
319. United States v. Alkhabaz, 104 F.3d 1492, 1495 (6th Cir. 1997). The Sixth Circuit later concluded that *Alkhabaz* wrongly read an "intimidation requirement" into the statute. *See* United States v. Doggart, 906 F.3d 506, 512 (6th Cir. 2018).

apparently sent e-mail messages to each other in an attempt to foster a friendship based on shared sexual fantasies."[320]

If Valle or Baker had communicated their sexual fantasies by making pornographic deepfake imagery of identifiable women and sharing it in online chats—as Valle had shared his verbal fantasies alongside their photographs and first names, and as Baker had done in text only—such conduct would undoubtedly be a "real" violation of many criminal and civil anti-deepfakes statutes. Yet several scholars have published works sympathizing with Valle and criticizing his prosecution as motivated by "discomfort, disgust, and confusion toward his online fantasy life" rather than thwarting dangerous criminality.[321] Valle's fantasies about torturing, raping, murdering, and cannibalizing female acquaintances surely strike many as deeply antisocial. In comparison to such ghoulishness, the fantasy that many pornographic deepfakes make manifest—"I'd like to see this person naked"—seems relatively innocuous.[322] Surely, then, scholars must be tripping over themselves to decry anti-deepfakes laws for "punishing sexual fantasy."[323] Right?

As far as I can tell, no scholars are on the record condemning anti-deepfakes laws as improperly punishing sexual fantasy.[324] Some scholars even explicitly frame

---

320. *Alkhabaz*, 104 F.3d at 1496. For a contrasting case, see Dayton v. Davis, 735 N.E.2d 939, 941–42 (Ct. App. Oh. 1999).

321. Andrew Gilden, *Punishing Sexual Fantasy*, 58 WM. & MARY L. REV. 419, 448–49 (2016); *see also generally* Thea Johnson & Andrew Gilden, *Common Sense and the Cannibal Cop Essays*, 11 STAN. J. C.R. & C.L. 313 (2015); Michal Buchhandler-Raphael, *Overcriminalizing Speech*, 36 CARDOZO L. REV. 1667, 1691 (2014); Kaitlin Ek, *Conspiracy and the Fantasy Defense: The Strange Case of the Cannibal Cop Note*, 64 DUKE L.J. 901, 941 (2014); *see also* Susan W. Brenner, *Complicit Publication: When Should the Dissemination of Ideas and Data Be Criminalized?*, 13 ALB. L.J. SCI. & TECH. 273, 379–83 (2002) (approving of the judgment in Alkhabaz). *But see* Nicholas Barnes, *The Cannibal Cop: Criminal Conspiracy in the Digital Age Notes & Comments*, 25 TEMP. POL. & CIV. RTS. L. REV. 1, 14 (2016) ("There can be no doubt that the Cannibal Cop agreed with others to commit a crime.").

322. The prosecution in *Valle* made this very point in its summation:

> There is a reason why the word 'fantasy' gets sprinkled over and over again through every cross-examination. . . . It is because [when] we think of fantasies, we normally have a positive idea. You think of Mariah Carey . . . . Gil Valle's fantasy is about seeing women executed. . . . . That's not a fantasy that is OK.

*Valle*, 301 F.R.D. at 107.

323. Gilden, *supra* note 318.

324. The EFF gets close when it suggests that "the already available legal remedies will . . . cover injuries caused by deepfakes" without acknowledging that these remedies probably do not

deepfakes as a kind of fantasizing that deserves to be punished.[325] The widespread intuition seems to be that a fantasy remains a fantasy when it's represented in writing, but it becomes "real" when represented in a photorealistic image.

But why would it be that disseminating a deepfake is a "real" harm, while disseminating gruesome text accompanied by non-pornographic photographs is constitutionally protected "fantasy?"[326] Valle even co-opted his targets' likenesses: though his portrayals of sadistic sexual conduct were strictly symbolic, he disseminated targets' photographs along with them.[327] What Valle and Baker didn't do, however, was manipulate an icon itself; Baker used text alone, and Valle paired innocuous photographs with appalling text. This is the only material difference between Valle's and Baker's chats and a deepfake offense. The dismissal of the cases against Baker and Valle, and the sympathy afforded to these defendants by some of the academic commentariat, imply that this difference is legally and morally dispositive. Unlike a verbal disquisition on cannibalism, a deepfake is an iconic desecration of a person's image. It is semiotics and the irrational power of images that explain why deepfakes are real criminal wrongs while Valle's and Baker's words were unactionable fantasies.

---

afford redress for non-deceptive deepfakes. David Greene, *We Don't Need New Laws for Faked Videos, We Already Have Them*, ELECTRONIC FRONTIER FOUNDATION (2018), https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them (last visited Apr 21, 2019).

325. *See, e.g.*, Jacquelyn Burkell & Chandell Gosse, *Nothing New Here: Emphasizing the Social and Cultural Context of Deepfakes*, FM (2019), https://firstmonday.org/ojs/index.php/fm/article/view/10287 (last visited Sep 17, 2023) ("[T]here is something powerfully disturbing and deeply wrong with being an involuntary participant in someone's sexual fantasies [made manifest] and having your likeness co-opted for the sexual purposes of an [unknown] other."); Regina Rini & Leah Cohen, *Deepfakes, Deep Harms*, 22 J. ETHICS & SOC. PHIL. 143, 147 (2022).

