Printed in U.S.A.
Vol. 116, No. 3

Articles

DEEPFAKE PRIVACY: ATTITUDES AND REGULATION

Matthew B. Kugler & Carly Pace

ABSTRACT—Using only a series of images of a person's face and publicly available software, it is now possible to insert the person's likeness into a video and show them saying or doing almost anything. This "deepfake" technology has permitted an explosion of political satire and, especially, fake pornography. Several states have already passed laws regulating deepfakes, and more are poised to do so. This Article presents three novel empirical studies that assess public attitudes toward this new technology. In our main study, a representative sample of the U.S. adult population perceived nonconsensually created pornographic deepfake videos as extremely harmful and overwhelmingly wanted to impose criminal sanctions on those creating them. Labeling pornographic deepfakes as fictional did not mitigate the videos' perceived wrongfulness. In contrast, participants considered nonpornographic deepfakes substantially less wrongful when they were labeled as fictional or did not depict inherently defamatory conduct (such as illegal drug use). A follow-up study showed that people sought to impose both civil and criminal liability on deepfake creation. A second follow-up showed that people judge the creation and dissemination of deepfake pornography to be as harmful as the dissemination of traditional nonconsensual pornography—otherwise known as revenge pornography—and to be slightly more morally blameworthy.

Based on the types of harms perceived in these studies, we argue that prohibitions on deepfake pornographic videos should receive the same treatment under the First Amendment as prohibitions on traditional nonconsensual pornography rather than being dealt with under the less-protective law of defamation. In contrast, nonpornographic deepfakes can likely only be dealt with via defamation law. Still, there may be reason to allow for enhanced penalties or other regulations based on the greater harm people perceive from a defamatory deepfake than a defamatory written story.

AUTHORS—Matthew B. Kugler is an Associate Professor at Northwestern Pritzker School of Law. Carly Pace is a J.D. Candidate at Northwestern

Pritzker School of Law. The authors thank Ana Blinder, Anne Boustead, Zachary Clopton, Jill Doherty, Ezra Friedman, Enrique Guerra-Pujol, Joshua Kleinfeld, Andrew Koppelman, Dustin Marlan, Kirsten Martin, Robert McAuliff, Benjamin McJunkin, Janice Nadler, Laura Pedraza-Fariña, Sarath Sanga, Max Schanzenbach, David Schwartz, Victoria Schwartz, Nadav Shoked, Alexis Shore, David Simon, Roseanna Sommers, Matthew Spitzer, Michael Tremeski, and Elizabeth Wayne for their comments on earlier versions of this Article, and Laynie Barringer for helpful research assistance.

INTRODUCTION

In 2020, actress Kristen Bell was shocked to discover a pornographic video of herself online. The reason Bell was so surprised was that she had never filmed the video. In an interview with Vox, Bell stated, "We're having this gigantic conversation about consent, and I don't consent, so that's why it's not okay . . . even if it's labeled as, 'This is not actually her,' it's hard to

think about that."[1] The video was what is known as a "deepfake." Deepfakes are videos that use machine-learning algorithms to digitally impose one person's face and voice onto videos of other people.[2] The resulting doctored videos show people doing and saying things they never did or said. The number of videos like the one Kristen Bell found of herself is increasing. From July 2019 to June 2020, there was an increase of over 330% in the number of deepfake videos found online.[3] And the deepfake of Bell is a typical example of the genre. Ninety-six percent of all deepfake videos online are pornographic, and those depicted in pornographic deepfakes are almost exclusively women.[4] Nonpornographic deepfake videos have depicted politicians, corporate figures, and celebrities.[5]

As the opening example of Bell illustrates, many deepfake subjects feel harmed by their depictions in these false videos. The emerging scholarly literature on deepfakes discusses them causing two types of harm: dignitary harms to the individuals depicted in the videos (whether viewers believe the videos or not)[6] and political and national security harms to society from successfully deceptive videos.[7] Yet the literature has noted that there are few legal protections for deepfake subjects under traditional privacy law, and what law does exist—for example, the law of defamation—tends to target

[1] Cleo Abram, *The Most Urgent Threat of Deepfakes Isn't Politics. It's Porn.*, VOX: RECODE (June 8, 2020), https://www.vox.com/2020/6/8/21284005/urgent-threat-deepfakes-politics-porn-kristen-bell [https://perma.cc/2MTD-6XHN].

[2] Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1758 (2019).

[3] Henry Ajder, *Deepfake Threat Intelligence: A Statistics Snapshot from June 2020*, SENSITY (July 3, 2020), https://sensity.ai/deepfake-threat-intelligence-a-statistics-snapshot-from-june-2020/ [https://perma.cc/ZHW5-53U7]; *see also* HENRY AJDER, GIORGIO PATRINI, FRANCESCO CAVALLI & LAURENCE CULLEN, DEEPTRACE, THE STATE OF DEEPFAKES: LANDSCAPE, THREATS, AND IMPACT 1 (2019) [hereinafter DEEPTRACE] (reviewing the current landscape and describing the rise over the last several years).

[4] DEEPTRACE, *supra* note 3, at 1–2. Although one study found that 100% of pornographic deepfake videos targeted women, *see id.* at 2, there are some pornographic deepfake videos of male celebrities, though these male videos are comparatively rare. Such videos do exist, however. MrDeepFakes.com has a small "Gay" section that features male celebrities such as Chris Pratt, Chris Evans, and Tom Holland. Notably, the category has only ninety-five videos as of June 2021, whereas many of the other categories have three- or four-digit video counts.

[5] *Id.* at 2.

[6] *See, e.g.*, Danielle Keats Citron, *Sexual Privacy*, 128 YALE L.J. 1870, 1886, 1925 (2019) (describing human dignity as encompassing the ability to manage access to one's "naked body and intimate information").

[7] *See, e.g.*, Chesney & Citron, *supra* note 2, at 1783–84 ("[D]eep fakes have utility as a form of disinformation supporting strategic, operational, or even tactical deception.").

only deception-related harms and not dignitary violations.[8] The general problem is that the major privacy torts target those who obtain or publicize information that is both true and private. These torts are a poor match for the typical case of pornographic deepfakes, where that which is true (the person's face) is not private, and that which is private (the sex act) is not true.[9]

Given that existing laws tend not to cover deepfake videos, several states have moved to create new regulations to address them. In 2019, California passed two measures: one creating a civil cause of action for those featured in pornographic deepfakes and the other prohibiting the dissemination of unlabeled altered videos containing political candidates in the weeks leading up to an election.[10] Similarly, Virginia expanded its nonconsensual-pornography statute to cover morphed videos,[11] and Texas protected candidates in the lead-up to elections.[12] Notably, one Texas candidate has already attempted to avail himself of that law's protection.[13] New York has recently passed new legislation expanding its nonconsensual-pornography law and providing limited protection against commercial uses of deepfakes.[14] Many other states, as well as the federal government, have also considered action in recent months.[15] As nonconsensual-pornography

[8] *See, e.g.*, *id.* at 1793–94 (discussing defamation as a remedy); Kareem Gibson, Note, *Deepfakes and Involuntary Pornography: Can Our Current Legal Framework Address This Technology?*, 66 WAYNE L. REV. 259, 272–282 (2020) (discussing the limitations of various tort actions as a remedy); Russell Spivak, *"Deepfakes": The Newest Way to Commit One of the Oldest Crimes*, 3 GEO. L. TECH. REV. 339, 368–83 (2019) (analyzing the viability of various tort actions); Rebecca A. Delfino, *Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act*, 88 FORDHAM L. REV. 887, 918–21 (2019) (discussing the inadequacy of current criminal laws in addressing deepfakes).

[9] *See* Citron, *supra* note 6, at 1939.

[10] CAL. CIV. CODE § 1708.86 (West 2020) (creating a civil cause of action for those nonconsensually depicted in altered videos that show them engaging in sexually explicit conduct); CAL. ELEC. CODE § 20010 (West 2020) (prohibiting unlabeled, altered videos featuring political candidates in the weeks prior to an election).

[11] VA. CODE ANN. § 18.2-386.2 (West 2019).

[12] TEX. ELEC. CODE ANN. § 255.004(d) (West 2019).

[13] Jasper Scherer, *Sylvester Turner Calls for Investigation into Tony Buzbee Ad, Citing 'Deep Fake' Law*, HOUS. CHRON. (Oct. 18, 2019, 8:44 PM), https://www.houstonchronicle.com/news/houston-texas/houston/article/Sylvester-Turner-calls-for-investigation-into-14545665.php [https://perma.cc/42XX-V49Q] ("Mayor Sylvester Turner has called for the district attorney to open a criminal investigation into Tony Buzbee's campaign over a television ad that appears to show edited photos of Turner and an allegedly fake text between the mayor and a 31-year-old intern who works at the airport.").

[14] N.Y. CIV. RIGHTS LAW §§ 50-F, 52-C (McKinney 2021).

[15] *See, e.g.*, David Ruiz, *Deepfakes Laws and Proposals Flood US*, MALWAREBYTES LABS (Jan. 23, 2020), https://blog.malwarebytes.com/artificial-intelligence/2020/01/deepfakes-laws-and-proposals-flood-us/ [https://perma.cc/ZE73-DV8A] (describing current legislative efforts).

laws proliferated greatly over the 2010s,[16] deepfake laws seem poised to expand in the 2020s.

Yet deepfakes present a difficult and novel challenge for courts and lawmakers. They raise fundamental questions about the moral wrongfulness of new and unusual technological acts that may harm others. How wrong is it to use a publicly available photo of a person's face? Is it problematic to make a deepfake that is pornographic? What about one that is not? Is it still harmful if people know the deepfake is fake? Currently, there is very little data on how the public views deepfakes and, particularly, how the public may view different types of deepfakes.

This lack of understanding of public attitudes is a substantial problem. Legal scholars have argued that laws—especially criminal laws—should reflect the views of the society that they govern.[17] Prior research has shown that both over- and under-criminalization can substantially degrade the law's legitimacy in the eyes of the public and reduce public compliance with legal rules.[18] People reading news reports of unjust laws express a greater willingness to engage in illegal activities,[19] they exhibit a greater inclination toward jury nullification in mock-juror studies,[20] and they are even more likely to cheat on experimental tasks and to steal pens.[21] There are, therefore, high costs to what some authors have called "disillusionment" with the law.[22] If we do not know how the public views the moral wrongfulness of deepfake production, then we cannot pass laws conforming to those beliefs.

Public perceptions also play a substantial role in parts of privacy law, further strengthening the case for researching deepfake attitudes. The language of several privacy and privacy-related causes of action explicitly references the attitudes of the community or the reasonable person. Two of the core privacy torts—intrusion upon seclusion and public disclosure of private facts—require that the privacy invasions or information disclosures

---

[16] *See generally* Mary Anne Franks, *"Revenge Porn" Reform: A View from the Front Lines*, 69 FLA. L. REV. 1251 (2017) (reviewing the rapid expansion of nonconsensual-pornography laws from 2013 to 2017).

[17] *See, e.g.*, Tom R. Tyler & John M. Darley, *Building a Law-Abiding Society: Taking Public Views About Morality and the Legitimacy of Legal Authorities into Account When Formulating Substantive Law*, 28 HOFSTRA L. REV. 707, 719–22 (2000) ("To sustain its moral authority, the law must be experienced as consistent with people's sense of morality.").

[18] *See* Janice Nadler, *Flouting the Law*, 83 TEX. L. REV. 1399, 1415–16 (2005); Paul H. Robinson, Geoffrey P. Goodwin & Michael D. Reisig, *The Disutility of Injustice*, 85 N.Y.U. L. REV. 1940, 2005–06 (2010).

[19] Nadler, *supra* note 18, at 1415–16.

[20] *Id.* at 1424–25.

[21] Elizabeth Mullen & Janice Nadler, *Moral Spillovers: The Effect of Moral Violations on Deviant Behavior*, 44 J. EXPERIMENTAL SOC. PSYCH. 1239, 1239–45 (2008).

[22] Robinson et al., *supra* note 18, at 2005.

be "highly offensive to a reasonable person."[23] Public perceptions are similarly critical for understanding obscenity, which is often at issue in cases involving sexual content. The meaning of obscenity depends on "community standards," particularly in determining what is "patently offensive" within a community.[24] Everyday people often resolve these questions, embodying the judgment of their communities, via the jury system,[25] and previous empirical research has examined the degree of correspondence between actual community attitudes and jury decisions in obscenity cases.[26] The jury is used in a similar fashion to embody the community's views in defamation actions, in which the jury determines whether a given statement about a person would harm their reputation either in general or in the eyes of some relevant subset of their peers.[27]

Outside the privacy tort context, many scholars have advocated using public opinion data to inform the Fourth Amendment's reasonable-expectations-of-privacy analysis.[28] Professors Christopher Slobogin and Joseph Schumacher pioneered this method by having respondents rate the intrusiveness of a variety of law enforcement information-gathering

---

[23] RESTATEMENT (SECOND) OF TORTS §§ 652B, 652D (AM. L. INST. 1977).

[24] Miller v. California, 413 U.S. 15, 23–24 (1973); Daniel Linz, Edward Donnerstein, Kenneth C. Land, Patricia L. McCall, Joseph Scott, Bradley J. Shafer, Lee J. Klein & Larry Lance, *Estimating Community Standards: The Use of Social Science Evidence in an Obscenity Prosecution*, 55 PUB. OP. Q. 80, 82 (1991).

[25] This issue is not generally a matter for expert testimony. *See, e.g.*, St. John v. N.C. Parole Comm'n, 764 F. Supp. 403, 408–10 (W.D.N.C. 1991) (citing cases that establish that expert testimony need not be introduced in obscenity cases). Instead, the jury is expected to fulfill this role. *See, e.g.*, Piepenburg v. Cutler, 649 F.2d 783, 792 (10th Cir. 1981) (noting that "when the material itself is introduced into evidence, the jury may judge for itself, using its own sense of community standards, whether the material is obscene; that is, the jury brings to the trial the community standard and no evidence is necessary to establish it").

[26] *See* Linz et al., *supra* note 24, at 80–82; *see also* Daniel Linz, Kenneth C. Land, Bradley J. Shafer, Arthur C. Graesser, Edward Donnerstein & Patricia L. McCall, *Discrepancies Between the Legal Code and Community Standards for Sex and Violence: An Empirical Challenge to Traditional Assumptions in Obscenity Law*, 29 LAW & SOC'Y REV. 127, 134 (1995) (discussing the "prosecution-induced intolerance" phenomenon, whereby jurors may assume that the community is less tolerant to sexually explicit material because of law enforcement's intolerance towards those materials).

[27] *See, e.g.*, Lyrissa Barnett Lidsky, *Defamation, Reputation, and the Myth of Community*, 71 WASH. L. REV. 1, 6–8 (1996) (expressing skepticism about this idea of a community while at the same time recognizing its ubiquity in the doctrinal discussion). A defendant in a defamation case may also seek to show that a plaintiff is a public figure—which changes the required *mens rea*—and one way of doing that is surveying the local community to determine their level of recognition. Waldbaum v. Fairchild Publ'ns, Inc., 627 F.2d 1287, 1295 (D.C. Cir. 1980) ("The judge can examine statistical surveys, if presented, that concern the plaintiff's name recognition.").

[28] For an extensive discussion justifying the use of such data, see Matthew B. Kugler & Lior Jacob Strahilevitz, *Actual Expectations of Privacy, Fourth Amendment Doctrine, and the Mosaic Theory*, 2015 SUP. CT. REV. 205, 224–44 (2016).

techniques.[29] Similarly, Professors Christine Scott-Hayward, Henry F. Fradella, and Ryan G. Fischer and Professors Bernard Chao, Ian Farrell, Christopher Robertson, and Ms. Catherine Durso have investigated Americans' opinions and beliefs about forms of electronic surveillance, finding, for example, that people generally expect privacy in data, such as their cell phone location records.[30]

There is therefore a rich tradition of considering the public's views both when setting the boundaries of criminal laws and when considering the scope of a person's privacy rights in civil actions. And there is some danger in setting policy in this area absent a better understanding of how people actually view deepfake videos. Yet, to date, the authors are aware of no other study that examines public opinion on different kinds of deepfakes. Two questions, in particular, are left unanswered. First, do people view deepfakes as wrongful even if they are labeled as fake (and thus are not deceptive)? Second, are nonpornographic deepfakes harmful if they do not depict defamatory conduct?

These questions are especially important given the First Amendment challenges of deepfake regulation. The government cannot prohibit speech merely because the speech is false; there must be some additional problem.[31] Given that mere falsity is not enough, we look to two potential frameworks that would allow for regulation for deepfakes. One is a defamation-style framework. This approach would allow for the prohibition of deepfakes that (1) are false, (2) are intended for viewers to perceive as true, and (3) cause harm to the target's reputation or standing in the community.[32] In such a framework, labeling the deepfake as fake would remove all liability; it would negate the second element. If people view labeled deepfakes as harmless, then they are implicitly taking this defamation-style approach.

Alternatively, one could take a privacy-violation approach to deepfake regulation. Drawing a parallel to the existing law of nonconsensual pornography, this approach would view the harm as coming from the

---

[29] Christopher Slobogin & Joseph E. Schumacher, *Reasonable Expectations of Privacy and Autonomy in Fourth Amendment Cases: An Empirical Look at "Understandings Recognized and Permitted by Society,"* 42 DUKE L.J. 727, 737 (1993); CHRISTOPHER SLOBOGIN, PRIVACY AT RISK: THE NEW GOVERNMENT SURVEILLANCE AND THE FOURTH AMENDMENT 110–11 (2007); *see also* Jeremy A. Blumenthal, Meera Adya & Jacqueline Mogle, *The Multiple Dimensions of Privacy: Testing Lay "Expectations of Privacy,"* 11 U. PA. J. CONST. L. 331, 343–45 (2009) (replicating Slobogin and Schumacher's main results).

[30] Christine S. Scott-Hayward, Henry F. Fradella & Ryan G. Fischer, *Does Privacy Require Secrecy? Societal Expectations of Privacy in the Digital Age*, 43 AM. J. CRIM. L. 19, 45–58 (2015); Bernard Chao, Catherine Durso, Ian Farrell & Christopher Robertson, *Why Courts Fail to Protect Privacy: Race, Age, Bias, and Technology*, 106 CALIF. L. REV. 263, 301 (2018).

[31] *See infra* notes 228–233 and accompanying text.

[32] *See infra* notes 97–101 and accompanying text.

appropriation of a person's identity and the depiction of them in a highly private position. Labeling the videos as fiction does not meaningfully remove this harm; the target's identity is still being appropriated. To the extent that people view the creation of pornographic deepfakes as highly harmful and this harm as not mitigated by labeling, it may be appropriate to assimilate pornographic deepfake regulation into the broader law of nonconsensual pornography. Though this is most likely to be an issue for pornographic deepfakes, people may also view the appropriation of people's identities in the nonpornographic context as highly offensive, shedding light on which framework is proper there as well.