326. Cf. Gilden, *supra* note 318 at 471–72; Buchhandler-Raphael, *supra* note 318 at 1691.

327. Valle also "possessed images and videos involving acts of sexual violence against women," although the district court opinion gives no indication that he modified images of his targets to impart sexual content. *Valle*, 301 F.R.D. at 96.

| | | *Bans held constitutional* | | *Not held un-constitutional* | | *Bans unconstitutional* | | |
|---|---|---|---|---|---|---|---|---|
| | Deepfakes | Morphed CSAM | Revenge Porn | Dilution by Tarnishment | Virtual CSAM | Flag burning | Effigy burning | Written fantasies |
| Individual harm | ✓ | ✓ | ✓ | ✓ | X | X | ✓ | ✓ |
| Sexual | ✓ | ✓ | ✓ | ✓ | ✓ | X | X | ✓ |
| Iconic de-facement | ✓ | ✓ | X | X | ✓ | ✓ | ✓ | X |
| Discloses private facts | X | X | ✓ | X | X | X | X | X |
| Established category of unprotected speech? | ? *(probably not)* | ✓ | X | ? | X | X | X | X |

**Table 1:** Characteristics of Regulations of Outrageous Expression

### C. In Summation

The disparate legal doctrines this Part surveyed each illuminate a distinct aspect of the typical anti-deepfakes law. Trademark dilution tells us that we are still bound today by laws that regulate non-deceptive uses of images because of their putative tendency to negatively affect emotional attitudes towards the referents of those images. CSAM jurisprudence reveals that federal courts of appeals have uniformly upheld the constitutionality of criminal prohibitions on non-deceptive, noncommercial uses of icons that do not disclose private facts. Flag-burning bans show that desecrating an icon or symbol can be distinctly more offensive than using disparaging words, but their constitutional invalidation also shows that non-individualized harm may not be sufficient grounds for regulating outrageous speech. The arc of defamation law shows actionable reputation-harming statements must be false and not merely outrageous. And the legal status of written sexual fantasies suggests that anti-deepfakes laws' focus on photorealistic subject matter may be an essential limitation.

The constellation of features that characterize constitutional and unconstitutional regulations of outrageous iconography also underscores the anti-deepfakes laws' constitutional precarity. On one hand, because anti-deepfakes laws remedy harmful uses of images that target specific individuals, they are meaningfully distinct from unconstitutional attempts to regulate written sexual fantasies or non-individualized harms like flag burning, virtual CSAM, or false claims of military decorations. On the other hand, because covered deepfakes neither disclose private facts nor defame, and likely do not fall into an established category of unprotected speech, they differ meaningfully from iconic manipulations that courts have held constitutional to ban. Given this legal context, Part IV explains what defensible regulation of deepfake pornography might look like.

## IV. The Law of Deepfakes is the Law of Icons, Not Indices

AI's ability to synthesize photorealistic pornography has policymakers deeply concerned—particularly when that pornography depicts children.[328] The technology also threatens to push the judiciary's tortured semiotic reasoning to a breaking point. One issue in particular has added urgency to concerns about AI-generated pornography: in late 2023, researchers disclosed that a major dataset used for training image-generating AI contained hundreds of CSAM images.[329] It is not yet clear how *Free Speech Coalition*'s holding applies to photorealistic, AI-generated CSAM, nor is it clear in what ways the presence or absence of indexical CSAM in training data will be legally significant. But interested parties are already calling for legislation to address photorealistic, AI-generated pornography, and these calls will only intensify as image- and video-generating technology continues to develop and proliferate.[330] Attempting to regulate all photorealistic imagery as if it were indexical results in doctrinal incoherence. Regulating deepfakes requires employing the legal theories that regulate iconic imagery for its iconic properties.

---

328. *See, e.g.*, Meg Kinnard, *Prosecutors in All 50 States Urge Congress to Strengthen Tools to Fight AI Child Sexual Abuse Images*, AP NEWS, Sep. 5, 2023, https://apnews.com/article/ai-child-pornography-attorneys-general-bc7f9384d469b061d603d6ba9748f38a (last visited Sep 17, 2023).

329. David Thiel, *Identifying and Eliminating CSAM in Generative ML Training Data and Models*, 8 (2023), https://purl.stanford.edu/kh752sm9123 (last visited Dec 20, 2023).

330. *See, e.g.*, Kinnard, *supra* note 8.

### A. The Law of Indices Cannot Address Deepfakes Coherently

One response to the AI-generated pornography crisis is to simply ignore the semiotic differences between iconic and indexical media and equate the two legally. This approach simplifies the regulatory gameplan: just add deepfakes-related clauses to existing revenge pornography statutes—as several states have done already[331]—and treat photorealistic AI-generated CSAM as equivalent to indexical CSAM. At least where AI-generated CSAM depicts identifiable children, prosecutions under the morphed-images prong of the child pornography statute could kick into high gear, since courts seem unbothered by the semiotic infirmities of this approach.[332]

The law of indexical images is a powerful machine, but the complicated ontology of AI-generated images is already starting to grind its gears. Consider the problem of image-generating models trained on CSAM. Riana Pfefferkorn concludes that the federal child pornography statute prohibits media "generated using training data that included photographic CSAM" because such "abuse-trained" CSAM "'involves the use of' actual abuse" and thus meets the statutory definition of child pornography.[333] But Pfefferkorn's reasoning has strange implications that she does not acknowledge: it entails that the federal statute prohibits *any* abuse-trained AI-generated image that depicts sexually explicit conduct at all, irrespective of whether it appears to depict minors or adults. Recall that the definition of "child pornography" includes "any visual depiction . . . of sexually explicit conduct" when its production "involves the use of a minor engaging in sexually explicit conduct."[334] "Sexually explicit conduct" is defined broadly, and it includes various forms of intercourse and nudity without reference to the participants' ages.[335] Thus, if Pfefferkorn is correct that having trained on photographic CSAM is all that is required for an image-generating AI's output to "involve[] the use of a minor engaging in sexually explicit conduct," then that output will fall within the federal statute's coverage so long as it depicts anything that meets the broad definition of "sexually explicit conduct."[336]

---

331. *See, e.g.*, S. 1042A (N.Y., 2023-24 Sess.); S.B. 309 (Haw. 2021); H.F. 2240 (Ia. 2024); 13 V.S.A. § 2606.
332. 18 U.S.C.A. § 2256(8)(C). For a discussion of the semiotic confusion of morphed-images prosecutions, see *supra*, Part III.A.2.c).
333. PFEFFERKORN, *supra* note 218 at 10.
334. 18 U.S.C. § 2256(8).
335. 18 U.S.C. § 2256(2)(A).
336. 18 U.S.C. § 2256(8)(A).

Pfefferkorn's analysis has bizarre ramifications—like turning every sexually explicit image generated by StableDiffusion into federally proscribed "child pornography"—because her reading of the statute gives indexicality more legal significance than the concept can support. The presence of indexical CSAM in AI training data is a dreadful thing, but liability that hinges on this fact is both wildly overinclusive and wildly underinclusive. If "involv[ing] the use of a minor engaging in sexually explicit conduct" is the evil, *simpliciter*, then *every* image produced by an abuse-trained AI model—no matter how innocuous its content—is tainted, because its production as a matter of fact involved the real-life sexual abuse of a minor. On the flipside, prohibitions on morphed CSAM and nonconsensual deepfakes show us that photorealistic iconicity alone suffices for banning an image, even if the image has no indexical relation to real-life abuse and deceives no one. Nonconsensual, photorealistic pornography is regulated because it is an outrageous use of icons, not because it indexically documents abuse.

Judges and lawmakers are already being invited to confront photorealistic, AI-generated pornography with the same haphazard semiotic analyses that morphed-images cases like *Anderson* and *Hotaling* employed.[337] In September 2023, the attorneys general of 54 states and territories sent a letter urging Congress to "act to deter and address child exploitation, such as by expanding existing restrictions on CSAM to explicitly cover AI-generated CSAM."[338] The letter warned,

> Even in situations where the CSAM images generated by AI . . . depict[] children who do not actually exist, these creations are still problematic for at least four reasons: 1) this AI-generated CSAM is still often based on source images of abused children; 2) even if some of the children in the source photographs have never been abused, the AI-generated CSAM often still resembles actual children, which potentially harms and endangers those otherwise unvictimized children . . . ; 3) even if some AI-generated CSAM images do not ultimately resemble actual children, the images support the growth of the child exploitation market by normalizing child abuse and stoking the appetites of those who seek to sexualize

337. *See supra*, text accompanying notes 236-264.

338. National Association of Attorneys General, *Letter Re: Artificial Intelligence and the Exploitation of Children*, 2 (2023), https://www.naag.org/wp-content/uploads/2023/09/54-State-AGs-Urge-Study-of-AI-and-Harmful-Impacts-on-Children.pdf (capitalization altered).

> children; and 4) just like deepfakes, these unique images are quick and easy to generate using widely available AI tools.[339]

In support of the letter, South Carolina's Attorney General Alan Wilson asserted that generating realistic-looking child pornography "creat[es] demand for the industry that exploits children."[340]

The production of photorealistic, pornographic depictions of children is a deeply troubling phenomenon. But it is troubling because it is an outrageous and antisocial use of icons. The prosecutors' letter, meanwhile, appeals to indexicality. Its first point is a strict indexicality argument; taking it seriously suggests that the prosecutors' concern with pornographic images is far too narrow, since *all* outputs of abuse-trained AI are "based on source images of abused children." The second point seems to rely on the same conflation of indices and icons that has been used to uphold morphed-images laws. The third point is an argument that *Free Speech Coalition* expressly rejected.[341] The fourth point holds no independent weight; that these images are "easy to generate" is only bad if the images are themselves bad for other reasons.