This Article presents the findings from three experimental studies that asked people to evaluate the wrongfulness of creating both pornographic and nonpornographic deepfake videos. Part I explains the rise of deepfake technology and the current scholarship on deepfake harms. It also reviews the current legal status of deepfakes and how it fits into holes in existing privacy laws. Part II introduces the three empirical studies. The primary study explores four main domains: pornographic videos and nonpornographic videos, either labeled fictional or unlabeled. Within both the pornographic and nonpornographic contexts, the study examines public reactions to a range of scenarios. This diverse set of scenarios allows us to consider the correspondence between public attitudes and both existing and proposed legal regimes.

This study finds that people are extremely critical of deepfakes, with many participants seeking to criminalize all types of deepfakes. Participants viewed deepfake videos as more wrongful and harmful than written accounts describing the same conduct. Though people regarded the production of nonpornographic deepfakes—which we call "attitudinal" deepfakes—as less wrongful when the videos were clearly marked as fictional, this was not the case for pornographic deepfakes. In fact, 92% of participants wanted to criminalize the dissemination of a pornographic deepfake even if the label indicated that it was fake. Pornographic deepfakes featuring celebrities (as opposed to everyday people) or non-nude but sexualized conduct were also all but universally condemned. These reactions do not merely reflect common opposition to pornography in all its forms: Prior research has shown that significantly fewer people, only about 30% of the public, want to criminalize pornography more generally.[33] In contrast, participants considered attitudinal deepfakes substantially less wrongful if they did not depict inherently defamatory conduct, such as illegal drug use. But many

[33] Charles Fain Lehman, *What Do Americans Think About Banning Porn?*, INST. FOR FAM. STUD. (Dec. 18, 2019), https://ifstudies.org/blog/what-do-americans-think-about-banning-porn [https://perma.cc/XUP9-DBCN].

participants still wished to assign criminal liability even for the creation of less obviously harmful attitudinal deepfake videos, such as one depicting a deceased scientist describing their life's work. A smaller follow-up study in Section II.D shows that participants generally support allowing for both civil and criminal causes of action against those who produce deepfakes. Finally, a second follow-up study reported in Section II.E shows that people judge pornographic deepfakes to be on par with traditional nonconsensual pornography. Specifically, they view the dissemination of a pornographic deepfake to be as harmful as the dissemination of traditional nonconsensual pornography, and they consider it marginally more morally blameworthy.

Part III considers the implications of these findings for legal reform. Whenever society seeks to regulate a new form of misconduct, one of its first tasks is to define what counts as wrong. Our data show that people are deeply skeptical of the involuntary sexualization that stems from pornographic deepfakes. They take a context-dependent view of the dignitary harms present in attitudinal deepfakes. The current civil and criminal regimes do not sufficiently reflect these moral intuitions. We proceed to explore whether attempts to bring the law into greater alignment with public attitudes would be constitutionally permissible under the First Amendment. Part III considers both the complexities of banning speech that is merely false as well as the kinds of harms that courts have recognized when considering cases involving nonconsensual pornography and morphed child pornography.[34] Ultimately, the fact that the harm perceived from pornographic deepfakes is not mitigated by labeling leads us to conclude that regulation of such videos should fall under the same First Amendment standards as regulation of nonconsensual pornography generally. The implications for nonpornographic deepfakes are less clear, and it may be proper to think of them primarily through the lens of defamation.

I. THE RISE OF DEEPFAKES AND THEORIES OF DEEPFAKE HARMS

Producing deepfake videos has gone from being extremely difficult to trivially easy in under five years.[35] This Part reviews the rise of deepfake technologies and then considers the kinds of societal and individual harms that may be caused by their increasing prevalence. It closes by reviewing the current legal status of deepfakes under various civil and criminal regimes.

---

[34] For definitions of these terms, see *infra* text accompanying note 236 (nonconsensual pornography), and *infra* text accompanying note 257 (morphed child pornography).

[35] For one indicator of the prevalence of generative adversarial networks (GANs), described below, see DEEPTRACE, *supra* note 3, at 3 (showing that a mere three academic papers mentioned GANs in their titles or abstracts in 2014 and over one thousand did so in 2019).

A. Deepfake Technology and the Rise of Consumer Use

Deepfake videos are generally created using generative adversarial networks (GANs), a technology created by Ian Goodfellow in 2014.[36] GAN technology involves the use of two neural networks in a dynamic that "mimics the back-and-forth between a picture forger and an art detective who repeatedly try to outwit one another."[37] The first network, known as the "generator," creates fake outputs until the second network, known as the "discriminator," cannot tell the difference between the generator's outputs and an original data set.[38] The result is a realistic-looking video. Essentially, the technology takes an image, such as a face, learns it, and inserts it into a video such that the substituted face appears seamlessly.

The rise of deepfake videos and consumer use of deepfake technology started in 2017 on the website Reddit. A user named "deepfake" posted doctored pornography that swapped the faces of celebrities and public figures with people in pornographic videos.[39] This user's posts became incredibly popular. A specialized Reddit page, known as a "subreddit," was dedicated exclusively to deepfake videos and quickly reached 90,000 community members.[40]

Although deepfake pornography has since been banned on Reddit,[41] the prevalence of deepfake videos on the internet is growing rapidly. One study found that in July 2019, there were 14,678 deepfake videos online, representing a near-100% increase from seven months earlier in December 2018.[42] As of June 2020, there were 49,081 deepfake videos online, representing an increase of over 330% in a year.[43] "Since December 2018, the number of deepfakes online is roughly doubling every six months,

---

[36] *See* Martin Giles, *The GANfather: The Man Who's Given Machines the Gift of Imagination*, MIT TECH. REV. (Feb. 21, 2018), https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/ [https://perma.cc/A7EX-QXQY].

[37] *Id.*

[38] *Id.*

[39] Meredith Somers, *Deepfakes, Explained*, MIT SLOAN (July 21, 2020), https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained [https://perma.cc/8U6Y-QCCH]; *Deepfakes*, KNOW YOUR MEME, https://knowyourmeme.com/memes/cultures/deepfakes [https://perma.cc/WMU2-YZR4].

[40] Mika Westerlund, *The Emergence of Deepfake Technology: A Review*, 9 TECH. INNOVATION MGMT. REV. 39, 41 (2019).

[41] Adi Robertson, *Reddit Bans 'Deepfakes' AI Porn Communities*, VERGE (Feb. 7, 2018, 1:28 PM), https://www.theverge.com/2018/2/7/16982046/reddit-deepfakes-ai-celebrity-face-swap-porn-community-ban [https://perma.cc/4CMF-A9ZN]; Arjun Kharpal, *Reddit, Pornhub Ban Videos that Use A.I. to Superimpose a Person's Face over an X-Rated Actor*, CNBC (Feb. 8, 2018, 6:44 AM), https://www.cnbc.com/2018/02/08/reddit-pornhub-ban-deepfake-porn-videos.html [https://perma.cc/HM9W-5U5H].

[42] DEEPTRACE, *supra* note 3, at 1, 16.

[43] Ajder, *supra* note 3.

confirming a continued exponential growth."[44] While this increase in the prevalence of deepfake videos can be attributed to consumer access to deepfake technology, it may also be attributed to its media coverage in recent years. Indeed, the media has often had the effect of popularizing dark corners of the internet. Take, for example, the case of Silk Road, the online marketplace that operated as a black market for guns, drugs, and other illicit goods and services.[45] Eventually, a journalist at *Gawker* discovered the website and published an article about it.[46] Within days, discussion of the website became part of the national discourse, customers flocked to the site, and the previously unknown website caught the attention of Congress and the Department of Justice.[47]

Some uses of deepfake technology have become mainstream. A simple Google search yields not only deepfake videos themselves, which are widely available on the internet, but also consumer access to the technology used to create these videos.[48] Independent phone applications can be downloaded to cell phones, where users can insert photos to create lifelike videos. Social media applications Snapchat and TikTok have integrated deepfake technology into their platforms as well.[49] For example, in December 2019, Snapchat announced a new tool called "Cameos," which allows users to insert their own pictures into a video setting to create a deepfake video.[50] However, these features generally limit what users can do with the deepfake technology. For example, the Cameos feature allows users to "jump into" preset scenes and customize captions.[51] These are generally intended to be fun or silly. One tutorial on Cameos shows how people can be inserted into

---

[44] *Id.*

[45] Caroline Sommers & Emily Bernstein, *Inside the FBI Takedown of the Mastermind Behind Website Offering Drugs, Guns and Murders for Hire*, CBS NEWS (Nov. 10, 2020, 11:03 PM), https://www.cbsnews.com/news/ross-ulbricht-dread-pirate-roberts-silk-road-fbi/ [https://perma.cc/VD9M-DNZF].

[46] NICK BILTON, AMERICAN KINGPIN: THE EPIC HUNT FOR THE CRIMINAL MASTERMIND BEHIND THE SILK ROAD 53 (2017).

[47] *Id.* at 56–58.

[48] Some of the top results from a Google search of "deepfake apps" in the summer of 2021 include Anya Zhukova, *7 Best Deepfake Apps and Websites*, ONLINE TECH TIPS (Aug. 24, 2020), https://www.online-tech-tips.com/cool-websites/7-best-deepfake-apps-and-websites/ [https://perma.cc/3X79-BY2D], and Beebom Staff, *10 Best Deepfake Apps and Websites You Can Try for Fun*, BEEBOM (Dec. 29, 2020), https://beebom.com/best-deepfake-apps-websites/ [https://perma.cc/SHC8-Y6DH].

[49] Michael Nuñez, *Snapchat and TikTok Embrace 'Deepfake' Video Technology Even as Facebook Shuns It*, FORBES (Jan. 8, 2020, 6:30 AM), https://www.forbes.com/sites/mnunez/2020/01/08/snapchat-and-tiktok-embrace-deepfake-video-technology-even-as-facebook-shuns-it/#3c01b4542c05 [https://perma.cc/JNL4-E7ZJ].

[50] *Introducing Cameos*, SNAP INC. (Dec. 9, 2019, 2:00 AM), https://newsroom.snap.com/introducing-cameos/ [https://perma.cc/FGB2-3L6X].

[51] *Id.*

videos showing them doing extreme sports, wearing a cat costume, or dressed as a Wicked Witch.[52]

Despite the growth of silly deepfakes through some more common applications, the overwhelming majority of deepfake videos on the internet are pornographic.[53] The majority of these deepfake videos are found on websites dedicated solely to deepfake pornography,[54] although deepfake videos are found on mainstream pornography websites as well.[55] One study found that 100% of these videos feature female subjects and that the majority depict famous women, such as actresses, musicians, and political figures,[56] but there are now pornographic deepfake videos that depict men as well.[57] Creators of pornographic deepfakes appear to be predominantly male, and pornographic deepfakes are sometimes used as a form of targeted harassment against women.[58] The use of deepfakes as a tool for harassment may explain why so many female political figures are the subjects of deepfakes.

In the nonpornographic context, the majority of deepfake videos depict famous people, such as those in the entertainment industry, politicians, and CEOs.[59] Often these nonpornographic deepfakes are intended to be satirical.[60] Unlike in the pornographic context, where the purpose of the video requires that the video appear realistic, the fact that a nonpornographic video is a deepfake can add to the joke. An oft-cited YouTube video of Bill Hader exemplifies the nature of these videos. The video shows a clip of Hader on the *Late Show with David Letterman* in 2008. Known for his celebrity impressions, Hader gives impressions of Tom Cruise and Seth Rogan, and each time he gives an impression, his face morphs into the face of the person he is impersonating.[61] The video, posted by YouTuber Ctrl Shift

---

[52] Techboomers, *How to Use Snapchat Cameos - New Feature!*, YOUTUBE (Jan. 14, 2020), https://www.youtube.com/watch?v=G11SL3azf6A [https://perma.cc/RM93-JGC2].

[53] DEEPTRACE, *supra* note 3, at 1.

[54] *Id.* at 6.

[55] *Id.*; Matt Burgess, *Porn Sites Still Won't Take Down Nonconsensual Deepfakes*, WIRED (Aug. 30, 2020, 9:00 AM), https://www.wired.com/story/porn-sites-still-wont-take-down-non-consensual-deepfakes/ [https://perma.cc/6ACE-PP87] (reporting that deepfake videos have been viewed millions of times, including on pornography sites that "rank in the top 10 biggest sites across the entire web").

[56] DEEPTRACE, *supra* note 3, at 2.

[57] *See supra* note 4.

[58] Sophie Compton, *More and More Women Are Facing the Scary Reality of Deepfakes*, VOGUE (Mar. 16, 2021), https://www.vogue.com/article/scary-reality-of-deepfakes-online-abuse [https://perma.cc/LN6R-ESAB]; *see also* Mary Anne Franks & Ari Ezra Waldman, *Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions*, 78 MD. L. REV. 892, 896–97 (2019) (commenting on the harassment possibilities of deepfakes).

[59] DEEPTRACE, *supra* note 3, at 2.

[60] *Id.* at 12.

[61] Ctrl Shift Face, *Bill Hader Channels Tom Cruise [DeepFake]*, YOUTUBE (Aug. 6, 2019), https://www.youtube.com/watch?v=VWrhRBb-1Ig&t=50s [https://perma.cc/FL6K-FBGT].

Face, has over eleven million views, and the title of the video labels it as a "[DeepFake]," meaning it is clearly labeled as fictional.[62]

Though most deepfake videos are of public figures, private individuals are also sometimes targeted. Social media gives deepfake producers access to images of private individuals in a way that was traditionally only true for celebrities.[63] This store of photos, coupled with the rise of consumer access to deepfake technology, makes the process of making deepfake videos of private individuals straightforward. There have already been a few cases of deepfake-facilitated harassment of private figures,[64] and nonconsensual deepfake pornography of private individuals is increasingly common.[65] Of course, people may create or consensually appear in deepfake videos in apparently innocuous contexts, such as through social media applications. In a relatively harmless case, a fifty-year-old man deepfaked himself as a young woman to increase the popularity of his video channel about motorbikes.[66]

B. Harms

The rising number of deepfake videos online has led to increased interest in the potential negative effects on deepfake subjects and society at large. The new scholarship on deepfakes has generally focused on two categories of harm associated with deepfake videos: individual harms to a deepfake subject's dignity and emotional well-being, and wider societal harms involving threats to national security and democratic institutions. Scholars have also sometimes discussed the macro-level implications of deepfakes and their contribution to the spread of misinformation.

---

[62] *Id.*

[63] Abram, *supra* note 1.

[64] *See, e.g.*, Jana Benscoter, *Pa. Woman Created 'Deepfake' Videos to Force Rivals off Daughter's Cheerleading Squad: Police*, PA. REAL-TIME NEWS (Mar. 12, 2021), https://www.pennlive.com/news/2021/03/pa-woman-created-deepfake-videos-to-force-rivals-off-daughters-cheerleading-squad-police.html [https://perma.cc/CGE2-R347] ("Police arrested a 50-year-old Bucks County woman March 4 for sending her teen daughter's cheerleading coaches fake photos . . . [of] her rivals . . . to try to get them kicked off the squad . . . .").

[65] *See, e.g.*, Giorgio Patrini, *Automating Image Abuse: Deepfake Bots on Telegram*, SENSITY (Oct. 20, 2020), https://sensity.ai/automating-image-abuse-deepfake-bots-on-telegram/ [https://perma.cc/B6WS-C84U] (reporting that, as of July 2020, a bot had "stripped" photos of over 100,000 women, which were then shared publicly); Matt Burgess, *Telegram Still Hasn't Removed an AI Bot That's Abusing Women*, WIRED (Nov. 18, 2020), https://www.wired.com/story/telegram-still-hasnt-removed-an-ai-bot-thats-abusing-women/ [https://perma.cc/8TSM-4PTH] ("Messaging app Telegram is under pressure to crack down on an AI bot that has generated tens of thousands of non-consensual images of women on its platform.").

[66] Tony Tran, *Young Female Twitter Star Turns Out to Be 50-Year-Old Man Using Deepfakes*, FUTURISM: THE BYTE (Mar. 21, 2021), https://futurism.com/the-byte/young-female-twitter-star-turns-out-50-year-old-man-using-deepfakes [https://perma.cc/T28L-7S8P].

1. Individual Harms

The potential for deepfakes to cause dignitary harms to deepfake subjects has almost exclusively been explored in the context of nonconsensual deepfake pornography.[67] These individual harms include both the harms associated with the video itself as well as the downstream emotional and reputational harm stemming from subsequent uses of the video and society's response to the person depicted. On the harms associated with the video itself, Professors Bobby Chesney and Danielle Citron highlight the intangible damage caused by the videos, which can "exploit an individual's sexual identity for other's gratification."[68]

As with other forms of nonconsensual pornography, nonconsensual deepfake pornography directly affects the sexual autonomy of the subjects it depicts. Citron notes that "[s]exual privacy concerns the social norms governing the management of boundaries around intimate life" and "involves the extent to which others have access to and information about people's naked bodies (notably the parts of the body associated with sex and gender); their sexual desires, fantasies, and thoughts; communications related to their sex, sexuality, and gender; and intimate activities (including, but not limited, to sexual intercourse)."[69] Although deepfakes do not depict the naked bodies of the deepfake subject—only the subject's face is taken—they still impinge on sexual autonomy by repurposing the subject's identity.

The core issue of nonconsensual pornography is consent, and deepfake pornography adds an additional layer because the individual depicted did not actually engage in the sexual behavior she is depicted as doing. Like the nonconsensual disclosure of pornography that depicts an individual engaging in activities they actually did, nonconsensual deepfake pornography is "an affront to the sense that people's intimate identities are their own to share or to keep to themselves."[70]

Sexual-privacy invasions can have profound effects. Victims report experiencing significant psychological impacts such as anxiety, depression,

---

[67] *E.g.*, Chesney & Citron, *supra* note 2, at 1772–75 (exploring the emotional consequences of sexually exploitative deepfakes but focusing on the practical and monetary harms of other deepfakes); *see also, e.g.*, Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 VA. J.L. & TECH. 1, 9 (2020) (noting that potential abuses of nonpornographic deepfakes could include depicting a president in such a way that interferes with an election or causes mass panic, but making no mention of the dignitary harms the individuals depicted could experience).

[68] Chesney & Citron, *supra* note 2, at 1772.

[69] Citron, *supra* note 6, at 1880.

[70] *Id.* at 1921.

loss of appetite, and suicidal ideation.[71] Although these impacts have not been widely studied, qualitative research on the psychological effects of nonconsensual pornography generally is consistent with these accounts and underscores their potential severity.[72] Further, victims of nonconsensual pornography experience harms in the form of societal reactions. For example, victims of nonconsensual pornography have reported experiencing job loss and barriers to employment as a result of appearing in these videos.[73] These secondary harms also exist in the deepfake context. In addition to the psychological impact caused by the creation of nonconsensual deepfake pornography, it has been used to threaten and harass victims.[74]

As Citron notes, "[w]hen the nude images of women and sexual minorities are posted online without consent, these individuals may be stigmatized."[75] This may be true even in the deepfake context, in which the images do not depict the actual bodies of the subjects, and the question remains whether labeling a deepfake video as fake ameliorates the harm to deepfake pornography victims. Public opinion data can shed light on the attitudes of everyday people toward these videos, and it can capture the reactions people have to videos even when they are labeled as fake. In Section II.E we explicitly contrast views toward deepfake pornography with views toward traditional nonconsensual pornography.