Finally, Attorney General Wilson's remarks contravene *Free Speech Coalition*'s admonition that "[p]rotected speech does not become unprotected merely because it resembles the latter."[342] The argument that photorealistic AI-generated CSAM "exploits children" even when it does not depict an identifiable child depends on mistaking icons for indices. A vegan Impossible Burger resembles a hamburger, but this does not entail that Impossible Burgers "creat[e] demand for the industry that exploits" cows. It has been suggested that AI-generated CSAM could stoke demand for indexical CSAM produced specifically for training AI.[343] But that possibility (a) seems fairly speculative and (b) is unmoored from what makes photorealistic CSAM outrageous. I doubt many people's alarm about AI-generated CSAM would be assuaged if they were assured that advances in image synthesis would not encourage the production of indexical CSAM for use as training data.[344] The problem, once again, is not that AI-generated nonconsensual

---

339. *Id.* at 3.

340. Kinnard, *supra* note 325.

341. *Free Speech Coalition*, 535 U.S. at 253 ("The mere tendency of speech to encourage unlawful acts is not a sufficient reason for banning it.").

342. *Id.* at 255.

343. PFEFFERKORN, *supra* note 218 at 11.

344. But some people would indeed be reassured! *See* Danielle Bernstein, *Could AI-Generated Porn Help Protect Children?*, WIRED, https://www.wired.com/story/artificial-intelligence-

pornography indexically records harms, but that it iconically signifies something abhorrent.

### B. The Law of Icons Coherently Addresses Deepfakes

Regulating AI-generated pornography doesn't require willful ignorance of semiotic realities. As applied to images, defamation law, privacy law, and CSAM doctrine paradigmatically regulate actual or perceived indices. Deepfakes are icons, and they defy the rationales used to regulate indices. But we already have two relevant doctrines that regulate icons *qua* icons: obscenity and appropriation of likeness. These doctrines, unlike those that focus on images *qua* indices, regulate the aspects of deepfakes that actually trouble us. Indeed, if a nondeceptive deepfake causes harm of the sort anti-deepfakes laws mean to redress, it does so either because it is obscene or because it wrongfully appropriates someone's likeness.

#### 1. Obscene Deepfakes Are Bad Because They Are Obscene

Obscenity is a constitutionally unprotected category of speech.[345] Unlike child pornography, however, obscenity is unprotected not because it indexically records harmful conduct, but because "obscene . . . . utterances are no essential part of any exposition of ideas, and are of such slight social value as a step to truth that any benefit that may be derived from them is clearly outweighed by the social interest in order and morality."[346] Whether material is obscene depends on what is known as the *Miller* test, which considers

> (a) whether the average person, applying contemporary community standards would find that the work, taken as a whole, appeals to the prurient interest; (b) whether the work depicts or describes, in a patently offensive way, sexual conduct specifically defined by the applicable state law; and (c) whether the work, taken as a whole, lacks serious literary, artistic, political, or scientific value.[347]

The *Miller* test is not a backwards-looking examination of the circumstances of media's production, but a forward-looking inquiry into what the media

---

csam-pedophilia/ (last visited Oct 30, 2023) (positing "that AI-generated child sexual material could actually benefit society in the long run by providing a less harmful alternative to the already-massive market for images of child sexual abuse").

345. United States v. Williams, 553 U.S. 285, 288 (2008).

346. Roth v. United States, 354 U.S. 476, 485 (1957) (quoting Chaplinsky v. New Hampshire, 315 U.S. 568, 571-572 (1942)); *see also* PFEFFERKORN, *supra* note 218 at 3.

347. United States v. Schales, 546 F.3d 965, 970 (9th Cir. 2008) (quoting Miller v. California, 413 U.S. 15, 24 (1973)).

communicates. The state regulates obscenity not because producing obscenity entails inflicting real-life abuse, nor because obscenity defames anyone or invades privacy, but because regulating obscenity fulfills a "governmental responsibility for communal and individual 'decency' and 'morality.'"[348]

An obscenity-based approach to regulating pornographic AI output is already on the books, and federal prosecutors are beginning to employ it against AI-generated CSAM.[349] While "child pornography" can be prohibited even if it is not obscene, a distinct "child obscenity" statute, 18 U.S.C. § 1466A, proscribes sexual images of children that are obscene.[350] The different ontologies of obscenity and CSAM correspond to meaningful semiotic differences between paradigmatic obscenity and paradigmatic CSAM. Unlike the law of child pornography, which chiefly regulates indices, the law of child obscenity (and obscenity in general) focuses on icons. For example, § 2256 specifically excludes "drawings, cartoons, sculptures, or paintings" from its coverage of "image[s] . . . indistinguishable from" indexical child pornography, while § 1466A explicitly *includes* "visual depiction[s] of any kind, including a drawing, cartoon, sculpture, or painting."[351] These semiotic differences track the different harms that obscenity and CSAM laws address. The foundational harm targeted by CSAM laws is the harm inherent in indexical records of abuse. Photographing a child being sexually abused requires a child to be harmed; drawing a cartoon of such abuse does not. By contrast, the core harm of obscenity isn't in how it's produced, but in what it represents. An obscene photograph is just as bad as an obscene painting because both images are iconic signs for a worthless, antisocial message.