There does not appear to be any writing on the individual dignitary harms associated with nonpornographic deepfakes. Nevertheless, it is easy to imagine having a visceral negative reaction to seeing oneself depicted saying a string of racial slurs, endorsing a terrorist group, or doing cocaine when one has not done so, for instance, and such videos could also cause downstream effects on employability. We seek to fill this gap in the literature by exploring views of different types of nonpornographic deepfakes in Part II.

---

[71] *Id.* at 1926; *see also* Sophia Ankel, *Many Revenge Porn Victims Consider Suicide—Why Aren't Schools Doing More to Stop It?*, GUARDIAN (May 7, 2018), https://www.theguardian.com/lifeandstyle/2018/may/07/many-revenge-porn-victims-consider-suicide-why-arent-schools-doing-more-to-stop-it [https://perma.cc/C9L8-T4QQ] (discussing emotional ramifications to adolescent victims of revenge pornography).

[72] *See, e.g.*, Samantha Bates, *Revenge Porn and Mental Health: A Qualitative Analysis of the Mental Health Effects of Revenge Porn on Female Survivors*, 12 FEMINIST CRIMINOLOGY 22, 30–34 (2017) (describing negative mental-health effects after victimization via revenge pornography).

[73] Citron, *supra* note 6, at 1927–28.

[74] *See* Drew Harwell, *Fake-Porn Videos Are Being Weaponized to Harass and Humiliate Women: 'Everybody Is a Potential Target,'* WASH. POST (Dec. 30, 2018, 9:00 AM), https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/ [https://perma.cc/D7BD-3GYD].

[75] Citron, *supra* note 6, at 1925.

2. Societal Harms

In contrast to the limited consideration of nonpornographic deepfakes in the domain of individual dignity, there has been a great deal of concern about the potential of political deepfake videos to interfere with elections, harm national security, and undermine democratic institutions. Hypotheticals are routinely proposed, including the possibility of the release of deepfake videos the night before an election, a deepfake video depicting a government official declaring war, or a deepfake video confirming a rumor about a politician.[76] Chesney and Citron note that deepfake videos could jeopardize national security in myriad ways, including their use in military operations and to distract intelligence agencies.[77]

Though we have yet to see a sophisticated deepfake informational campaign, deepfake videos of political figures have already been made. In April 2018, director Jordan Peele and Buzzfeed CEO Jonah Peretti released a deepfake video depicting President Barack Obama saying outrageous things, such as "Ben Carson is in the sunken place," and "Stay woke, bitches."[78] Of course, President Obama has not said those things publicly, and the video ultimately reveals Jordan Peele as the voice actor. The video serves as a public service announcement to viewers about being "more vigilant with what we trust from the internet."[79] A similar video was created of Prime Minister Boris Johnson that depicted him endorsing his then-political opponent. As with the Obama deepfake video, the deepfaked version of Boris Johnson reveals the video is a deepfake and warns viewers that "the unregulated power of technologies like this risk fueling misinformation, eroding trust, and compromising democracy."[80]

---

[76] Brown, *supra* note 67, at 9.

[77] Chesney & Citron, *supra* note 2, at 1777 (identifying seven dimensions of societal harms, including "distortion of democratic discourse on important policy questions; manipulation of elections; erosion of trust in significant public and private institutions; enhancement and exploitation of social divisions; harm to specific military or intelligence operations or capabilities; threats to the economy; and damage to international relations").

[78] Aja Romano, *Jordan Peele's Simulated Obama PSA Is a Double-Edged Warning Against Fake News*, VOX (Apr. 18, 2018, 3:00 PM), https://www.vox.com/2018/4/18/17252410/jordan-peele-obama-deepfake-buzzfeed [https://perma.cc/5XCQ-DWRC].

[79] BuzzFeedVideo, *You Won't Believe What Obama Says in This Video!*, YOUTUBE (Apr. 17, 2018), https://www.youtube.com/watch?v=cQ54GDm1eL0&feature=emb_title [https://perma.cc/PU4T-BBU5].

[80] Darren Altman, Future Advocacy & Bill Posters, *DeepFake Boris Johnson*, YOUTUBE (Nov. 13, 2019), https://www.youtube.com/watch?v=gHbF-4anWbE [https://perma.cc/8RZW-QB3H].

Deepfake videos depicting politicians have generally remained satirical and have yet to undermine an American election,[81] but there have been instances when doctored videos have been the subject of national news. For example, a doctored video of House Speaker Nancy Pelosi emerged online in May 2019.[82] Also known as a "shallowfake," this video was slightly altered to depict Pelosi slurring her words.[83] While the video was identified as altered by media outlets, its release and subsequent reporting highlighted the implications of deepfake technology.[84]

At the core of the concern for deepfake technology is the spread of misinformation. Scholars have highlighted the acute issue this poses for journalists.[85] Chesney and Citron note that news organizations may encounter challenges to authenticating evidence, which leads to a chilling effect on news reporting.[86] Professor Nina Brown highlights a broader effect of deepfake technology: erosion of public trust.[87] She suggests that when people can no longer believe what they see, people will "deny actual events captured on video" and "be disinclined to trust *any* video evidence, whether offered as part of a news story, or as evidence in a courtroom."[88] Similarly, Professor Regina Rini argues "that backstop crises triggered by contested deepfakes will lead to erosion of the reliability that recordings provide to our testimonial practices."[89] Americans are already reported to mistrust the media,[90] so the rise in deepfake technology may exacerbate this mistrust.

---

[81] Gary Grossman, *Deepfakes May Not Have Upended the 2020 U.S. Election, but Their Day Is Coming*, VENTURE BEAT (Nov. 1, 2020, 2:22 PM), https://venturebeat.com/2020/11/01/deepfakes-may-not-have-upended-the-2020-u-s-election-but-their-day-is-coming/ [https://perma.cc/82DS-738P].

[82] *Doctored Nancy Pelosi Video Highlights Threat of "Deepfake" Tech*, CBS NEWS (May 26, 2019, 9:26 AM) [hereinafter *Doctored Nancy Pelosi Video*], https://www.cbsnews.com/news/doctored-nancy-pelosi-video-highlights-threat-of-deepfake-tech-2019-05-25/ [https://perma.cc/MLZ9-B9G7].

[83] Jane Lytvynenko & Craig Silverman, *Why the Altered Videos of Pelosi Will Never Go Away*, BUZZFEED NEWS (May 27, 2019), https://www.buzzfeednews.com/article/janelytvynenko/altered-videos-of-pelosi-will-never-go-away [https://perma.cc/Q7MV-VW54].

[84] *See Doctored Nancy Pelosi Video*, *supra* note 82; Drew Harwell, *Faked Pelosi Videos, Slowed to Make Her Appear Drunk, Spread Across Social Media*, WASH. POST (May 24, 2019, 3:41 PM), https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/ [https://perma.cc/LYV8-PPBH]; Maheen Sadiq, *Real v Fake: Debunking the 'Drunk' Nancy Pelosi Footage – Video*, GUARDIAN (May 24, 2019, 12:38 PM), https://www.theguardian.com/us-news/video/2019/may/24/real-v-fake-debunking-the-drunk-nancy-pelosi-footage-video [https://perma.cc/NV8J-5YR2].

[85] Brown, *supra* note 67, at 12; Chesney & Citron, *supra* note 2, at 1784.

[86] Chesney & Citron, *supra* note 2, at 1784–85.

[87] Brown, *supra* note 67, at 8–14.

[88] *Id.* at 11.

[89] Regina Rini, *Deepfakes and the Epistemic Backstop*, 20 PHILOSOPHERS' IMPRINT 1, 11 (2020).

[90] Megan Brenan, *Americans Remain Distrustful of Mass Media*, GALLUP (Sept. 30, 2020), https://news.gallup.com/poll/321116/americans-remain-distrustful-mass-media.aspx [https://perma.cc/

Professors Jessica Silbey and Woodrow Hartzog actually refer to this as an "upside" of deepfakes in that they expose the existing rot in our journalistic and electoral institutions and may stimulate broader reforms.[91]

C. Existing Civil and Criminal Frameworks

Despite this growing discussion of deepfake harms, there are few remedies under current law. This Section reviews the various civil remedies that might be available to victims of deepfakes, paying specific attention to unlabeled deepfakes because falsity is often determinative in privacy law.

Traditional tort and privacy law causes of action such as public disclosure of private fact and intrusion upon seclusion are generally not applicable in the deepfake context. Public disclosure of private fact involves the disclosure of a private matter that is "highly offensive to a reasonable person" and "not of legitimate concern to the public."[92] But deepfakes are not facts—they are entirely made up—so they cannot be private facts. Intrusion upon seclusion claims involve an intentional intrusion, "physically or otherwise, upon the solitude or seclusion of another or his private affairs or concerns" that "would be highly offensive to a reasonable person."[93] When distributors create deepfake videos using photographs found on the internet, no intrusion is required.[94] This is even clearer in the celebrity context, where a deepfake creator need commit no fresh intrusion to repurpose internet photographs taken by paparazzi or posted on social media.[95] From a privacy-as-information standpoint, there is not even a privacy intrusion: all that is being used is a person's face, which is generally not private.[96]

G652-D6JC] (reporting that 27% of Americans trust the media "not very much" and 33% trust the media "not at all").

[91] *See* Jessica Silbey & Woodrow Hartzog, *The Upside of Deep Fakes*, 78 MD. L. REV. 960, 964–65 (2019) ("Perhaps the vivid threat of deep fakes can muster will to salvage journalism from the ravages of an economic system transformed by technology that appears to value viral lies over truth by subsidizing a free press with public funds and incentivizing the reestablishment of the journalistic profession.").

[92] RESTATEMENT (SECOND) OF TORTS § 652D (AM. L. INST. 1977).

[93] *Id.* § 652B.

[94] Chesney & Citron, *supra* note 2, at 1795 ("Deep-fakes usually will not involve invasions of spaces (either physical or conceptual like email inboxes) in which individuals have a reasonable expectation of privacy."); *see also* Nader v. Gen. Motors Corp., 255 N.E.2d 765, 770 (N.Y. 1970) (discussing how some acts are not intrusions upon seclusion because they are not done to obtain information).

[95] Spivak, *supra* note 8, at 379 ("In many (though not all) cases, the deepfake subject has either put the photos into the public by posting them online or consented to their collection by posing for paparazzi. Deepfakers have not violated anyone's personal space to obtain the necessary information to create and publish their work.").

[96] *See, e.g.*, United States v. Dionisio, 410 U.S. 1, 14 (1972) ("No person can have a reasonable expectation that others will not know the sound of his voice, any more than he can reasonably expect that his face will be a mystery to the world.").

Victims of nonconsensual deepfake videos may have more success with defamation or false light claims *if it is unclear that the videos are fake*. Defamation requires the publication of a false fact that harms the reputation of another.[97] False light is a similar cause of action that requires one to be portrayed falsely in a manner that is "highly offensive to a reasonable person."[98] So there could easily be liability if a convincing deepfake showed a person committing a crime or engaging in disreputable conduct. Courts are also likely to find unlabeled pornographic deepfakes defamatory given the reputational harms of being in a pornographic video.[99] Similarly, courts may uphold a false light claim by concluding that falsely depicting a person as engaging in sexual conduct is highly offensive to a reasonable person.[100] Though public figures generally face additional burdens under defamation law, these barriers likely will not pose substantial obstacles here.[101]

Private citizens and public figures may therefore be successful in bringing defamation or false light claims for unlabeled pornographic deepfakes and unlabeled nonpornographic deepfakes that depict disreputable conduct. Most likely, the dispute in a particular case would be over whether the deepfake video was presented as if it were real. However, satirical deepfakes are likely more challenging cases. Though deepfake videos that depict a person engaging in illegal or extreme behavior are more likely to harm a person's reputation—qualifying for defamation liability—parody or satirical deepfake videos that depict an individual engaging in merely embarrassing behavior likely do not inflict the same reputational harm.

A final tort possibility is intentional infliction of emotional distress.[102] This tort is generally difficult to satisfy—because it requires extremely

---

[97] RESTATEMENT (SECOND) OF TORTS §§ 558–59.

[98] *Id.* § 652E.

[99] Chesney & Citron, *supra* note 2, at 1772–75 (describing how being depicted in fake pornography videos may be expected to have collateral consequences for future social and employment prospects given existing research on nonconsensual pornography).

[100] *See* Kareem Gibson, Note, *Deepfakes and Involuntary Pornography: Can Our Current Legal Framework Address This Technology?*, 66 WAYNE L. REV. 259, 278 (2020).

[101] In *New York Times Co. v. Sullivan*, the Supreme Court held that public figures must prove a heightened *mens rea* of "actual malice"—that the statement was made "with knowledge that it was false or with reckless disregard of whether it was false or not." 376 U.S. 254, 279–80 (1964). But the creator of a deepfake knows it is fake, so this requirement would generally be satisfied. Unlike with defamation, the Supreme Court has not decided whether the heightened standard applies to false light claims, but some jurisdictions have concluded that it does. *See, e.g.*, West v. Media Gen. Convergence, Inc., 53 S.W.3d 640, 647 (Tenn. 2001) ("We hold that actual malice is the appropriate standard for false light claims when the plaintiff is a public official or public figure, or when the claim is asserted by a private individual about a matter of public concern.").

[102] *See* RESTATEMENT (SECOND) OF TORTS § 46 ("One who by extreme and outrageous conduct intentionally or recklessly causes severe emotional distress to another is subject to liability for such emotional distress, and if bodily harm to the other results from it, for such bodily harm.").

outrageous conduct—and it faces substantial First Amendment problems when applied to public figures or speech on public issues. In *Hustler Magazine, Inc. v. Falwell*, for example, the famous pastor Jerry Falwell sued *Hustler Magazine* for, among other things, intentional infliction of emotional distress for publishing what might be considered the written equivalent of a deepfake—a parody advertisement that said Falwell had engaged in sexual conduct with his mother in an outhouse.[103] Noting that the advertisement in question was a departure from traditional caricatures of political figures, the Court nevertheless protected the speech to avoid chilling political dialogue.[104] Similarly, extreme anti-gay-rights protests adjacent to a military funeral were held to not give rise to intentional infliction of emotional distress because they concerned a major public issue and violated no other laws.[105] This tort would therefore be a hard sell in any politically charged case.

Consequently, tort law provides little protection against deepfakes unless the deepfakes purport to be accurate depictions of facts. A deepfake that announces itself as fake is immune to the major privacy torts, fails the test for defamation, and is unlikely to be extreme enough to qualify for intentional infliction of emotional distress. Some states may provide some relief through right-of-publicity laws, but these often protect against the exploitation of a person's likeness in advertising and commerce, rather than in general.[106] A minority of states provide broader protection here, however, that may apply to deepfakes.[107]

Statutory protection under nonconsensual-pornography laws is little better in almost all states. State laws that do not explicitly address deepfakes seldom apply to deepfakes. For example, Texas's nonconsensual-pornography statute criminalizes the nonconsensual disclosure of "visual material depicting another person with *the person's* intimate parts exposed or engaged in sexual conduct."[108] Statutes written in this manner likely do not apply to deepfake pornography because those videos usually do not depict the real body of the victim. Some states statutes, for example, North Dakota's, are broader and prohibit the dissemination of a "visual depiction"

---

[103] 485 U.S. 46, 48 (1988).

[104] *Id.* at 55–57.

[105] Snyder v. Phelps, 562 U.S. 443, 454–58 (2011).

[106] *See, e.g.*, 765 ILL. COMP. STAT. 1075/5, /35 (requiring a use for "a commercial purpose"); VA. CODE ANN. § 8.01-40 (2021) (requiring use "for advertising purposes or for the purposes of trade"); CAL. CIV. CODE § 3344 (West 2021) (requiring use "for purposes of advertising or selling, or soliciting purchases of, products, merchandise, goods or services").

[107] See, for example, OKLA. STAT. tit. 12, § 1450 (2021), an anti-catfishing statute that allows for a cause of action against those who engage in impersonation online with an intent to harass.

[108] TEX. PENAL CODE ANN. § 21.16 (West 2019) (emphasis added).

or "any intimate image" that depicts nudity or sexual conduct.[109] A deepfake pornographic video fits under that definition. The North Dakota statute, however, further requires that the dissemination of the image or video be in violation of a reasonable expectation of privacy.[110] Although there are inherent privacy concerns with deepfake pornography, deepfake pornography is often made without the victim's knowledge, so statutes requiring that the victim intended that an image be kept private do not translate to the deepfake context. This type of requirement is common in nonconsensual-pornography statutes. New York's statute includes as an element that the "still or video image was taken under circumstances when the person depicted had a reasonable expectation that the image would remain private and the actor knew or reasonably should have known the person depicted intended for the still or video image to remain private."[111] Similarly, Connecticut's statute requires that an image be disseminated with the knowledge that the person depicted "understood that the image would not be so disseminated."[112] A recent analysis by Professors Jonathan Sales and Jessica Magaldi found that thirty nonconsensual-pornography statutes have a similar expectation of privacy requirements.[113]

Several new laws specifically targeting deepfakes were passed in 2019 and 2020. These laws are highly targeted and still few in number. Virginia, for example, amended its nonconsensual-pornography statute to address deepfakes specifically.[114] Section 1708.86 of the California Civil Code provides a civil cause of action for an individual who is depicted in a pornographic deepfake video without their consent. The statute imposes civil liability on anyone who either creates and distributes the deepfake or who distributes the deepfake knowing it was created without consent.[115] The statute carves out exceptions to liability, including when the deepfake is "[a] matter of legitimate public concern" or "[a] work of political or newsworthy value or similar work."[116] Notably, that the deepfake video is labeled as fake

---

[109] N.D. CENT. CODE § 12.1-17-07.2 (2015).

[110] *Id.*

[111] N.Y. PENAL LAW § 245.15 (McKinney 2019).

[112] CONN. GEN. STAT. § 53a-189c (2021).

[113] Jonathan S. Sales & Jessica A. Magaldi, *Deconstructing the Statutory Landscape of "Revenge Porn": An Evaluation of the Elements that Make an Effective Nonconsensual Pornography Statute*, 57 AM. CRIM. L. REV. 1499, 1524 (2020).

[114] *See* VA. CODE ANN. § 18.2-386.2 (West 2019). In 2019, Virginia amended its revenge pornography statute to include "any videographic or still image created by any means whatsoever that depicts another person." *Id.*

[115] CAL. CIV. CODE § 1708.86(b)(1)–(2) (West 2020).