The child obscenity statute, and the catch-all obscenity statutes, §§ 1460-62, would seem like the perfect vehicles for addressing the harms of AI-generated pornography. Yet obscenity law is in desuetude; Pfefferkorn reports that in the past 20 years "there have been over fifty-fold more federal court decisions citing 2252A than 1466A."[352] Why might this be? Pfefferkorn suggests that child-

348. Louis Henkin, *Morals and the Constitution: The Sin of Obscenity*, 63 COLUM. L. REV. 391, 391 (1963).; *VanBuren*, 214 at 800 ("The purposes underlying government regulation of obscenity and of nonconsensual pornography are distinct . . . .").

349. Federal prosecutors have already filed at least one indictment prosecuting AI-generated CSAM under §1466A, the child obscenity statute. *See* Sealed Indictment, Dkt. No. 2, United States v. Steven Anderegg, No. 24-cr-50-jdp (W.D. Wis. May 15, 2024) [hereinafter "Anderegg Indictment"].

350. *See* United States v. Williams, 553 U.S. 285, 297 (2008).

351. *Compare* 18 U.S.C. § 2256(11) *with id.* § 1466A(a)-(b).

352. PFEFFERKORN, *supra* note 218 at 8.

obscenity offenses are harder to prosecute than child-pornography offenses: "The knowing possession, receipt, or distribution of (photographic) CSAM is tantamount to a strict liability offense, whereas an obscenity case entails the more probing inquiry of the three-pronged *Miller* test." [353] But Pfefferkorn also anticipates that § 1466A may become a more appealing vehicle for prosecutors as AI tools for producing photorealistic images proliferate and make it harder for the government to prove that the defendants' images depict a real-life child.[354]

Pivoting to obscenity law to address photorealistic pornography has risks. One risk is that obscenity might not cover the breadth of cases that current child pornography or anti-deepfakes laws would. For example, a recent Fifth Circuit decision held that graphic iconic and written depictions of violent sexual abuse of children were obscene, but a drawing of "an adolescent girl alone, reclining and apparently masturbating" with "no indication . . . [of] being forced to perform a sexual act" was not.[355] When it comes to non-child pornography, commentators are divided over obscenity's present-day viability: some maintain that "[t]oday, pornography is ubiquitous and essentially legal," while others call such arguments "short-sighted and, in many respects, incorrect."[356] The *Miller* test is surely harder to satisfy for pornography that only depicts adults, and obscenity prosecutions for pornography depicting adults are indeed rare—but they aren't non-existent.[357]

But at least when AI-generated pornography depicts no identifiable person, *Miller*'s demands are a feature, not a bug. Unlike indexical CSAM, which is abhorrent because of what it records, iconic CSAM is abhorrent because of what it

353. *Id.* at 8–9.

354. *Id.* at 9, 16–20.

355. United States v. Arthur, 51 F.4th 560, 570 (5th Cir. 2022).

356. *Compare* Brian L. Frye, *The Dialectic of Obscenity*, 35 HAMLINE L. REV. 229, 236 (2012); *with* Jennifer M Kinsley, *The Myth of Obsolete Obscenity*, 33 CARDOZO ARTS & ENT. L.J. 607, 610 (2016). *Cf. also* Kendra Albert, *Imagine a Community: Obscenity's History and Moderating Speech Online*, 25 YALE J.L. & TECH. SPECIAL ISSUE 59, 71 (2023) ("Courts and commentators generally agree that the First Amendment protects most pornography.").

357. *See, e.g.*, United States v. Stagliano, 693 F. Supp. 2d 25, 27 (D.D.C. 2010); United States v. Ragsdale, 426 F.3d 765, 769 (5th Cir. 2005) (affirming obscenity convictions for two videos, each depicting rape of a "young woman," one of whom is described as "about 20 [years] old"). The district court acquitted Stagliano and reportedly "called the government's case 'woefully lacking' or 'woefully inadequate.'" *See* Judgment of Acquittal as to John A. Stagliano, Dkt. No. 95, United States v. Stagliano, No. 1:08-cr-00093 (D.D.C. Sept. 2, 2010); Josh Gerstein, *DOJ Stumbles Prompt Porn Purveyor's Acquittal*, POLITICO (Jul. 16, 2010), https://www.politico.com/blogs/under-the-radar/2010/07/doj-stumbles-prompt-porn-purveyors-acquittal-028102 (last visited Feb 18, 2024).

depicts. If AI-generated pornography depicts no recognizable person and the prosecution fails to prove that it is obscene, then the material is no more harmful than the non-obscene drawing in *Arthur*.[358] Unlike doctrines made to address indexical images, obscenity doctrine actually measures the relevant variable for nondeceptive deepfakes with no identifiable subject: a popular consensus that expression has gone beyond the pale.