[116] *Id.* § 1708.86(c)(1)(B)(i)–(ii).

is not a permissible defense.[117] A victim has the option to recover either economic and non-economic damages caused by the deepfake video or substantial statutory damages.[118] The statutory damages range from $1,500 to $30,000 unless the distributor acted with malice, in which case a victim can recover up to $150,000.[119] A victim may also recover punitive damages and attorneys' fees, as well as receive injunctive relief.[120]

Section 20010 of the California Elections Code creates a civil cause of action for a political candidate who appears in a deepfake video. The statute prohibits the distribution of unlabeled "materially deceptive audio or visual media" featuring "a candidate for elective office [who] will appear on the ballot" with "the intent to injure the candidate's reputation or to deceive a voter" within sixty days of an election.[121] The statute defines "materially deceptive audio or visual media" as any audio or video of a candidate that has been intentionally manipulated so that it appears authentic to a reasonable person and causes "a reasonable person to have a fundamentally different understanding or impression of the expressive content" than if they were to hear or see the unedited image, audio, or video.[122] However, the statute permits distribution if the media constitutes parody or satire[123] or is labeled with the following message: "This [image, video, or audio] has been manipulated."[124] A candidate appearing in the manipulated media may seek injunctive relief to stop the distribution.[125] Texas has passed a similar provision that protects candidates in the lead-up to elections.[126] Neither of these statutes provides any protection to the common citizen against nonpornographic deepfakes, however. In contrast with the law of defamation—where public figures are disadvantaged compared to private figures[127]—here, only public figures are protected and only in a particular time frame.

---

[117] *Id.* § 1708.86(d).

[118] *Id.* § 1708.86(e)(1)(B)(i)–(ii).

[119] *Id.* § 1708.86(e)(1)(B)(ii)(I)–(II).

[120] *Id.* § 1708.86(e)(1)(C)–(E).

[121] CAL. ELEC. CODE § 20010(a) (West 2020).

[122] *Id.* § 20010(e)(1)–(2).

[123] *Id.* § 20010(d)(5).

[124] *Id.* § 20010(b)(1).

[125] *Id.* § 20010(c)(1).

[126] TEX. ELEC. CODE ANN. § 255.004(d) (West 2019).

[127] *See, e.g.*, N.Y. Times Co. v. Sullivan, 376 U.S. 254, 279–80 (1964) (requiring elevated *mens rea* for a person to be liable for defamation of a public figure).

One of the most recent state laws on deepfakes was passed in New York on November 30, 2020.[128] This action provided two new protections against deepfake videos. First, it expanded the New York right-of-publicity law to cover digitally manipulated likenesses and allow for protection to run for forty years after the depicted person's death. But this right-of-publicity statute, like most others, only applies to limited commercial uses. Specifically, it bars uses in advertising or on products.[129] This would cover very few current deepfakes, as most existing deepfakes are either satirical or pornographic, rather than commercial. The statute also provides limited protection against the use of unauthorized deepfakes in audiovisual works unless the works include a conspicuous disclaimer.[130]

The second form of new protection provided by New York is against pornographic deepfakes. These are prohibited in language similar to that of the new California statute: it is a violation to distribute unauthorized deepfakes of a person showing them "nude, meaning with an unclothed or exposed intimate part . . . or appearing to engage in, or being subjected to, sexual conduct."[131] This provision specifically says that a disclaimer saying the representation is fake is not a defense against liability.[132] Interestingly, this statute further provides that consent to appear in deepfake pornography is valid only if obtained through a rigorous process, with substantial notice to the subject and a right to revoke consent.[133]

Looking at the variations across these new deepfake laws gives a sense of the broad range of options that will confront legislatures over the next several years. Depending on which harms, and which victims, most concern a state, the state could ban deepfake pornography, deepfake election interference, deepfake commercial exploitation, or all three. This range of possibilities highlights the need to determine which deepfakes are viewed as morally wrong and practically harmful by the public. Part II begins to answer those questions.

---

[128] *Governor Cuomo Signs Legislation Establishing a "Right to Publicity" for Deceased Individuals to Protect Against the Commercial Exploitation of Their Name or Likeness*, N.Y. STATE (Nov. 30, 2020), https://www.governor.ny.gov/news/governor-cuomo-signs-legislation-establishing-right-publicity-deceased-individuals-protect [https://perma.cc/PAJ7-EG9A].

[129] N.Y. CIV. RIGHTS LAW § 50-f(2)(a) (McKinney 2021).

[130] *Id.* § 50-f(2)(b) (prohibiting use "in a scripted audiovisual work as a fictional character or for the live performance of a musical work . . . if the use is likely to deceive the public into thinking it was authorized by the person [or their representatives]" and clarifying that "[a] use shall not be considered likely to deceive the public . . . if the person making such use provides a conspicuous disclaimer in the credits").

[131] *Id.* § 52-c(1)(e).

[132] *Id.* § 52-c(2)(b).

[133] *Id.* § 52-c(3)(a)–(b).

II. THREE STUDIES OF DEEPFAKE ATTITUDES

Given the possibility of substantial future legislative activity in this area and the unsettled literature on deepfake harms, it is essential to better understand how the public views deepfakes. Are all deepfakes problematic, or only ones that are pornographic or depict certain kinds of conduct? Are deepfakes of all people problematic, or only ones of people who are not politicians and celebrities? One can easily see how pornographic deepfakes, or Nazi-promoting attitudinal deepfakes, can harm the dignity of those depicted. But not all deepfakes are of that sort. If someone creates a deepfake of the president doing Fortnite dances, is that similarly an affront to dignity? After all, Jordan Peele was not widely condemned for participating in the creation of a comedic deepfake of President Barack Obama.[134]

Further, American law places great faith in the marketplace of ideas. False claims about a person can lead to liability, but American law recognizes that public figures do not have a right to avoid being the subjects of satire, however little they may enjoy the experience.[135] Likewise, the publication of a publicly taken photograph of a person generally does not run afoul of state privacy laws.[136] Before creating what may amount to a new privacy right, we should first carefully mark the boundaries of what we seek to protect.

Very little is known about the attitudes of everyday people toward deepfakes. One nonacademic survey of an unrepresentative sample showed that people thought that deepfakes would do more harm than good and that a majority wanted to criminalize deepfakes.[137] Yet this study did not address any of the above questions about how different deepfakes would be

---

[134] Most reporting on this took a very matter-of-fact approach. For an example of a matter-of-fact tone used in reporting on the comedic deepfake of President Obama, see James Vincent, *Watch Jordan Peele Use AI to Make Barack Obama Deliver a PSA About Fake News*, VERGE (Apr. 17, 2018, 1:14 PM), https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed [https://perma.cc/QMA2-BQST]. The authors have not found any articles describing the creation as inappropriate.

[135] *See, e.g.*, Hustler Magazine, Inc. v. Falwell, 485 U.S. 46, 55–57 (1988) (holding that a public figure cannot sustain a claim of intentional infliction of emotional distress against the publisher of a parody depicting the plaintiff because the "outrageous" standard of conduct as applied to political cartoons would invite juries to impose their own "tastes or views" in violation of the First Amendment).

[136] *See, e.g.*, Gill v. Hearst Pub. Co., 253 P.2d 441, 444–45 (Cal. 1953) (holding that plaintiffs waived their right to privacy by "expos[ing] themselves to public gaze in a pose open to the view of any persons who might then be at or near" them, and therefore publication of their photograph did not invade their right of privacy).

[137] Toni Allen, *Dodging Deception & Seeking Truth Online [Survey Results]*, WHO IS HOSTING THIS (Sept. 19, 2019), https://www.whoishostingthis.com/blog/2019/09/02/seeking-trust-online/ [https://perma.cc/2LJN-D3UP]. The survey was conducted of 981 "internet users," from whom few demographics were reported. *Id.*

viewed.[138] It did not ask about differences between pornographic and nonpornographic deepfakes or bring up the idea of labeling deepfakes as fake—which appear to be the two main distinctions discussed by current legislative proposals. Therefore, it provides little guidance for future legislation.

To fill this gap and explore how everyday people view different kinds of deepfake videos, we conducted a study with a representative sample of the U.S. adult population. This Part discusses the design and methodology of the study, our sample, and findings from the study.

To conduct our primary study, we wrote scenarios that captured attitudes toward deepfake videos in the pornographic and nonpornographic contexts independently. Further, we wrote a range of scenarios for each context, sampling broadly from the universe of possible uses of deepfakes. One of our main goals was to determine if labeling the deepfake as fake mattered. The question of labeling is particularly important in this context because some proposals would only ban unlabeled deepfake videos.[139] Further, whether a deepfake video is labeled has implications for a victim's ability to seek redress under theories of defamation, false light, or intentional infliction of emotional distress.[140]

We also included scenarios that depicted the victims as either public figures or private individuals. Public figures are treated differently under various tort laws, and courts have provided substantial protection for speech concerning them.[141] The question remains whether the same considerations are consistent in the context of visual deepfake depictions.

For this study, a sample of American adults were recruited by Dynata, an online survey firm with an established panel of respondents.[142] The demographics of the sample were set to match the U.S. Census proportions on the dimensions of age, sex, region, education, race, and ethnicity. Full

---

[138] The study provided a brief description of deepfakes, saying that they were AI-produced videos depicting people saying or doing things that they did not say or do. It then asked, "Do you believe deepfaking someone without consent should be illegal or legal?" *Id.* The study does not appear to have provided subjects with any particular examples of deepfakes.

[139] *See, e.g.*, CAL. ELEC. CODE § 20010(b)(1) (West 2020) (providing no liability for labeled videos); Ruiz, *supra* note 15 (noting one federal bill would require a deepfake "watermark" label).

[140] *See supra* Section I.C.

[141] *See, e.g.*, Hustler Magazine, Inc. v. Falwell, 485 U.S. 46, 54–56 (1988) ("From the viewpoint of history it is clear that our political discourse would have been considerably poorer without [satirical cartoons]."); N.Y. Times Co. v. Sullivan, 376 U.S. 254, 279–81 (1964) ("The importance to the state and to society of [discussing the character and qualifications of candidates for their suffrages] is so vast, and the advantages derived are so great, that they more than counterbalance the inconvenience of private persons whose conduct may be involved . . . ." (quoting Coleman v. MacLennan, 98 P. 281, 286 (Kan. 1908))).

[142] DYNATA, PANEL BOOK 5–6 (2020).

demographics are reported in Appendix A. The final sample contained 1,141 individuals.[143] The study was conducted in October 2020 through Qualtrics. Respondents received an email from Dynata inviting them to participate in the survey. If they clicked on the provided link, then they were routed to a Qualtrics survey hosted by Northwestern University. By monitoring the demographics of those completing the survey, Dynata targeted waves of survey invitations to create a final sample consistent with the desired quotas.

This study had two basic parts. The first part presented participants with vignettes that described people making deepfake videos of various types. Participants were asked to rate these scenarios on several dimensions and decide whether it should be possible to criminally punish the person making the video. The purpose of using vignettes in this part was to introduce participants to deepfakes, a concept with which many of them might have been unfamiliar, and to give them examples of how deepfake technology could be used. This reduced the chance that participants would imagine drastically different conduct when thinking about deepfakes. The second part of the study asked a series of questions about the harmfulness of deepfakes, more generally, outside the context of a particular set of facts.

Study participants were randomly assigned to receive vignettes about one of four different types of deepfake videos: pornographic or attitudinal deepfakes that were either labeled as fake or not. The pornographic vignettes all included sexualized content, with the deepfake subject depicted either having sex or engaged in sexual behavior. By contrast, the deepfakes we called attitudinal incorporated a range of different contents—from the silly to the defamatory to the totally mundane. We termed these attitudinal because the key behavior in the videos was often expressive—the deepfake subject was made to convey attitudes or facts.

In addition to being pornographic or attitudinal in content, the videos were either labeled or unlabeled. Labeled videos were described as clearly identified as fake by the video maker. For unlabeled videos, in contrast, it was clearly stated that the video creator did not indicate the video was fake. The following was the default unlabeled pornographic deepfake scenario:

---

[143] Inattentive participants were screened from the final sample based on two criteria. First, participants who did not give the appropriate response to an attention-check question—a question asking participants to give a particular response—or a CAPTCHA item were unable to complete the study. Second, participants were screened from the final sample if they finished the study in less than one-third of the time taken by the median participant or if they wrote gibberish in a comment box. Of the participants who completed the study, 3.7% were screened on the basis of time or gibberish. For a discussion of attention checks in legal surveys, see Matthew B. Kugler & R. Charles Henn, *Internet Surveys in Trademark Cases: Benefits, Challenges, and Solutions*, *in* TRADEMARK AND DECEPTIVE ADVERTISING SURVEYS (Shari Seidman Diamond & Jerre B. Swann eds., 2d ed. forthcoming 2021).

> Imagine Jane is a friend of Will. Will finds a series of photos of Jane online. Will takes the photos and uses an app to merge her face onto a pornographic video. The final video shows Jane's face on the body of a naked woman having sex with a man. The video shows the entirety of the naked woman's body. Jane's face is clearly identifiable in the video. Will posts the video online publicly, and he includes Jane's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

The scenario makes it clear that the deepfake video used publicly available photos of the video subject, that it included graphic sex, that it looked genuine, and that it was posted publicly in a way that made it easily linked to the real identity of the video subject. The labeled version replaced the last sentence with, "In the video title and as a caption on the video, Will writes 'This is fake' to show that it is fake." This disclaimer was intended to be completely unambiguous and as permanent as any digital watermark could reasonably be. Each participant received only one type of vignette. For example, every vignette read by Participant A was about pornographic deepfakes that were labeled, and every vignette read by Participant B was about attitudinal deepfakes that were unlabeled. The full text of the unlabeled scenarios is available in Appendix B. In each case, the labeled version differed only in the last sentence, as in the above example.

Within each of these four conditions, participants rated multiple scenarios in a random order. For each scenario, the participant answered three questions:

(1) How morally blameworthy was the video maker's conduct (1: Not at All to 6: Very Much)?;
(2) How harmful was this to the deepfake video subject (same scale)?; and
(3) How, if at all, should it be possible to punish the person making the video?

This last question was answered on the following scale:

(1) It should not be possible to punish him; this should not be a crime;
(2) It should be punished with a fine (less than $500);
(3) It should be punished like a minor crime (a year or less in jail); and
(4) It should be punished like a major crime (up to 10 years in jail).

We will review the results for the pornographic deepfakes before turning to the attitudinal deepfakes and closing with the overall questions about deepfake harmfulness. Table 1 shows the full list of scenarios used in the study. Participants received either the pornographic or attitudinal scenarios (if attitudinal, they saw both "private" and "politician" videos) that

were either labeled as fake or not. In total, 283 participants received the unlabeled attitudinal scenarios, 281 the labeled attitudinal scenarios, 287 the unlabeled pornographic scenarios, and 290 the labeled pornographic scenarios.

TABLE 1: FULL LIST OF SCENARIOS USED IN THE STUDY

Type	Scenario
Pornographic	Written Pornographic Story, Friend
	Deepfake (DF) Pornographic Video, Friend (Default Condition)
	DF Pornographic Video, Celebrity
	DF Pornographic Video, Sexualized Voice
	DF Pornographic Video, No Nudity, BDSM
	DF Pornographic Video, Personal Use, No Consent, Friend
	DF Pornographic Video, Personal Use, Consent, Friend
Attitudinal, Private	Written Cocaine-Use Story
	DF Cocaine-Use Video
	DF Self-Insult
	DF Scientist Biography, Dead
	DF Scientist Biography, Living
Attitudinal, Politician	Written Handshake-with-Child-Molester Story
	DF Handshake-with-Child-Molester Video
	DF Terror Endorsement
	DF Silly Song, No Consent
	DF Silly Song, Consent
	DF Polling Place, No Consent
	DF Polling Place, Consent

Analyses for these results took the form of a series of Analysis of Variance (ANOVA) tests on each of the dependent measures. ANOVAs test whether scores from two or more samples differ systematically enough that the samples are likely to be statistically distinct. Comparisons across labeling condition, looking at the effect of labeled versus not, were between-subject because different people saw the labeled and unlabeled vignettes. Comparisons across different labeled scenarios—such as comparing the default pornography deepfake condition to several of the other pornographic variants—were within-subject: the same people rated each of the labeled pornographic scenarios. Most of the analyses that follow are therefore mixed ANOVAs. For example, the first analysis below is a mixed 2x2 ANOVA that looks at the difference between a pornographic deepfake video and a pornographic written story (within-subject comparison, the same people saw both) and the difference between those scenarios being labeled as fake or not labeled as fake (a between-subject comparison with different people seeing

each possibility), as well as their interaction term. So this ANOVA tests whether the written story is different than the deepfake video (the main effect of video), whether labeled stories or videos are different than unlabeled stories or videos (the main effect of labeling), and whether the effect of labeling differs for stories and videos (the interaction between labeling and video).

A. Impressions of Pornographic Deepfakes

The default deepfake pornographic condition—in which our protagonist makes a deepfake video of a female friend that depicts the friend having sexual intercourse with a man, without labeling it as fake, and posts the video online—was viewed as highly blameworthy, extremely harmful to the person depicted, and deserving of substantial punishment (see Table 2). The first analysis here contrasts the protagonist making a deepfake pornographic video about his friend with the protagonist creating a written story describing the same conduct. Though writing and posting a pornographic story featuring the same conduct was viewed as less blameworthy, harmful, and deserving of punishment,[144] that act was also rated as quite serious, with only 10.5% not wanting to punish it criminally accompanied by relatively high blameworthiness and harm scores (Table 2).

TABLE 2: COMPARISON OF DEEPFAKE PORNOGRAPHIC VIDEO TO WRITTEN STORY

		Unlabeled		**Labeled**	
Deepfake Pornographic Video, Friend	Blameworthy	5.44	(1.25)	5.36	(1.27)
	Harm	5.43	(1.20)	5.43	(1.14)
	Punishment	3.08	(0.94)	2.91	(0.92)
	Percentage not a crime	7.3%		8.0%	
Written Pornographic Story, Friend	Blameworthy	5.31	(1.33)	4.96	(1.54)
	Harm	5.29	(1.20)	5.08	(1.41)
	Punishment	2.78	(0.95)	2.45	(0.99)
	Percentage not a crime	10.5%		18.3%	

Note. Means (standard deviations in parentheses). Blameworthiness and harmfulness were rated on 6-point scales. Punishment was on a 4-point scale. The proportion of respondents choosing the lowest punishment option, "It should not be possible to punish him; this should not be a crime," is reported in the bottom row for each scenario.

[144] A 2x2 ANOVA test (video or written as a within-subjects factor, labeled versus not as a between-subjects factor) revealed a significant main effect for the content being a video on each of the three measures. Blameworthiness: $F(1, 571) = 23.36$, $p < 0.001$ $\eta^2 = 0.04$. Harm: $F(1, 571) = 22.24$, $p < 0.001$ $\eta^2 = 0.04$. Punishment: $F(1, 571) = 108.79$, $p < 0.001$ $\eta^2 = 0.16$.

The effect of labeling this story or video as fake depended on whether the content was written or a deepfake video.[145] Labeling helped significantly for the written story—causing participants to view it as less harmful, less wrongful, and deserving of less punishment—but mattered much less for the video. Labeling the video produced only a small significant effect on punishment, and that effect was one-third the size of the effect for the written story.[146] There was no significant effect of labeling on the perceived harmfulness or blameworthiness of the video.

The remaining deepfake pornographic cases were then compared to this default friend deepfake video case (see Table 3).[147] In one, the deepfake was of a celebrity rather than a friend. Everything else was the same: the video was still posted online and still clearly identified the celebrity. Targeting a celebrity rather than a friend was viewed as mitigating on each of the three dependent measures, but only very slightly. A full 90.2% of the sample still wanted to criminalize this conduct in the unlabeled condition. Two other variants that included sexualized behavior but no nudity—spanking in one and seductive speaking in the other—were also viewed only slightly more leniently than the default case.

[145] The mixed ANOVA tests revealed an interaction effect between labeling and content type. Blameworthiness: $F(1, 571) = 6.54$, $p < 0.05$ $\eta^2 = 0.01$. Harm: $F(1, 571) = 4.30$, $p < 0.05$ $\eta^2 = 0.01$. Punishment: $F(1, 571) = 5.03$, $p < 0.05$ $\eta^2 = 0.01$.

[146] A simple effects analysis looking at the effect of labeling for the written and video scenarios separately revealed significant effects of labeling on the written scenario: $F(1, 571) = 8.61$, $p < 0.001$ $\eta^2 = 0.02$. Harm: $F(1, 571) = 3.67$, $p = 0.05$ $\eta^2 = 0.01$. Punishment: $F(1, 571) = 16.58$, $p < 0.001$ $\eta^2 = 0.03$. But only a significant effect on punishment for the video: $F(1, 571) = 0.55$, ns. Harm: $F(1, 571) = 0.00$, ns. Punishment: $F(1, 571) = 4.65$, $p < 0.05$ $\eta^2 = 0.01$.

[147] This was a series of mixed ANOVA tests with labeling as a between-subjects factor and the type of scenario (default versus celebrity; default versus no nudity, BDSM; default versus sexualized voice) as a within-subjects factor. Table 3's "Comparison with Default" column reports the F-values of the within-subjects scenario factor.

TABLE 3: COMPARISON OF VARIANTS TO DEFAULT DEEPFAKE PORNOGRAPHIC VIDEO

		Unlabeled		Labeled		Compared to Default Condition (Collapsing Across Labeling Categories)
Deepfake Pornographic Video, Celebrity	Blameworthy	5.30	(1.37)	5.31	(1.36)	$F(1, 571) = 4.51^{*}\ \eta^2 = 0.01$
	Harm	5.27	(1.29)	5.21	(1.36)	$F(1, 571) = 15.58^{***}\ \eta^2 = 0.03$
	Punishment	2.94	(0.99)	2.89	(0.91)	$F(1, 571) = 6.18^{*}\ \eta^2 = 0.01$
	Pct. not a crime	9.8%		9.0%		
Deepfake Pornographic Video, No Nudity, BDSM	Blameworthy	5.35	(1.27)	5.32	(1.24)	$F(1, 571) = 2.06\ \eta^2 = 0.00$
	Harm	5.35	(1.20)	5.22	(1.26)	$F(1, 571) = 17.07^{***}\ \eta^2 = 0.03$
	Punishment	2.85	(0.93)	2.63	(0.89)	$F(1, 571) = 56.62^{***}\ \eta^2 = 0.09$
	Pct. not a crime	8.4%		10.0%		
Deepfake Pornographic Video, Sexualized Voice	Blameworthy	5.29	(1.39)	5.26	(1.34)	$F(1, 570) = 6.86^{**}\ \eta^2 = 0.01$
	Harm	5.21	(1.30)	5.19	(1.34)	$F(1, 570) = 24.31^{***}\ \eta^2 = 0.04$
	Punishment	2.86	(0.97)	2.60	(0.93)	$F(1, 570) = 58.02^{***}\ \eta^2 = 0.09$
	Pct. not a crime	10.1%		12.8%		
Deepfake Pornographic Video, Personal Use, No Consent	Blameworthy	5.47	(1.07)	5.01	(1.56)	$F(1, 270) = 4.77^{*}\ \eta^2 = 0.02$
	Harm	5.14	(1.37)	4.71	(1.67)	$F(1, 270) = 29.22^{***}\ \eta^2 = 0.10$
	Punishment	2.77	(1.04)	2.48	(1.09)	$F(1, 270) = 41.68^{***}\ \eta^2 = 0.13$
	Pct. not a crime	15.1%		23.8%		
						Compared to No Consent
Deepfake Pornographic Video, Personal Use, Consent[148]	Blameworthy	3.86	(2.13)	3.86	(2.09)	$F(1, 567) = 85.88^{***}\ \eta^2 = 0.13$
	Harm	3.78	(2.05)	3.91	(2.01)	$F(1, 567) = 50.85^{***}\ \eta^2 = 0.08$
	Punishment	2.08	(1.13)	1.96	(1.13)	$F(1, 567) = 43.91^{***}\ \eta^2 = 0.07$
	Pct. not a crime	43.6%		51.2%		

Note. Means (standard deviations in parentheses). Blameworthiness and harmfulness were rated on 6-point scales. Punishment was on a 4-point scale. Statistical significance is indicated as * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The proportion of respondents choosing the lowest punishment option, "It should not be possible to punish him; this should not be a crime," is reported in the bottom row for each scenario.[149]

These two no-nudity scenarios address a question that arises under the current California statute on pornographic deepfakes. This statute prohibits videos depicting individuals who are "nude" or engaging in "sexual conduct."[150] Sexual conduct is in turn described as masturbation, several

[148] This analysis was between-participants, as each person got either the personal-use-with-consent or personal-use-without-consent scenario.

[149] Due to incomplete data for a few participants, not all comparisons have the same N. This did not affect the means for the comparison deepfake case by more than two one-hundredths for any comparison except the personal-use case, which was only shown to half the sample. For that analysis, the means for the default case were: Blameworthiness unlabeled (M = 5.53, SD = 1.10), labeled (M = 5.27, SD = 1.36); Harm unlabeled (M = 5.53, SD = 1.07), labeled (M = 5.26, SD = 1.34); Punishment unlabeled (M = 3.14, SD = 0.94), labeled (M = 2.87, SD = 0.95).

[150] CAL. CIV. CODE § 1708.86(a)(14) (West 2020) (defining sexually explicit material).

different kinds of sexual intercourse, sexual penetration of the vagina or rectum, ejaculation on a person, and "[s]adomasochistic abuse involving the depicted individual."[151] A spanking scene would likely qualify under this last prong, despite the lack of penetration or nudity. The sexualized-voice scene does not depict the speaker engaging in any of those forms of sexual conduct, and therefore would be outside the scope of the statute. Participants, however, viewed all of these as equivalently problematic. Though there are slight statistical differences between these and the default scenario, they are quite small. All of the scenarios received blameworthiness and harm ratings of above 5 on a 6-point scale. All earned criminalization ratings of above 85%.

The largest difference in preference for punishment, across all these pornographic scenarios, was for the final scenario: where the maker of the deepfake did not distribute it but instead kept it for his own personal use. But that was still criminalized by 84.9% of respondents in the unlabeled nonconsensual case and viewed as extremely blameworthy and harmful. This undistributed creation would *not* fall within the scope of the California or New York statutes, as they target only the disclosure of deepfake videos.[152]

In an additional wrinkle, half of the participants evaluating this personal-use variant were presented with a version in which the maker of the deepfake asked for and received the consent of the deepfake subject. The other half was presented with a version in which the deepfake subject was not asked for consent, consistent with the other pornographic scenarios. This consent manipulation mattered a great deal. Ratings on all three measures were significantly lower in the consent condition than in the condition where consent was not mentioned (and the video was still unpublished): 43.6% of participants in the unlabeled condition and 51.2% of participants in the labeled condition did not seek to criminalize or punish this conduct when consent was obtained (Table 3). Further, the distribution of blameworthiness responses was markedly different here than in the other conditions. In the default pornographic deepfake condition, only 4.3% chose the lowest blameworthiness option. In the nonconsensual personal-use condition, 4.0% chose that option. In the consensual personal-use condition, the distribution is bimodal: 28.6% chose the lowest option, indicating that they believed the protagonist did not do something morally wrong, and 38.9% chose the worst option, with the remainder irregularly scattered between.

As discussed in Section I.C, the law of defamation would have little difficulty punishing a statement that was false, looked as if it were meant to

[151] *Id.* § 1708.86(a)(13) (defining sexual conduct).

[152] *Id.* § 1708.86(b)(1) (creating a civil cause of action against anyone who "[c]reates *and* intentionally discloses" (emphasis added)); N.Y. CIV. RIGHTS LAW § 52-c(2)(a) (McKinney 2021).

be taken as true, and caused harm to a person's reputation. Labeling that account as false would generally prevent liability, however. But this kind of labeling does not have much effect on the perceived blameworthiness and harmfulness of pornographic deepfake videos. Across all scenarios, labeling mattered very little. In the four main variants (default friend, celebrity, spanking, and speaking), there were no significant effects on labeling in the analysis on harm or blameworthiness, and only an inconsistent mitigation effect on punishment.[153]

Overall, then, people view the pornographic deepfake scenarios as extremely blameworthy, harmful, and deserving of punishment. The written stories, especially the written story labeled as fiction, are viewed more leniently on each dimension than the videos. People still find them troubling, however. Among the deepfake videos, three of the four variants (celebrity, spanking, and sexualized voice) were barely different than the baseline scenario in which the actor made a pornographic deepfake of a friend. Making the victim a celebrity did not have a substantial mitigating effect, nor did the two variants that excluded nudity but included sexualized content. Also, across all of these scenarios, labeling only intermittently mattered. Even deepfakes labeled as deepfakes were viewed as blameworthy, harmful, and deserving of punishment.

B. Impressions of Attitudinal Deepfakes

In addition to the pornographic deepfake scenarios, we also asked about attitudinal scenarios. These varied greatly in content. Some depicted the deepfake subject doing something morally questionable, some of them doing something silly, and some neither. None included sex or sexualized conduct, however.

The main scenarios here depicted an everyday person or a politician doing something morally blameworthy. The everyday person, described as a friend, was depicted as doing cocaine. The politician was depicted as shaking hands with a convicted child molester. Again, our first analysis here contrasts the deepfake videos with written stories describing the same content (Table 4). Two major patterns emerged. First, the videos were significantly worse

---

[153] See *supra* note 146 for the results labeling had on the default friend condition. In the celebrity condition, labeling had no effect on blameworthiness, $F(1, 573) = 0.01$, ns $\eta^2 = 0.00$; harm $F(1, 573) = 0.28$ ns $\eta^2 = 0.00$; or punishment $F(1, 573) = 0.47$, ns $\eta^2 = 0.00$.

In the no-nudity, BDSM condition, labeling had no effect on blameworthiness $F(1, 572) = 0.05$, ns $\eta^2 = 0.00$, or harm $F(1, 572) = 0.02$, ns $\eta^2 = 0.00$, but there was an effect on punishment such that labeling led to lower punishments $F(1, 572) = 11.00$, $p < 0.001$ $\eta^2 = 0.02$.

In the sexualized-voice condition, labeling had no effect on blameworthiness $F(1, 574) = 0.07$, ns $\eta^2 = 0.00$, or harm $F(1, 574) = 1.79$, ns $\eta^2 = 0.00$, but there was an effect on punishment $F(1, 574) = 8.47$, $p < 0.001$ $\eta^2 = 0.02$.

on blameworthiness, harm, and punishment than the written stories regardless of whether they were labeled.[154] Second, and in contrast to the pornographic scenarios, here, there was a significant labeling effect on each of the three dependent measures, with labeling lowering the severity on each for both written and video variants.[155]

TABLE 4: REACTIONS TO MAIN ATTITUDINAL SCENARIOS

			Unlabeled		**Labeled**	
Private, Cocaine Use	Video	Blameworthy	5.05	(1.48)	4.83	(1.46)
		Harm	5.14	(1.35)	4.92	(1.36)
		Punishment	2.73	(0.95)	2.44	(0.94)
		Percentage not a crime	12.0%		16.0%	
	Written	Blameworthy	5.03	(1.37)	4.64	(1.52)
		Harm	5.11	(1.32)	4.69	(1.43)
		Punishment	2.58	(0.98)	2.23	(0.93)
		Percentage not a crime	16.6%		24.7%	
Politician, Handshake with Child Molester	Video	Blameworthy	4.93	(1.58)	4.71	(1.49)
		Harm	5.08	(1.34)	4.77	(1.43)
		Punishment	2.66	(1.00)	2.34	(0.97)
		Percentage not a crime	14.6%		21.8%	
	Written	Blameworthy	4.98	(1.52)	4.58	(1.56)
		Harm	5.03	(1.43)	4.71	(1.45)
		Punishment	2.66	(1.00)	2.29	(0.94)
		Percentage not a crime	16.3%		21.4%	

Note. Means (standard deviations in parentheses). Blameworthiness and harmfulness were rated on 6-point scales. Punishment was on a 4-point scale. The proportion of respondents choosing the lowest punishment option, "It should not be possible to punish him; this should not be a crime," is reported in the bottom row for each scenario.

There were very few other significant effects in this first analysis. It was slightly less blameworthy to write a story about or make a deepfake of a politician than an everyday person; though, here, whether the person was a

[154] The analyses took the form of mixed ANOVA tests with labeling as a between-subjects factor and politician (versus person) and video (versus written) as within-subjects factors. There were significant effects on each of the three dependent variables for whether the content was a deepfake video. Blameworthiness: $F(1, 555) = 4.05$, $p < 0.05$ $\eta^2 = 0.01$. Harm: $F(1, 555) = 6.28$, $p < 0.05$ $\eta^2 = 0.01$. Punishment: $F(1, 555) = 18.05$, $p < 0.001$ $\eta^2 = 0.03$.

[155] Blameworthiness: $F(1, 555) = 7.58$, $p < 0.01$ $\eta^2 = 0.01$. Harm: $F(1, 555) = 9.89$, $p < 0.01$ $\eta^2 = 0.02$. Punishment: $F(1, 555) = 23.75$, $p < 0.001$ $\eta^2 = 0.04$. There was an interaction effect, by which labeling reduced blameworthiness more for written content, though labeling was also significant for video. Interaction: $F(1, 555) = 5.76$, $p < 0.05$ $\eta^2 = 0.01$. Written: $F(1, 557) = 13.05$, $p < 0.001$ $\eta^2 = 0.02$. Video: $F(1, 555) = 3.87$, $p = 0.05$ $\eta^2 = 0.01$. The interactions on harm and punishment were not significant.

politician was confounded with the type of morally questionable conduct depicted.[156] Whether the content was video or written mattered less for punishment in the politician case than it did for the everyday person, though the base rate was high: more than 85% of people wanted to criminalize the unlabeled politician video.[157]

Two additional scenarios concerned everyday people. In one, our protagonist makes a deepfake of his friend calling herself a jerk.[158] This self-insult variant was viewed as less blameworthy, less harmful, and deserving of less punishment than the default cocaine scenario but was still generally criminalized (see Table 5).[159] Comparing labeled and unlabeled self-insult condition, labeling again helped.[160]

The second everyday-person scenario described our protagonist running a science-enthusiast website. As part of this website, they created a video of a scientist describing their own life and accomplishments. This was intended to push the boundaries of deepfake harm by making the video as inoffensive as possible. Though this was viewed as less problematic on each measure than the default cocaine video,[161] most people still sought to criminalize it (see Table 5). Comparing labeled and unlabeled scientist condition, labeling again helped.[162] In a further variant, the scientist in question was either described as having died ten years earlier or having just recently retired; participants in the attitudinal condition saw one variant or the other of this vignette. This was intended to keep constant the approximate recency of the scientist—the scientist is not Newton or Einstein and also not still active—while manipulating whether the scientist is still alive, a factor

---

[156] $F(1, 555) = 6.70$, $p = 0.01$ $\eta^2 = 0.01$.

[157] Interaction $F(1, 555) = 10.78$, $p < 0.001$ $\eta^2 = 0.02$. Politician $F(1, 557) = 0.49$, ns. Person: $F(1, 555) = 31.76$, $p < 0.001$ $\eta^2 = 0.05$.

[158] This was inspired by a scene in *Scrubs*. In that scene, the protagonist fantasizes about a recently met and annoying character saying, "I'm a tool. I'm a tool. I'm a tool, tool, tool, an unbelievably annoying tool." *Scrubs: My First Day* (ABC television broadcast Oct. 2, 2001) (transcript available at https://scrubs.fandom.com/wiki/My_First_Day_transcript [https://perma.cc/MJ8G-CFE4]).

[159] Mixed ANOVA tests were conducted with the cocaine and self-insult vignettes as within-subjects factors and labeling as a between-subjects factor. There was a significant effect of scenario on each of the three measures. Blameworthy: $F(1, 560) = 28.79$, $p < 0.001$ $\eta^2 = 0.049$. Harm: $F(1, 560) = 66.31$, $p < 0.001$ $\eta^2 = 0.106$. Punishment: $F(1, 560) = 74.13$, $p < 0.001$ $\eta^2 = 0.117$.

[160] Blameworthy: $F(1, 560) = 6.52$, $p < 0.05$ $\eta^2 = 0.012$. Harm: $F(1, 560) = 7.02$, $p < 0.01$ $\eta^2 = 0.012$. Punishment: $F(1, 560) = 17.59$, $p < 0.001$ $\eta^2 = 0.03$.

[161] Mixed ANOVA tests were conducted with the cocaine and self-insult vignettes as within-subjects factors and labeling as a between-subjects factor. There was a significant main effect of scenario on each of the three measures. Blameworthy: $F(1, 561) = 44.57$, $p < 0.001$ $\eta^2 = 0.074$. Harm: $F(1, 561) = 132.97$, $p < 0.001$ $\eta^2 = 0.192$. Punishment: $F(1, 561) = 95.64$, $p < 0.001$ $\eta^2 = 0.146$.