A distinct risk is that obscenity might be *too* inclusive of expression that the public unjustly disfavors. Pfefferkorn warns that "increased reliance on obscenity law risks enshrining regressive social norms about sex, sexuality, and sexual orientation."[359] This is indeed a hazard, but it is a hazard inherent to the project of regulating the outrageous use of icons. Insofar as they proscribe non-deceptive media, anti-deepfakes laws are quite self-consciously responding to the power that "regressive social norms" hold over our lives. Regressive social norms are what make nudity and sexuality a uniquely sensitive and shameful topic; they are what trigger harmful repercussions for sexual presentations that deviate from socially prescribed standards.[360] Just as they are premised on the power of regressive social norms, anti-deepfakes laws are also premised on the notion that some uses of icons trigger irrational beliefs in reasonable people. A perfectly rational viewer who encounters an obvious deepfake won't impute a photorealistic avatar's actions to the person the avatar resembles. If this were the expected reaction to outrageous images, there would be no need for anti-deepfakes laws to cover non-deceptive media, just as there would be no need for dilution law to cover tarnishing but non-confusing uses of trademarks. That anti-deepfakes laws *do* cover non-deceptive media shows that these laws assume that an ordinary person who encounters a deepfake will have an affective response rooted in sexual mores and irrational beliefs about iconography. If anti-deepfakes laws' very theory of harm derives from these widespread social biases and irrational beliefs, how can we expect to administer such laws without reference to those same biases and irrationalities?

Predicting exactly how obscenity law will address nonconsensual, pornographic deepfakes is outside this article's scope. My guess is that obscenity law will effectively regulate AI-generated pornography depicting children, whether

358. *See* 51 F.4th at 570.

359. PFEFFERKORN, *supra* note 218 at 21.

360. Citron is quite careful to acknowledge this very point. *See* Citron, *supra* note 5 at 1898 ("The recognition that intimate activity and nudity can be viewed as discrediting and shameful--and result in discrimination--is not to suggest that intimate behaviors and nudity are discrediting and shameful.").

or not those children are identifiable.[361] Obscenity law will probably be less effective at regulating deepfake pornography of identifiable adults.[362] In all events, an honest semiotic analysis shows that it is the law of icons, not the law of indices, that we must bring to bear on outrageous, non-deceptive deepfakes.

### 2. Appropriative and Offensive Deepfakes Are Bad Because They Are Appropriative and Offensive

Even before anti-deepfakes laws, we had some legal scaffolding to address non-obscene, pornographic deepfakes: the appropriation tort, discussed in Part I.[363] To effectively combat deepfakes, appropriation needs to encompass noncommercial uses of likeness—which Citron has proposed[364]—and it may need to be buttressed with criminal penalties. And, as it happens, this is exactly what anti-deepfakes laws do. Anti-deepfakes laws are best understood as an extension of appropriation to all dissemination of certain outrageous imagery, irrespective of whether the dissemination realizes an "advantage" for the defendant. This is a place courts, citing the First Amendment, have previously feared to venture.[365]

Appropriation regulates images *qua* icons, not *qua* indices: for example, Muhammad Ali used New York's appropriation statute to enjoin *Playgirl* magazine from publishing "a full frontal nude drawing" purporting to depict him.[366] And appropriation recognizes that the relevant harm is a hijacking of identity without consent, not the creation of expression that would be objectionable even if consensual. Functionally equivalent is Goldberg and Zipursky's proposal to extend false light to cover highly offensive speech that is not false.[367]

This reform requires not legislation that equates deepfakes with revenge porn, nor legislation that bans "false" depictions of persons, but legislation that bans

---

361. *See generally* Anderegg Indictment. *But see supra*, text accompanying note 355.

362. *See VanBuren*, 214 A.3d at 800-02 (holding that nonconsensual pornography is not necessarily obscene), *accord Casillas*, 952 N.W.2d at 639; *Austin*, 155 at 455; *Fairchild-Porche*, 638 S.W.3d at 782-83.

363. For discussion of criminal penalties for speech and strict-scrutiny analysis, see *VanBuren*, 214 A.3d at 812; Danielle Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345, 376-77 (2014).

364. CITRON, *supra* note 136 at 137.

365. Indeed, courts have interpreted the First Amendment to protect authors of expressive works *irrespective* of whether the work is ultimately sold for profit. *See, e.g.*, *De Havilland*, 21 Cal. App. 5th at 860; Guglielmi v. Spelling-Goldberg Prods., 25 Cal. 3d 860, 869, 603 P.2d 454, 460 (1979) (Bird, C.J., concurring). *See also* note 127, *supra*.

366. *See, e.g.*, Ali v. Playgirl, Inc., 447 F. Supp. 723, 726, 729 (S.D.N.Y. 1978).

367. Zipursky and Goldberg, *supra* note 106 at 18.

highly offensive appropriations of likeness. This is exactly what a new Australian criminal anti-deepfakes statute does overtly: unlike its American counterparts, the law simply states that its prohibition "does not apply if . . . a reasonable person would consider transmitting the material to be acceptable, having regard to the" totality of the circumstances.[368] Moreover, existing American law already redresses similar harms: the appropriation tort and trademark dilution redress non-deceptive, offensive uses of icons in commerce; and the law of morphed images criminalizes non-deceptive sexual depictions of children even in the absence of physical abuse. If the laws of morphed CSAM and dilution by tarnishment have thus far avoided constitutional invalidation, then this expanded version of appropriation should be able to coexist with them. And if we can't abide banning offensive iconography in the manner required to combat deepfakes, then we should reconsider the laws of morphed CSAM and tarnishment, too.