[162] Blameworthy: $F(1, 561) = 7.47$, $p <).01$ $\eta^2 = 0.013$. Harm: $F(1, 561) = 6.64$, $p < 0.05$ $\eta^2 = 0.012$. Punishment: $F(1, 561) = 11.34$, $p < 0.001$ $\eta^2 = 0.02$.

which matters a great deal elsewhere in privacy law.[163] This manipulation had no effect on any measure, so the Table 5 analysis combines these two conditions.[164]

TABLE 5: REACTIONS TO FURTHER NONCONSENSUAL DEEPFAKES

		Unlabeled		Labeled	
Self-Insult	Blameworthy	4.82	(1.49)	4.44	(1.61)
	Harm	4.70	(1.43)	4.36	(1.56)
	Punishment	2.41	(0.97)	2.11	(0.92)
	Percentage not a crime	19.9%		28.5%	
Scientist (Living and Dead Combined)	Blameworthy	4.70	(1.54)	4.31	(1.69)
	Harm	4.35	(1.61)	4.04	(1.75)
	Punishment	2.29	(1.00)	2.11	(0.97)
	Percentage not a crime	25.5%		31.7%	
Politician, Terror Endorsement	Blameworthy	5.06	(1.48)	4.74	(1.54)
	Harm	5.12	(1.34)	4.87	(1.38)
	Punishment	2.80	(1.02)	2.48	(0.99)
	Percentage not a crime	13.8%		18.1%	

Note. Means (standard deviations in parentheses). Blameworthiness and harmfulness were rated on 6-point scales. Punishment was on a 4-point scale. The percent choosing the lowest punishment option, "It should not be possible to punish him; this should not be a crime," is reported in the bottom row for each scenario.

For everyday people, then, there is substantial opposition to deepfakes of all sorts. There is substantially less opposition to more satirical deepfakes than ones depicting obviously blameworthy conduct (illegal drug use), and this difference is further helped by labeling. But most participants still sought to prohibit even the most harmless deepfake that we could design—the scientist saying true facts about their own life.

Because of the emphasis in the literature on election interference, and the California and Texas statutes specifically targeting deepfakes of politicians,[165] we included a second scenario that misrepresented a politician behaving in a morally questionable way. In this vignette, the politician is

[163] *See, e.g.*, Natalie M. Banta, *Death and Privacy in the Digital Age*, 94 N.C. L. REV. 927, 935 (2016) ("[P]rivacy torts are generally available to protect privacy during life; however, the majority of courts do not allow this right to extend beyond death." (footnotes omitted)); Warren v. Colvin, 744 F.3d 841, 843–44 (2d Cir. 2014) ("[Plaintiff] correctly asserts that deceased individuals generally do not enjoy rights under the Privacy Act.").

[164] A between-subjects ANOVA using alive versus dead and labeled versus not as factors found no significant effect of whether the scientist was alive on any measure. Blameworthy: $F(1, 559) = 2.79$, ns. Harm: $F(1, 559) = 1.03$, ns. Punishment: $F(1, 559) = 0.12$, ns.

[165] *See supra* note 10; *supra* notes 121–126 and accompanying text.

represented as "saying they support a known terrorist organization."[166] Compared to the politician-handshake case, this was viewed as more morally blameworthy and deserving of more punishment, but not significantly more harmful (see Table 5).[167] Comparing labeled and unlabeled politician endorsement condition, labeling again helped.[168]

Finally, there were two additional politician scenarios. As with the scientist, here we sought to present less overtly harmful fact patterns. For one, the video of the politician told people where their local polling places were. The study was conducted in advance of the 2020 general election, so this scenario was timely. The other scenario was of a politician singing an unspecified silly song.

In each case, we manipulated whether the politician had given general consent to produce the deepfake. For the polling-place scenario, the consensual vignette began, "Imagine a state-level politician has invited her constituents to make and share videos of her telling people the location of their local polling place." The consensual version of the silly song vignette was similar, "Imagine a state-level politician has invited her constituents to make and share silly videos of her for her campaign." Each participant in an attitudinal condition received either the consensual or nonconsensual version of each of these vignettes.

This consent was not as explicit and specific as it could have been. In general, one could easily imagine a politician consenting to have their image used in personalized get-out-the-vote messaging. Former President Barack Obama, for instance, phone-banked on behalf of Joseph Biden in the 2020 general election.[169] It would not be that great a stretch to imagine him working with the national party committee to produce personalized messages. A former president, however, likely would have been leery of

---

[166] If this scenario seems extreme, recall that Representative Peter King (R-N.Y.) endorsed the Irish Republican Army. In 1985, he said: "If civilians are killed in an attack on a military installation, it is certainly regrettable, but I will not morally blame the I.R.A. for it." Elspeth Reeve, *Peter King Supported the IRA Before Hunting for Terrorists*, ATLANTIC (Mar. 9, 2011) https://www.theatlantic.com/politics/archive/2011/03/peter-king-loved-terrorism-when-it-was-done-irish-people/348691/ [https://perma.cc/9DSZ-HLW6].

[167] Mixed ANOVA tests were conducted with the handshake and terror vignettes as within-subjects factors and labeling as a between-subjects factor. There was a significant effect of scenario on two of the measures, and a nonsignificant trend on perceived harmfulness. Blameworthy: $F(1, 557) = 4.15$, $p < 0.05$ $\eta^2 = 0.007$. Harm: $F(1, 557) = 3.34$, $p = 0.07$ $\eta^2 = 0.006$. Punishment: $F(1, 557) = 16.93$, $p < 0.001$ $\eta^2 = 0.03$.

[168] Blameworthy: $F(1, 557) = 5.61$, $p < 0.05$ $\eta^2 = 0.01$. Harm: $F(1, 557) = 7.61$, $p < 0.01$ $\eta^2 = 0.013$. Punishment: $F(1, 557) = 18.33$, $p < 0.001$ $\eta^2 = 0.032$.

[169] Sirena Bergman, *Voter Shares Adorable Video of Obama Chatting to Her New Baby on the Phone While Canvassing for Biden*, INDY100 (Nov. 1, 2020, 2:45 PM). https://www.indy100.com/article/obama-phone-banking-biden-viral-video-pennsylvania-election-9724055 [https://perma.cc/M3HF-QKMB].

granting their supporters as broad a license to make deepfake videos as did our hypothetical politician. The president would presumably want some editorial control to ensure quality and appropriateness. Here, we glossed over that issue.

TABLE 6: REACTIONS TO CONSENSUAL ATTITUDINAL POLITICIAN DEEPFAKES

		Unlabeled		Labeled	
Polling Place, No Consent	Blameworthy	4.49	(1.66)	4.13	(1.77)
	Harm	4.26	(1.74)	3.92	(1.80)
	Punishment	2.39	(1.08)	2.07	(0.94)
	Percentage not a crime	26.1%		31.6%	
Polling Place, Consent	Blameworthy	4.05	(1.84)	3.48	(1.83)
	Harm	3.78	(1.87)	3.32	(1.85)
	Punishment	2.07	(1.12)	1.84	(1.00)
	Percentage not a crime	43.4%		50.0%	
Silly Song, No Consent	Blameworthy	4.65	(1.67)	4.24	(1.73)
	Harm	4.41	(1.63)	3.91	(1.78)
	Punishment	2.23	(0.99)	2.09	(1.02)
	Percentage not a crime	27.8%		34.4%	
Silly Song, Consent	Blameworthy	3.74	(1.90)	3.57	(1.83)
	Harm	3.83	(1.84)	3.54	(1.77)
	Punishment	2.01	(1.05)	1.90	(0.95)
	Percentage not a crime	42.4%		43.6%	

Note. Means (standard deviations in parentheses). Blameworthiness and harmfulness were rated on 6-point scales. Punishment was on a 4-point scale. The percent choosing the lowest punishment option, "It should not be possible to punish him; this should not be a crime," is reported in the bottom row for each scenario.

As can be seen in Table 6, consent greatly reduced the perceived wrongfulness and harmfulness, as well as the desire to punish, for both scenarios.[170] Labeling was somewhat effective at alleviating concerns in the polling-place scenario, though the effect was not significant on every

[170] Separate ANOVA tests were conducted for the polling-place and silly-song vignettes with the same design. Both consent and labeling were between-subjects factors. For each, there was a strong effect of consent.

Polling place: Blameworthy: $F(1, 556) = 13.32$, $p < 0.001$ $\eta^2 = 0.023$. Harm: $F(1, 556) = 12.18$, $p < 0.001$ $\eta^2 = 0.021$. Punishment: $F(1, 556) = 9.91$, $p < 0.01$ $\eta^2 = 0.018$.

Silly Song: Blameworthy: $F(1, 559) = 27.09$, $p < .001$ $\eta^2 = 0.023$. Harm: $F(1, 559) = 10.27$, $p < 0.001$ $\eta^2 = 0.021$. Punishment: $F(1, 559) = 5.80$, $p < 0.05$ $\eta^2 = 0.018$.

measure for the silly-song scenario.[171] Nevertheless, people were still often willing to criminalize these deepfakes.

As with the consensual personal-use scenario, the consensual voting-announcement and silly-song videos also increased the proportion of people viewing the deepfakes as not wrongful at all. The consensual voting announcement was viewed as minimally blameworthy by 18.6% of respondents (11.4% for nonconsensual), and the consensual song video by 20.1% (9.7% for nonconsensual).

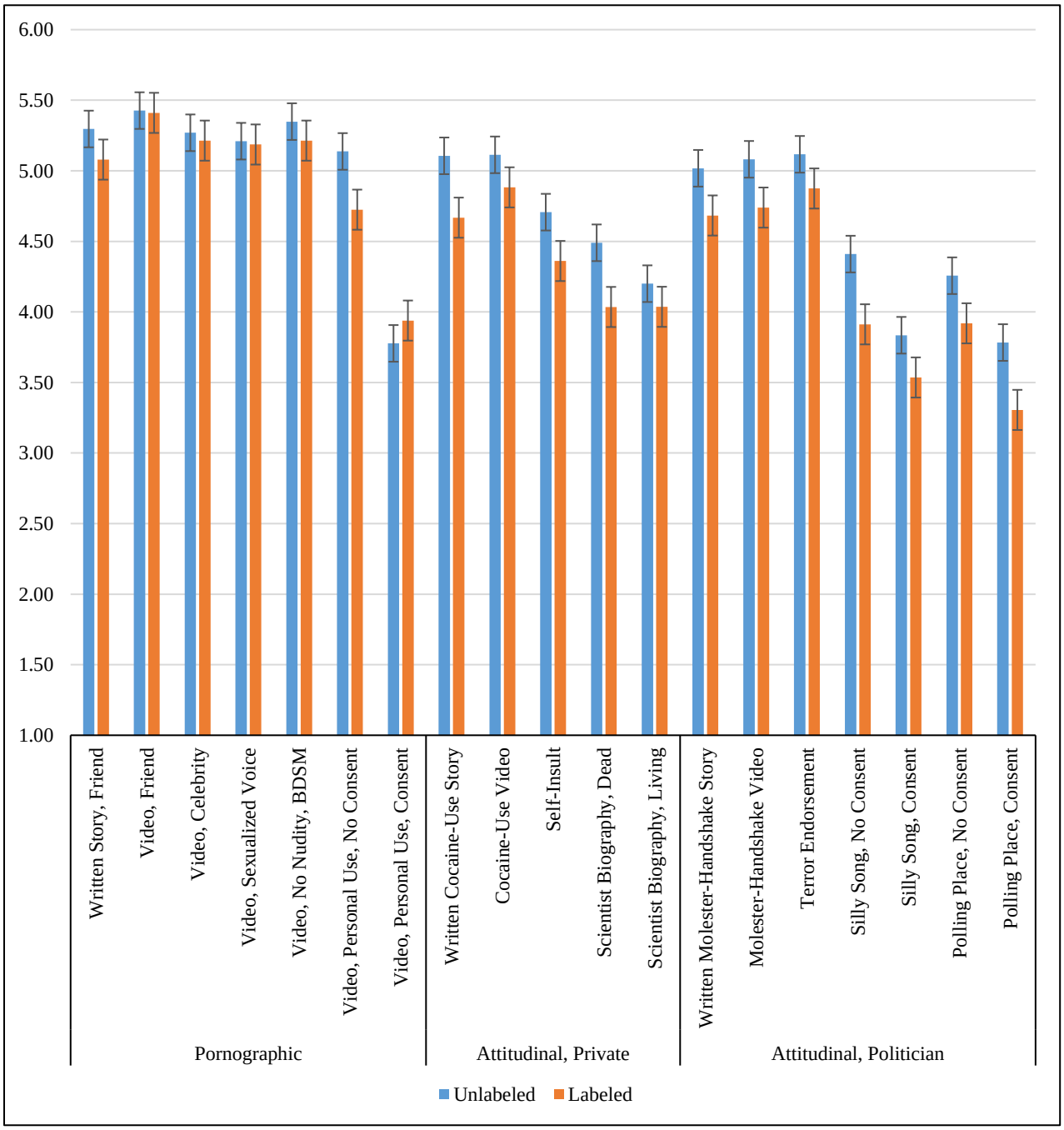
Figure 1 summarizes the main cross-scenario differences by showing the perceived harmfulness of each. The overall differences are stark. The consensual scenarios attract much lower harmfulness scores, and the nonconsensual pornographic videos attract particularly high scores. Attitudinal deepfakes worry a great many people, but this worry is reduced in the cases that are more satirical or somewhat harmless and by labeling. Pornographic deepfakes, however, are seen as very harmful by almost everyone. Labeling has a minimal effect—generally no effect—and no amount of variation in the scenarios matters much, even the ones that did not depict nudity.

The role of consent in these scenarios is somewhat unexpected. Consent always helped substantially, but it did not reduce the perceived harmfulness to nothing. There could be many reasons for this. For one, perhaps participants were not clear on the scope of consent—did the deepfake subject truly understand and agree to what actually happened? We comment further on the psychology of consent in this context in Part III.

[171] Polling place: Blameworthy: $F(1, 556) = 9.80$, $p < 0.01$ $\eta^2 = 0.017$. Harm: $F(1, 556) = 6.77$, $p < 0.05$ $\eta^2 = 0.012$. Punishment: $F(1, 556) = 10.16$, $p < 0.01$ $\eta^2 = 0.018$.

Silly Song: Blameworthy: $F(1, 559) = 3.60$, $p = 0.06$ $\eta^2 = 0.017$. Harm: $F(1, 559) = 7.20$, $p < 0.01$ $\eta^2 = 0.012$. Punishment: $F(1, 559) = 2.11$ $p = 0.15$ $\eta^2 = 0.018$.

FIGURE 1: PERCEIVED HARMFULNESS OF EACH TYPE OF DEEPFAKE



Note. Error bars represent standard errors. Harmfulness ratings are on a 1–6 scale.

C. Views on Deepfake Policies and Gender

Following the vignettes, these same participants were asked a series of policy-style questions. These questions explicitly defined deepfake videos and asked participants to think about the kinds of deepfake videos discussed in the scenarios they just read.[172] For example, in the unlabeled pornographic condition, participants were told:

> Think about **pornographic deepfake videos** that show people saying and doing things they did not say or do. These are the types of videos referred to earlier in the study. **So these are videos that include people nude, having sex, or engaged in sexual activities**. How harmful do you think this kind of video is **if the viewers think the video is real?**

Given that participants had just finished working through the scenarios reported in the preceding section, it was likely that these instructions were interpreted in terms of the use cases they had read.

The first question asked participants to make an overall assessment of harm for deepfake videos in their category on a 0–100 scale. As can be seen in Figure 2, pornographic videos were viewed as significantly more harmful than attitudinal videos; additionally, labeled videos—videos the viewer would know were false—were less harmful than unlabeled ones.[173] There was also a marginally significant interaction between attitudinal versus pornographic and labeling.[174] Consistent with the scenario results, labeling reduced perceived harmfulness more for the attitudinal scenarios.[175]

---

[172] The following definition was used:

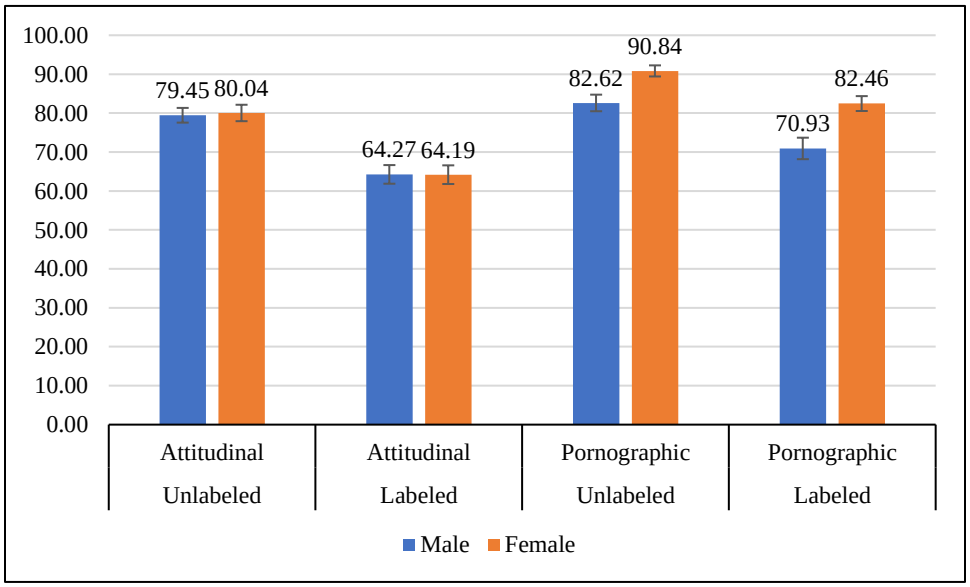
> A deepfake video is a realistic-looking video that has been edited to depict someone saying or doing something they never said or did. In a deepfake video, a person from one photo or video is inserted into another video. These videos can imitate people's faces and voices so well that they look and sound real.

[173] ANOVA tests were conducted looking at the factors pornographic versus attitudinal, labeled versus unlabeled, and male versus female. There were significant main effects for labeled, $F(1, 1111) = 71.54$, $p < 0.001$ $\eta^2 = 0.061$, pornographic, $F(1, 1111) = 41.44$, $p < 0.001$ $\eta^2 = 0.036$, and gender $F(1, 1111) = 11.26$, $p < 0.001$ $\eta^2 = 0.01$.

[174] $F(1, 1111) = 3.30$. $p = 07$ $\eta^2 = 0.003$.

[175] Attitudinal $F(1, 548) = 49.27$, $p < 0.001$ $\eta^2 = 0.082$. Pornographic: $F(1, 563) = 23.68$, $p < 0.001$ $\eta^2 = 0.040$.

FIGURE 2: PERCEPTIONS OF HARM FOR EACH TYPE OF DEEPFAKE SCENARIO BY GENDER



Note. Bars represent scores on a 0–100 scale. Error bars are standard errors.

There was also a significant effect of gender—women thought that deepfake videos were more harmful—but this was entirely driven by the pornographic deepfakes; there was a gender effect in the pornographic condition but not the attitudinal.[176] This gender pattern was also observed in the main pornographic and attitudinal scenarios. The female participants viewed the baseline pornographic scenario as more blameworthy, harmful, and deserving of punishment than the male participants did. However, there were no significant effects of gender for the baseline attitudinal scenario.[177] Previous research has observed that support for criminalizing nonconsensual

[176] There was a significant interaction between gender and pornographic versus attitudinal. $F(1, 1111) = 10.14$, $p < 0.001$ $\eta^2 = 0.009$. A simple effects analysis revealed that there was a significant effect of gender for the pornographic conditions, $F(1, 563) = 22.96$, $p < 0.001$ $\eta^2 = 0.039$, but not for the attitudinal conditions, $F(1, 548) = 0.01$ ns.

[177] Pornographic: Blameworthy: $F(1, 571) = 10.24$, $p < 0.001$ $\eta^2 = 0.018$, Male ($M = 5.21$, $SD = 1.40$), Female ($M = 5.55$, $SD = 1.11$). Harm: $F(1, 571) = 13.71$, $p < 0.001$ $\eta^2 = 0.023$, Male ($M = 5.23$, $SD = 1.32$), Female ($M = 5.59$, $SD = 1.00$). Punishment: $F(1, 571) = 9.05$, $p < 0.01$ $\eta^2 = 0.016$, Male ($M = 2.86$, $SD = 0.97$), Female ($M = 3.10$, $SD = 0.89$).

Attitudinal (cocaine): Blameworthy: $F(1, 559) = 1.36$ ns, Male ($M = 4.77$, $SD = 1.46$), Female ($M = 4.91$, $SD = 1.46$). Harm: $F(1, 559) = 2.70.101$ ns, Male ($M = 4.81$, $SD = 1.39$), Female ($M = 5.00$, $SD = 1.37$). Punishment: $F(1, 559) = 0.570.45$ ns, Male ($M = 2.37$, $SD = 1.01$), Female ($M = 2.43$, $SD = 0.93$).

pornography also differs by gender,[178] so it is not surprising that we observed this pattern of gender difference here.

This study did not include extensive measures of study participants' individual differences. The basic demographic questions on political orientation and educational attainment did not significantly relate to perceptions of overall harmfulness in any condition.[179]

Participants were also asked to rate the extent to which they thought each kind of video would cause particular kinds of harm. Specifically, they were asked to rate whether the videos would interfere with the video subjects' prospects for employment, cause them emotional harm, hurt their reputation, or damage their election chances. On each of these questions, participants rating pornographic scenarios assigned higher scores (between 5 and 5.5 out of 6 for each question) than did those participants rating nonpornographic scenarios (between 4.7 and 5).[180] Based on their responses, participants expected labeling to help somewhat on employment and, nonsignificantly, on election chances, but labeling had no effect on emotional harm or reputation.[181] Further, female participants thought all deepfake scenarios were more likely to cause these negative effects than did male participants.[182]

---

[178] *See, e.g.*, Sarah Esther Lageson, Suzy McElrath & Krissinda Ellen Palmer, *Gendered Public Support for Criminalizing "Revenge Porn,"* 14 FEMINIST CRIMINOLOGY 560, 577 (2019) (reporting greater "support for criminalizing nonconsensual pornography among" those "respondents who identify as women").

[179] These results are available from the authors upon request.

[180] ANOVA tests were conducted looking at the factors pornographic versus attitudinal, labeled versus unlabeled, and male versus female. These are the effects for the main effect of pornographic versus attitudinal.

Employment: Attitudinal ($M = 4.72$, $SD = 1.43$), Pornographic ($M = 5.3$, $SD = 1.26$). $F(1, 1119) = 48.26$, $p < 0.001$ $\eta^2 = 0.04$.

Emotional harm: Attitudinal ($M = 4.89$, $SD = 1.35$), Pornographic ($M = 5.4$, $SD = 1.15$). $F(1, 1119) = 43.79$, $p < 0.001$ $\eta^2 = 0.04$.

Reputation: Attitudinal ($M = 4.93$, $SD = 1.35$), Pornographic ($M = 5.38$, $SD = 1.22$). $F(1, 1119) = 32.12$, $p < 0.001$ $\eta^2 = 0.03$.

Election chances: Attitudinal ($M = 4.92$, $SD = 1.34$), Pornographic ($M = 5.38$, $SD = 1.20$). $F(1, 1119) = 33.7$, $p < 0.001$ $\eta^2 = 0.03$.

[181] Employment: Unlabeled ($M = 5.13$, $SD = 1.33$), Labeled ($M = 4.91$, $SD = 1.42$). $F(1, 1119) = 8.05$, $p < 0.01$ $\eta^2 = 0.01$.

Election chances: Unlabeled ($M = 5.23$, $SD = 1.25$), Labeled ($M = 5.09$, $SD = 1.34$). $F(1, 1119) = 3.46+$ $\eta^2 = 0$.

[182] Employment: Male ($M = 4.84$, $SD = 1.48$), Female ($M = 5.18$, $SD = 1.26$). $F(1, 1119) = 14.94$, $p < 0.001$ $\eta^2 = 0.01$.

Emotional harm: Male ($M = 4.99$, $SD = 1.33$), Female ($M = 5.3$, $SD = 1.21$). $F(1, 1119) = 13.47$, $p < 0.001$ $\eta^2 = 0.01$.

Overall, then, participants felt that deepfake scenarios were quite harmful. This was especially true for pornographic scenarios and unlabeled attitudinal scenarios, but even labeled attitudinal scenarios were believed to cause harm (64 points out of 100) (see Figure 2). In terms of the kinds of harm that might result from these scenarios, people endorsed all of them to a high degree (approximately 5 out of 6 on all measures across all conditions). Deepfake views are also gendered, as women believe that pornographic deepfakes are more harmful than men do, though even men rate them as extremely harmful.

D. Follow-Up Study: Deepfakes and the Civil–Criminal Divide

Some states that have laws addressing nonconsensual pornography allow for both government-administered criminal punishment as well as private civil lawsuits.[183] One limitation of the primary study is that it focused on the criminal justice system. Participants who sought to punish deepfakes could only do so by suggesting a criminal sanction; there was no civil alternative. This design may have obscured a willingness among our participants to impose a less-than-criminal (or at least different-than-criminal) punishment.

Based on the results of the primary study, there is reason to think that participants would have been inclined to allow for both civil and criminal remedies in most cases. In general, criminal law is intended to punish morally blameworthy conduct, whereas the civil system is intended to compensate victims for wrongful injuries.[184] The questions in the first study, asking participants to rate the moral blameworthiness of the acts and their potential for causing harm, implicitly reflect these two related goals. Prior work has shown that people's preference for retributive punishment tracks the perceived wrongfulness of a transgression, whereas preference for compensatory damages is affected primarily by the amount of harm caused

Reputation: Male ($M = 5.00$, $SD = 1.36$), Female ($M = 5.31$, $SD = 1.22$). $F(1, 1119) = 13.74$, $p < 0.001$ $\eta^2 = 0.01$.

Election chances: Male ($M = 5.01$, $SD = 1.36$), Female ($M = 5.29$, $SD = 1.22$). $F(1, 1119) = 11.36$, $p < 0.001$ $\eta^2 = 0.01$.

[183] This may be in the form of two separate statutes or one statute. For example, Colorado has separate criminal and civil statutes. COLO. REV. STAT. ANN. §§ 13-21-1401–1409 (West 2019) (providing "Civil Remedies for Unauthorized Disclosure of Intimate Images"); *id.* §§ 18-7-107–108 (criminal statute). Vermont has a single statute that provides both criminal penalties and a civil cause of action. VT. STAT. ANN. tit. 13, § 2606 (West 2015).

[184] *See* OLIVER W. HOLMES, THE COMMON LAW 50–51 (Boston, Little, Brown & Co. 1881).

by the transgression.[185] Based on the blameworthiness and harm ratings from the first study, therefore, one would expect people to be seeking to both punish the video creator criminally as well as allow for civil compensatory recovery by the deepfake target.

Nevertheless, the first study does not provide firm evidence on whether people would have a strong preference between the civil and criminal systems. We therefore conducted a second study to specifically answer the question of whether people would prefer to deal with deepfake wrongs through the civil regime, the criminal regime, or both. This study employed only a subset of the scenarios employed in the first study, allowing us to ask this more complicated question without exhausting participant attention.

A sample of American adults was recruited in January 2021 by CloudResearch, another online survey firm with an established panel.[186] The demographics of the sample were set to match U.S. Census proportions on the dimensions of age and sex, but race, ethnicity, and educational attainment could freely vary.[187] This produced a sample that was somewhat more white, less Hispanic, and more educated than in the first study. The sample was, however, as politically neutral and gender- and age-balanced as the representative data collection in the primary study. Full demographics are reported in Appendix A. The final sample contained 395 individuals.[188] The changes in sample size and provider were aimed at reducing the cost of the survey.

The procedure for this study mirrored that of the first. After completing the demographic questions, participants were told that they would be asked to rate four scenarios. To test a range of different possibilities, we set up four scenarios: one pornographic (friend video), one attitudinal and defamatory (cocaine video), one attitudinal and non-defamatory (living-scientist video),

---

[185] John M. Darley, Lawrence M. Solan, Matthew B. Kugler & Joseph Sanders, *Doing Wrong Without Creating Harm*, 7 J. EMPIRICAL LEGAL STUD. 30, 41–43 (2010) (presenting an experimental study showing that more blameworthy states of mind produced higher punitive damages and proposed prison terms, whereas greater realized harm produced higher compensatory damages); Joseph Sanders, Matthew B. Kugler, Lawrence M. Solan & John M. Darley, *Must Torts Be Wrongs? An Empirical Perspective*, 49 WAKE FOREST L. REV. 1, 25–27 (2014) (presenting an empirical study showing that people were willing to assign compensatory, but generally not punitive, damages to innocent agents who caused harm).

[186] *See The Easiest Way to Find Participants for Academic Research*, CLOUDRESEARCH, https://www.cloudresearch.com/industries/students-universities/ [https://perma.cc/F8SZ-S95X].

[187] Recall that the only major demographic effect in the first study was on gender, which is still representative here.

[188] As in the first study, inattentive participants were screened from the final sample based on two criteria. First, participants who did not give the appropriate response to an attention check question—a question asking participants to give a particular response—or a CAPTCHA item were unable to complete the study. Second, participants were screened from the final sample if they finished the study in less than one-third of the time taken by the median participant.

and one defamatory and political (politician-terror-endorsement video).[189] Participants saw these four scenarios in a random order. As in the first study, participants were told that the protagonist, Will, had either labeled all his videos as fake or that he had done nothing to show the videos were not genuine. Following each scenario, the key new question asked:

> How, if at all, should it be possible to punish Will for making and distributing the video?
>
> (A) Will should not be punished.
>
> (B) [Deepfake subject] should be able to sue Will, have the video taken down, and get money in compensation for any harm they/she might suffer from the video.
>
> (C) It should be a crime for Will to do this, meaning that the government should be able to prosecute him. This might result in having the video taken down, a fine, and/or a prison sentence.
>
> (D) Both B and C (Will may be sued by [deepfake subject] and be criminally prosecuted).

Both the civil and criminal options here left open the possibility of a remedial injunction: removing the video. The main differences between the two are who is bringing the action (the state or the victim) and whether a prison sentence is possible. For simplicity, participants were not asked to give a magnitude judgment for either the criminal or civil punishment.

TABLE 7: PREFERENCE FOR CIVIL AND CRIMINAL REMEDIES FOR NONCONSENSUAL DEEPFAKES

	Pornographic, Friend		**Cocaine Use, Friend**		**Scientist, Living**		**Politician, Terror Endorsement**	
	Labeled	Not	Labeled	Not	Labeled	Not	Labeled	Not
No Punishment	3.6%	0.5%	3.1%	0.5%	12.2%	6.6%	7.1%	3.0%
Civil Punishment	17.3%	15.2%	26.5%	18.2%	33.2%	20.7%	18.9%	13.6%
Criminal Punishment	8.7%	10.1%	8.7%	10.1%	10.7%	13.1%	12.8%	10.6%
Both Civil and Criminal	70.4%	74.2%	61.7%	71.2%	43.9%	59.6%	61.2%	72.7%

Note. Values reflect the percentage of participants choosing each punishment option.

As can be seen in Table 7, participants generally wished to allow for both civil and criminal punishments. Providing participants with the option of a civil remedy had the effect of slightly lowering the percentage of

[189] The living-scientist scenario was modified slightly to say that the scientist was currently employed at a major university (rather than to have retired recently).

participants opting for criminalization and substantially lowered the percentage opting for no punishment as compared to the first study. In the unlabeled pornographic case, for instance, 92.7% of the respondents in the first study wished to criminalize the conduct, and 7.3% wished to assign no punishment. Here, 84.3% wished to criminalize (criminal punishment or both civil and criminal), and only 0.5% wished to assign no punishment, with the rest offering an exclusive civil remedy. There was a similar pattern for the labeled video of the scientist. In the first study, 69.6% of the sample wished to criminalize the conduct, and 30.4% wished to assign no punishment.[190] Here, 54.6% wished to criminalize (criminal punishment or both civil and criminal), and only 12.2% wished to assign no punishment, with the rest offering an exclusive civil remedy.

These results suggest that a small portion of those wishing to punish the creation and dissemination of deepfake videos would be satisfied with a civil rather than criminal remedy. Comparing the ratings here to those from the first study shows that the decline in desire to criminalize is, on average, 8.8 percentage points.[191] Conversely, the portion of the sample opting for no punishment also declines sharply, with only a single participant in the pornographic unlabeled condition opting to forgo any remedy.[192]

E. Follow-Up Study: Explicit Comparison to Traditional Nonconsensual Pornography

The prior two studies have shown substantial condemnation of pornographic deepfakes, whether labeled as fake or not, but they have not allowed an explicit comparison to traditional nonconsensual pornography where a picture or video showing someone's nude body is shared without their permission. Since so many states have laws prohibiting nonconsensual pornography, it would be helpful to know whether people view deepfake pornography as being on par with this already-regulated practice.

A short follow-up study was therefore conducted in July 2021. The sample for this study was also recruited by CloudResearch. The demographics of the sample were set to match U.S. Census proportions on the dimensions of age and sex, but race, ethnicity, and educational attainment

[190] Recall that this is the living-scientist variant, not the combination of the dead and living conditions (Schrodinger's Scientist) reported in *supra* Table 5.

[191] The Study 1 values are reported in *supra* Tables 2, 4, and 5, except for the living-scientist scenario (69.6% for labeled, 75.6% for unlabeled). Study 2 compared like scenario to like scenario, combining the criminal-punishment and both-civil-and-criminal-punishment options: 83.73 – 74.96 = 8.76, which rounds to 8.8.

[192] It is somewhat misleading to report the average for this decline (11.7 points), given the restricted range. Specifically, the average is greater than the small percentage of respondents opting against criminalization in the first study's pornographic condition.

could freely vary. Again, this produced a sample that was reasonably but not perfectly representative. Full demographics are reported in Appendix A. The final sample contained 417 individuals.[193]

The procedure for this study mirrored that of the first and second. After completing the demographic questions, participants were told that they would be asked to rate two scenarios. These were a modified version of the friend deepfake and a comparable traditional nonconsensual-pornography scenario. Participants saw these two scenarios in a random order. As in the second study, participants had the option of punishing the actor civilly or criminally if they so wished. They also rated the blameworthiness and harmfulness of the video.

The changes in the deepfake condition were relatively minor. The deepfake subject was described as a former romantic partner rather than as a friend, and the deepfake video was of the subject masturbating rather than having sexual intercourse.[194] The deepfake creator was said to have made and posted the video after the end of the romantic relationship. To maintain consistency with the other scenario, the video was not said to be labeled as fake. In the traditional nonconsensual-pornography condition, a woman, Mary, had sent her romantic partner, James, a video of herself masturbating. James was said to have requested this video and promised to keep it private. Again, the former partner posted the video online after the breakup. This condition was intended to fall within the scope of many nonconsensual-pornography laws by explicitly noting the expectation of confidentiality.[195] The text of both scenarios is included in Appendix C.

[193] As in the first study, inattentive participants were screened from the final sample based on two criteria. First, participants who did not give the appropriate response to an attention check question—a question asking participants to give a particular response—or a CAPTCHA item were unable to complete the study. Second, participants were screened from the final sample if they finished the study in less than one-third of the time taken by the median participant.

[194] The switch to masturbation was done to avoid any question of joint creation in the traditional nonconsensual-pornography case.

[195] See *supra* notes 108–113 and accompanying text for a discussion of state-by-state variations in nonconsensual-pornography provisions.

TABLE 8: PREFERENCE FOR CIVIL AND CRIMINAL REMEDIES

	Deepfake of Ex-Partner		**Traditional Nonconsensual Pornography of Ex-Partner**	
Blameworthy	5.51	(1.13)	5.35	(1.25)
Harmful	5.48	(1.14)	5.49	(1.06)
No Punishment	5.8%		7.2%	
Civil Punishment	20.6%		26.6%	
Criminal Punishment	12.9%		12.0%	
Both Civil and Criminal	60.7%		54.2%	

Note. For blameworthiness and harm: means (standard deviations in parentheses). On the punishment question, each row is reporting the proportion of the sample choosing that option.

As can be seen in Table 8, the deepfake and traditional nonconsensual-pornographic video were both viewed as highly morally blameworthy.[196] Posting the deepfake video was viewed as slightly more blameworthy, though, given the high scores, this difference may not be practically important.[197] There was no significant difference in the perceived harmfulness of each, though, again, both means are quite high.[198] In terms of desired remedy, the median participant would have allowed for both civil and criminal enforcement for each. Approximately equal proportions of participants wished to allow for civil and criminal remedies in each case. Slightly more participants wanted to allow for criminal sanctions in the deepfake case than in the traditional nonconsensual-pornography case, however.[199] Overall, there is somewhat less reliance on criminal remedies in this study than in the previous one. This may be due to using an ex-romantic partner as the deepfake subject rather than a friend or stranger.

Our participants, therefore, tended to view deepfake pornography as on par with traditional nonconsensual pornography. Compared to traditional nonconsensual pornography, creating and posting deepfake pornography

[196] The within-subjects ANOVA analyzing the harm and blameworthiness measures included order as a factor. There was a main effect of order on both measures. Blameworthiness: $F(1, 415) = 4.23$, $p = 0.04$ $\eta^2 = 0.01$. Harm: $F(1, 415) = 5.91$, $p = 0.015$ $\eta^2 = 0.014$. In each case, this was due to both scenarios being rated as worse when the traditional nonconsensual-pornography scenario came first. This is odd given that the traditional nonconsensual-pornography scenario was rated as less blameworthy in both orders; we might expect that when the worse-rated scenario is shown first, participants will be primed to rate the next scenario as more harmful and blameworthy, but the opposite occurred. There was no significant interaction between order and scenario condition (deepfake or not) on either measure. Blameworthiness: $F(1, 415) = 0.10$, $p = 0.753$ $\eta^2 = 0.000$. Harm: $F(1, 415) = 2.76$, $p = 0.097$ $\eta^2 = 0.007$.

[197] $F(1, 415) = 9.23$, $p = 0.003$ $\eta^2 = 0.022$.

[198] $F(1, 415) = 0.07$, $p = 0.785$ $\eta^2 = 0.000$.