What, then, about the First Amendment? Almost certainly, at least some nonconsensual, pornographic deepfakes depicting adults are protected speech.[369] By prohibiting only sexual deepfakes, the typical anti-deepfakes law is a content-based restriction of protected speech.[370] Anti-deepfakes laws will force jurists to

368. CRIMINAL CODE AMENDMENT (DEEPFAKE SEXUAL MATERIAL) BILL 2024 (enacted Sept. 2, 2024), *available at* https://www.aph.gov.au/Parliamentary_Business/Bills_LEGislation/Bills_Search_Results/Result?bId=r7205. Of course, one should think carefully about who the "reasonable person" contemplated by this law is. *See* note 137, *supra*, and accompanying text.

369. *Cf.* note 268, *supra* (discussing revenge porn).

370. The weight of authority holds that the statutes most analogous to anti-deepfakes laws—revenge porn statutes—are content-based restrictions on speech. *VanBuren*, 214 A.3d at 811; *Katz*, 179 N.E.3d at 455; *Fairchild-Porche*, 638 S.W.3d at 782. *See also* State v. Culver, 918 N.W.2d 103, 108 n.7 (Wis. App. 2018) (prosecution and defense stipulated that revenge porn "statute is content-based"). The Supreme Court of Illinois held, over a dissent, that Illinois's revenge-porn statute was content neutral because it "distinguishes . . . based on whether the disseminator obtained the image under circumstances in which a reasonable person would know that the image was to remain private and knows or should have known that the person in the image has not consented to the dissemination." *Austin*, 155 N.E.3d at 456-58. This analysis is unpersuasive. "[T]he content of the image is precisely the focus of [the challenged statute]." *Id.* at 475 (Garman, J., dissenting). Similarly, the Ninth Circuit has held that California's right-of-publicity law is a content-based speech restriction. Sarver v. Chartier, 813 F.3d 891, 905-06 (9th Cir. 2016). Moreover, a conclusion that anti-deepfakes laws are content-based aligns with Supreme Court precedent. Laws that regulate only depictions of sex—as the paradigmatic anti-deepfakes law does—are content-based restrictions on speech. *See* United States v. Playboy Ent. Grp., Inc., 529 U.S. 803, 806, 811-12 (2000). So are laws that regulate speech based

assess whether cases like *Free Speech Coalition*, *Stevens*, *Johnson*, *Tam*, and *Brunetti* truly forbid banning expression simply because it is outrageous.[371] Courts thus far have managed to dodge this question when considering deepfakes' closest cousins, revenge porn and morphed CSAM. In revenge porn cases, courts could emphasize that the regulated media indexically documents true, private facts. In morphed images cases, courts could make expedient use of the categorical First Amendment exception for "child pornography," even though the rationales that justified that exception for indexical images do not extend easily to morphed images. When trademark dilution faces constitutional scrutiny, courts may invoke property rights.[372] Deepfakes offer none of these offramps.

In other words, what anti-deepfakes laws will do is force courts to consider whether American law can regulate outrageous, non-deceptive, non-obscene iconography *per se*—something it has reliably done and continues to do today—*even when jurists must admit that this is what the law is doing*. Pornographic deepfakes may or may not be a category of speech that has "historic[ally] and traditional[ly]" been unprotected.[373] But in all events, deepfakes bear critical resemblances to forms of non-deceptive, iconic signification that have either been held categorically unprotected—such as morphed CSAM and obscenity—or whose regulation has thus far avoided constitutional invalidation—such as dilution by tarnishment and the tort of appropriation of likeness. This resemblance may or may not save the full sweep of a typical anti-deepfakes statute from a constitutional challenge, but it does show the government's established interest in the regulation of outrageous iconography *per se*. And if anti-deepfakes laws' continuity with historical prohibitions on the outrageous treatment of icons cannot save them, it's doubtful anything else could. At its least useful, then, my analysis gives proponents of anti-deepfakes laws an honest way to lose a constitutional challenge they were doomed to lose; at its most useful, it gives them an honest way to win.

---

on its "emotive impact . . . on its audience." Boos v. Barry, 485 U.S. 312, 321 (1988); *see Johnson*, 491 U.S. at 412.

371. *Cf. Tam*, 582 U.S. at 223 (stating "a bedrock First Amendment principle: Speech may not be banned on the ground that it expresses ideas that offend.").

372. Cf. Lemley and Tushnet, *supra* note 201 at 107 n.93. Perhaps an explicit likeness-as-property regime, which state anti-deepfakes statutes typically eschew but which the draft NO FAKES Act provides, *see* NO FAKES Act of 2024 § 2(b)(2)(A)(i), would provide a suitable offramp.

373. *Stevens*, 559 U.S. at 468 (quoting Simon & Schuster, Inc. v. Members of N.Y. State Crime Victims Bd., 502 U.S. 105, 127, (1991) (Kennedy, J., concurring in judgment)).

## Conclusion

States have enacted a welter of brand-new anti-deepfakes laws, and similar federal legislation is pending. Existing defamation and privacy doctrine can explain much of these laws' scope. But privacy and defamation doctrine can't account for the full breadth of a typical anti-deepfakes law. This doctrinal mismatch shows not that anti-deepfakes laws are overinclusive, but rather that they redress an injury distinct from the injuries redressed by defamation and paradigmatic privacy law. Anti-deepfakes laws prohibit non-deceptive, outrageous uses of iconic signs, which cause harm independent of any factual proposition that they might communicate.

Although images are usually regulated for their indexical qualities—such as their ability to record harmful events or reveal private facts—or for deceptively resembling indexical records, several areas of American law regulate icons *qua* icons. Trademark dilution doctrine proscribes tarnishing uses of marks not because they cause confusion, but because even non-deceptive uses threaten to change observers' attitudes. The law of morphed CSAM images posits that "a child suffers [harm] from appearing as the purported subject of pornography in a digital image . . . regardless of the image's verisimilitude."[374] The harm to a brand that is tarnished, or to a child that is depicted in an obviously fake morphed image, is a harm rooted not in detached rationality but instead in the emotional power of images. That doesn't make the harm any less real.[375]

Thus, anti-deepfakes laws do not challenge us to examine whether our law can have any solicitude at all for irrational beliefs about images. The doctrines of trademark dilution and morphed images, as well as historical regulation of flag desecration and effigy burning, show that our law already extends solicitude to such beliefs. Rather, anti-deepfakes laws challenge us to decide whether the law will extend that solicitude to the particular irrational beliefs about images that harm people—women, overwhelmingly [376] —who appear in pornographic deepfakes without their consent. If the law can empower Coca-Cola to enjoin a (hypothetical) non-confusing "Coca-Cola Strip Club," might the law also empower an individual to enjoin non-deceptive, photorealistic, pornographic deepfakes that depict her?

374. *Anderson*, 759 F.3d at 896.

375. "If [people] define situations as real, they are real in their consequences." WILLIAM I. THOMAS & DOROTHY SWAINE THOMAS, THE CHILD IN AMERICA: BEHAVIOR PROBLEMS AND PROGRAMS 572 (1928), http://archive.org/details/in.ernet.dli.2015.155699 (last visited Apr 23, 2024) (I thank James Grimmelmann for bringing the "Thomas Theorem" to my attention.).

376. AJDER ET AL., *supra* note 2 at 2.

And if the law can't abide one of these possibilities, should it really abide either of them?

We cannot justify the regulation of photorealistic, AI-generated icons using the same rationales we have used to regulate indexical images. Regulating non-deceptive deepfakes is not about deciding what private facts may be disclosed, or what lies may be told, or what abuse may be recorded—all of which are questions that a law of indices can answer. It is about deciding how our society will tolerate its members to be depicted. This is a question only a law of icons can answer.

## Appendix: State Anti-Deepfakes Laws

**Covering depictions of adults*

| State | Citation | Civil/Criminal |
|---|---|---|
| Alabama | ALA. CODE § 13A-6-240 | Criminal |
| California | CAL. CIV. CODE § 1708.86 | Civil |
| Arizona | A.R.S. § 16-1023 | Civil |
| Colorado | COLO. REV. STAT. § 18-7-107 | Criminal |
| Delaware | DEL. CODE ANN. TIT. 10, § 7802 | Criminal |
| Florida | FLA. STAT. § 836.13(4) | Criminal |
| Georgia | GA. CODE § 16-11-90 | Criminal |
| Hawaii | HAW. REV. STAT. § 711-1110.9 | Criminal |
| Idaho | IDAHO CODE § 18-6606 | Criminal |
| Illinois | 740 ILL. COMP. STAT. 190/10 | Civil |
| Indiana | IND. CODE § 35-45-4-8 | Criminal |
| Indiana | IND. CODE § 34-21.5-2-1; IND. CODE § 34-21.5-3-1. | Civil |
| Iowa | IOWA CODE § 708.7 | Criminal |
| Kentucky | KY. REV. STAT. ANN. § 531.120; KY. REV. STAT. § 531.010(8) | Criminal |
| Louisiana | LA. STAT. § 14:73.13 | Criminal |
| Massachusetts | MASS. GEN. LAWS CH. 265, § 43A | Criminal |
| Minnesota | MINN. STAT. § 604.32 | Civil |
| New Hampshire | N.H. REV. STAT. § 644:9-A | Criminal |
| New York | N.Y. PENAL LAW § 245.15 | Criminal |
| New York | N.Y. CIV. RIGHTS LAW § 52-C | Civil |
| North Carolina | N.C. GEN. STAT. § 14-190.5A | Civil and criminal |
| Pennsylvania | 18 PA. STAT. AND CONS. STAT. § 3131 | Criminal |
| Pennsylvania | 42 PA. STAT. AND CONS. STAT. § 8316.1 | Civil |
| South Dakota | S.D. CODIFIED LAWS § 22-21-4 | Criminal |
| Texas | TEX. PENAL CODE § 21.165 | Criminal |
| Utah | UTAH CODE § 76-5B-205 | Criminal |
| Vermont | VT. STAT. TIT. 13, § 2606 | Civil and criminal |
| Virginia | VA. CODE § 18.2-386.2 | Criminal |
| Washington | WASH. REV. CODE § 9A.86.030 | Criminal |
| Washington | WASH. REV. CODE § 7.110.020 | Civil |