[199] This difference is significant χ^2 $(1, N = 417) = 5.48$, $p = 0.019$.

was viewed as marginally more morally blameworthy, approximately as harmful, and slightly more likely to be deserving the attention of the criminal justice system. It is unclear why some participants appear to have viewed deepfakes as worse. This may be a result of victim-blaming in the traditional nonconsensual-pornography condition, but it could also be due to many other factors. For instance, greater effort is involved in fabricating a fake video rather than posting an already-available real one.

III. FITTING DEEPFAKE ATTITUDES INTO THE LAW

The consistent message of these surveys is that people overwhelmingly find pornographic and attitudinal deepfakes to be very harmful. Clearly labeling the deepfake as fake mitigated the harm for attitudinal deepfakes but not for pornographic ones. And respondents were nearly unanimous in wishing to allow for civil punishment, criminal punishment, or both of those making pornographic deepfakes. Our final study shows that people were inclined to treat pornographic deepfakes much like traditional nonconsensual pornography.

Thinking back to the relatively limited legal options for deepfake subjects discussed in Section I.C, there is a remarkable divergence between the moral expectations of our sample and the remedies available under privacy tort law. Our participants believe that pornographic deepfakes cause substantial injuries. These videos were believed to affect employment chances, emotional well-being, and general reputation.[200] Participants are almost definitionally correct in their belief that depiction in these deepfakes causes harm to a person's dignity: if people believe something is demeaning—"[c]ausing someone to lose their dignity and the respect of others"[201]—then it is. These findings would therefore substantially support the argument that being unwillingly featured in a pornographic deepfake is "highly offensive to a reasonable person." But even success on this argument would be of only limited help; the other elements of each of the key privacy torts of intrusion upon seclusion and public disclosure of private facts are not satisfied.[202]

Defamation and false light claims are also not helpful in supporting the moral intuitions of the sample. The survey respondents rated *labeled* deepfake videos—especially pornographic ones—as incredibly harmful. Yet

---

[200] *See supra* note 180 and accompanying text.

[201] *Demeaning*, OXFORD LEXICO DICTIONARY, https://www.lexico.com/en/definition/demeaning [https://perma.cc/Q5WT-LGG3]. The definition from Merriam-Webster is similar: "damaging or lowering the character, status, or reputation of someone or something." *Demeaning*, MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/demeaning [https://perma.cc/6EB4-FADF].

[202] *See supra* Section I.C.

both causes of action require a falsity,[203] and victims will not be able to pursue either claim when the video is obviously fake, such as when it is labeled as fake or uploaded to a website dedicated to deepfake videos. This returns us to our opening example of Kristen Bell. She explained that labeling a pornographic deepfake of her as fake did not cure her harm; the issue was that she had not consented.[204]

Statutory causes of action are similarly unhelpful in most states; deepfakes are beyond the reach of most current nonconsensual-pornography statutes.[205] But this is likely to be the subject of legislative consideration over the next several years. This Part, therefore, does two things. First, it attempts to understand the psychology behind some of the more puzzling findings from Part II. Second, it considers how the empirical results from Part II should inform our understanding of the First Amendment's limitations on deepfake regulation.

A. Contextualizing Deepfake Punitiveness

Across all scenarios, people were extremely willing to punish those who made and distributed deepfake videos. Somewhat surprisingly, many survey respondents viewed as blameworthy and harmful even deepfakes made with consent or deepfakes that did not create obvious harm, such as a deepfake depicting a scientist talking about their life's work or a deepfake depicting a politician singing a silly song. This Section considers how these puzzling results of the main study can be understood within two frameworks: moral psychology and feminist legal scholarship. The moral-psychology approach will explore how the condemnation of consensual deepfakes may be an explicable judgment error. The feminist-legal-scholarship approach will explore how condemnation of consensual deepfakes may be a sensible view given the bare-bones consent process described in our scenarios.

1. Moral Psychology: From Disgust to Harm

Though it is easy to justify the moral wrongfulness of the core deepfake cases, it is somewhat harder to explain how a consensual deepfake can be morally blameworthy. If the problem with a pornographic deepfake is that it

[203] This is slightly more complicated in the case of false light, where the accused message merely needs to convey a false impression. A woman was able to win a false light claim against a pornographic magazine when it published her (clothed) picture surrounded by lascivious images, because this arguably implied things about her character. Braun v. Flynt, 726 F.2d 245, 254 (5th Cir. 1984). Nonconsensual, labeled deepfake creations imply nothing in particular about the character of those depicted, however, so it would be harder for labeled deepfakes to serve as the basis for a false light claim.

[204] Abram, *supra* note 1.

[205] *See supra* notes 108–112 and accompanying text.

takes away the agency of the person depicted, then consent should remove that as a concern.

One possible explanation is that this is a kind of moral-judgment error. The person believes that deepfakes are bad, perhaps thinking of the nonconsensual pornographic deepfakes of celebrities. When confronted with a deepfake that is consensual and nonpornographic, the person may still have a negative feeling about the deepfake due to cognitive bleed over from the more common and more distasteful example. If this is occurring, it may be an example of what is called moral dumbfounding.[206] Moral dumbfounding can generally be defined as "the stubborn and puzzled maintenance of a judgment without supporting reasons."[207] The quintessential moral-dumbfounding study takes something that almost everyone believes is wrong (cannibalism, incest, or bestiality) and removes by fiat all of the factors that one would normally use to argue that the conduct is harmful.[208] For example, Professors Jonathan Haidt, Fredrik Björklund, and Scott Murphy asked survey participants to evaluate a scenario in which a medical research assistant eats a human cadaver that has been donated to a medical lab and will be incinerated the next day.[209] Moral dumbfounding occurs when people cannot articulate a reason for why cannibalism is wrong in that context but still maintain that it is morally objectionable.[210] Haidt and colleagues believe that this type of dumbfounding is common and that it shows that people often leap from intuitive feelings of disgust to judgments of moral wrongfulness without stopping to consider coherent philosophical theories of harm.[211] A moral-dumbfounding account of deepfake attitudes would suggest that people have an intuitive negative reaction to deepfakes generally, based on a number of factors, and that they fail to sufficiently correct their understandings when some of those factors are no longer present.

Perhaps contributing to this negative "gut reaction" against the idea of *any* deepfake videos is the novelty of the technology. Deepfake technology is relatively new, and the concept of inserting someone's face into a video to

---

[206] Cillian McHugh, Marek McGann, Eric R. Igou & Elaine L. Kinsella, *Searching for Moral Dumbfounding: Identifying Measurable Indicators of Moral Dumbfounding*, 3 COLLABRA: PSYCH. 1, 1–2 (2017) (noting that "[i]t is apparent from the literature that there is no single, agreed definition of moral dumbfounding" but that "an absence of reasons for, or an inability to justify or defend, a moral judgement, is consistently identified across definitions").

[207] Jonathan Haidt, Fredrik Björklund & Scott Murphy, Moral Dumbfounding: When Intuition Finds No Reason 1 (Aug. 10, 2000) (unpublished manuscript) (on file with journal).

[208] *See, e.g.*, *id.* at 5–6 (describing various moral-dumbfounding studies); McHugh et al., *supra* note 206, at 1.

[209] Haidt et al., *supra* note 207, at 18.

[210] *Id.* at 11–12; *see also* McHugh et al., *supra* note 206, at 5–6 (describing the Haidt et al. vignettes).

[211] *See* Haidt et al., *supra* note 207, at 11.

make them do or say something is strange and unusual. Research by Professors Kurt Gray and Jonathan Keeney has shown that people view morally questionable acts as more morally wrongful and as indicative of worse character if the person performing them engages in weird but morally irrelevant conduct (in this study, painting themselves red and putting on a hair cape).[212] Whether it is morally acceptable to make a deepfake pornographic video of a friend, or a deepfake biopic of a scientist, it is certainly uncommon. Put another way, "who does that?"

Both moral dumbfounding and this weirdness effect are part of a general literature in moral-psychology research that links moral judgment to perceptions of harm and feelings of disgust.[213] Within this literature, there are two general sorts of theories of how disgust, harm, and moral judgment are linked. Professors Jonathan Haidt and Matthew A. Hersh's direct disgust model, which grows out of work on moral dumbfounding, suggests that "[m]oral judgment (at least in the domain of sexual morality) is better predicted by affective reactions than by informational assumptions about harm."[214] These "affective reactions such as disgust and discomfort . . . are later cloaked by harm-based rationalizations."[215] Under this approach, anything that makes people uncomfortable will come to be viewed as wrong, and people will then generate theories of harm to justify their reactions post hoc. The theories of harm are, therefore, somewhat inconsequential; what actually matters is the initial gut reaction.

A competing theory—the theory of dyadic morality—takes the theories of harm far more seriously. Psychologists Chelsea Schein and Kurt Gray suggest two principles that explain moral judgment: "what seems harmful seems wrong" and "what seems wrong seems harmful."[216] Schein and Gray suggest that these two principles interact to create a dyadic feedback loop, amplifying the perceived harmfulness and wrongfulness of certain issues.[217] Rather than theories of harm being irrelevant justifications for visceral reactions, under this approach, they play a substantial independent role. That

---

[212] Kurt Gray & Jonathan E. Keeney, *Impure or Just Weird? Scenario Sampling Bias Raises Questions About the Foundation of Morality*, 6 SOC. PSYCH. & PERSONALITY SCI. 859, 864–65 (2015).

[213] For a discussion on the background of moral psychology research, see Chelsea Schein, Ryan S. Ritter & Kurt Gray, *Harm Mediates the Disgust-Immorality Link*, 16 EMOTION 862, 862–63 (2016).

[214] Jonathan Haidt & Matthew A. Hersh, *Sexual Morality: The Cultures and Emotions of Conservatives and Liberals*, 31 J. APPLIED SOC. PSYCH. 191, 213 (2001).

[215] *Id.* at 212 (citation omitted).

[216] Chelsea Schein & Kurt Gray, *Moralization and Harmification: The Dyadic Loop Explains How the Innocuous Becomes Harmful and Wrong*, 27 PSYCH. INQUIRY 62, 62 (2016).

[217] *Id.* ("This feedback loop has the power to amplify the perceived levels of both harm and immorality: what seems harmful seems wrong, and what seems wrong seems *more* harmful, and what seems more harmful becomes *more* wrong, and so on.").

which feels disgusting will initially be viewed as wrongful, but this feeling may either deepen or depart depending on whether the person can construct a theory of harm to justify their initial reaction. Similarly, that which appears initially harmful may come to be seen as disgusting.

This feedback cycle may further help explain our survey results. Survey respondents clearly viewed deepfake videos as harmful, which may have led them to view the behavior as blameworthy. The dyadic framework suggests that if individuals have an "inkling of an intuition of harm" in a given context, they will view it as "somewhat immoral," which will then cause them to perceive more harm,[218] which might culminate in "deepening moral judgments."[219] The harm perceived in the more blatantly harmful deepfake videos may therefore have "deepen[ed] and expand[ed] to related concepts,"[220] such as the less blatantly harmful deepfake videos. In short, participants may have been so persuaded by the generally problematic nature of deepfakes that they neglected to fully discount their feelings of disgust in the presence of consent.

2. Scope of Consent and Feminist Legal Scholarship

There are also philosophical arguments that support viewing even consensual deepfakes as harmful. Here it is helpful to consider the perspective of antipornography feminism. Traditionally, antipornography feminists have condemned pornography based on its perceived harmful impact on women. Professor A. W. Eaton describes this "harm hypothesis" of antipornography feminist theory, noting that traditional antipornography feminism connects pornography to harm through both the production and the postproduction of pornography.[221] Essentially, this "harm hypothesis" concludes that "by harnessing representations of women's subordination to a ubiquitous and weighty pleasure, pornography is especially effective at getting its audience to internalize its inegalitarian views."[222]

Deepfakes often depict pornography, and although the product does not subject the depicted woman to physical exploitation in the same way that making live pornography might, the final product still depicts a woman's likeness. Recall that the scenarios in the study were intentionally written to reflect the current trends in pornographic deepfakes: men created the videos, and in the pornographic-deepfake context, all of the videos created were of women. Even when the woman has consented, the survey respondents might

218 *Id.*

219 *Id.*

220 *Id.* at 63.

221 A. W. Eaton, *A Sensible Antiporn Feminism*, 117 ETHICS 674, 677 (2007).

222 *Id.* at 680.

be uncomfortable with another having control over a woman's likeness to create sexualized depictions. This would be consistent with prior scholarship that critiques the genuineness of consent in a patriarchal society.[223] It also reflects a potential view that the protagonist should not have even wanted to produce the video.

One need not accept this particular brand of feminist critique to have concerns about the consent depicted in these deepfake scenarios. As we mentioned in Part II, it might not have been clear to the survey respondents that the people consenting to deepfake creation were making a free and informed choice. The scenarios are silent on whether the participant consented to the specific contents of the videos or even knew how deepfakes worked. One could easily imagine a participant having genuine concerns that the allegedly consenting party did not know to what they were agreeing. Also, given the high harmfulness scores for pornographic deepfakes, survey respondents might be concerned with the postproduction consequences of the deepfakes. Neither the deepfake subject nor the deepfake creator has full control over the distribution of a video once it has been publicly posted.

Indeed, scholars have raised a similar concern about the genuineness of consent in the privacy context more generally. Professor Daniel Solove, for example, notes that although consent is at the core of privacy self-management, individuals often do not meaningfully consent to the collection, use, and disclosure of their data due to flawed decision-making and structural problems, such as the vast number of entities collecting data and the unanticipated impacts of aggregated data.[224] Survey respondents may hold similar concerns about deepfakes. The consent-skeptical responses of survey respondents are therefore not entirely unreasonable, even if we would be slow to endorse them as a policy matter.

Notably, one existing deepfake statute already contains provisions responsive to a consent-skeptical view. The recently passed New York deepfake statute says that a person may only consent to the creation or dissemination of pornographic deepfake "by knowingly and voluntarily signing an agreement written in plain language that includes a general description of the sexually explicit material and the audiovisual work in which it will be incorporated."[225] This consent process is more detailed than

---

[223] *See, e.g.*, Morrison Torrey, *Feminist Legal Scholarship on Rape: A Maturing Look at One Form of Violence Against Women*, 2 WM. & MARY J. WOMEN & L. 35, 41 (1995) ("In general, feminist critiques of the legal definition of consent to sexual activity fall into three categories: (1) true consent is not possible until women are no longer subordinated by men; (2) consent is often presumed or implied in non-stranger rape; and (3) prevalent sexual mythology encourages men to disbelieve women when they say 'no.'").

[224] Daniel J. Solove, *Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880, 1880–82 (2013).

[225] N.Y. CIV. RIGHTS LAW § 52-c(3)(a)–(b) (McKinney 2021).

that in our scenarios and would result in more thorough notice to the deepfake subject.

B. Deepfakes and the First Amendment

Because current law often does not vindicate the privacy interests identified by our subjects—except to a degree in states like California and New York—it is important to consider whether an expansion of current law could do so. The most substantial area where our subjects would wish to grant new protection is in the context of labeled pornographic deepfakes. We analyze existing First Amendment doctrine in the context of falsity, nonconsensual pornography, and morphed child pornography to understand how courts might approach expanded deepfake laws that seek to give protection in this area.

1. The Current First Amendment Framework

The Supreme Court has defined categories of speech that fall outside First Amendment protection—speech "likely[] to incite imminent lawless action," obscenity, defamation, "speech integral to criminal conduct," fighting words, child pornography, fraud, threats, and "speech presenting some grave and imminent threat the government has the power to prevent."[226] Deepfake videos as a whole do not fall within these categories, although specific deepfake videos can depict content that does. So, the fact that a video is a deepfake does not make it obscene, but a deepfake might depict obscenity. Because of this, any statute that bans deepfake videos outside these categories will likely have to be narrowly tailored to serve a compelling state interest to withstand strict scrutiny.[227]

Against this backdrop, banning deepfake videos will not be without challenges. Deepfake videos cannot be banned merely because they are false in nature. In *United States v. Alvarez*, the Supreme Court struck down the Stolen Valor Act, which made it a crime to make false statements about receiving military decorations or medals.[228] The Court reasoned that it had

---

[226] United States v. Alvarez, 567 U.S. 709, 717 (2012).

[227] *See, e.g.*, Reed v. Town of Gilbert, 576 U.S. 155, 163–64 (2015) (noting that laws which "cannot be 'justified without reference to the content of the regulated speech'" must face strict scrutiny (quoting Ward v. Rock Against Racism, 491 U.S. 781, 791 (1989))).

[228] *Alvarez*, 567 U.S. at 715. The relevant part of the Act read:

> "Whoever falsely represents himself or herself, verbally or in writing, to have been awarded any decoration or medal authorized by Congress for the Armed Forces of the United States . . . shall be fined under this title, imprisoned not more than six months, or both. . . . If a decoration or medal involved in an offense under subsection (a) or (b) is a Congressional Medal of Honor . . . the offender shall be fined under this title, imprisoned not more than 1 year, or both."

Id. at 715–16 (quoting 18 U.S.C. § 704(b)–(c)).

never held that falsity alone was outside First Amendment protection.[229] Rather, false statements fall outside First Amendment protection when there are additional considerations, such as "some other legally cognizable harm associated with [the] false statement"[230] or "[w]here false claims are made to effect a fraud or secure moneys or other valuable considerations, say, offers of employment."[231]

In the deepfake context, *Alvarez* would prohibit an outright ban on all deepfake videos and also a ban on deepfake videos that have no cognizable harms associated with them. Notably, the participants in the study wanted to criminalize unlabeled attitudinal deepfakes, but under *Alvarez*, unlabeled attitudinal deepfakes cannot be prohibited for merely promoting falsehoods.[232] For example, a deepfake of a politician singing a silly song could not be prohibited unless there was some problem with it beyond mere falsity.[233] Survey respondents seemed to associate all deepfake videos with harm, rating both labeled and unlabeled deepfakes as incredibly harmful and indicating a belief that both labeled and unlabeled deepfakes could interfere with the subject's employment prospects, cause emotional and reputational harm, and, where applicable, interfere with the subject's election chances. However, regulations on deepfake videos can likely not be this expansive.[234] The kind of election-proximity protection offered to candidates in California and Texas may be constitutional based on prior case law that limits

[229] *Id.* at 719 ("The Court has never endorsed the categorical rule the Government advances: that false statements receive no First Amendment protection. . . . Even when considering some instances of defamation and fraud, moreover, the Court has been careful to instruct that falsity alone may not suffice to bring the speech outside the First Amendment. The statement must be a knowing or reckless falsehood.").

[230] *Id.*

[231] *Id.* at 723.

[232] *See id.* ("Were the Court to hold that the interest in truthful discourse alone is sufficient to sustain a ban on speech, absent any evidence that the speech was used to gain a material advantage, it would give government a broad censorial power unprecedented in this Court's cases or in our constitutional tradition.").

[233] *See id.* at 721 (noting that "[s]tatutes that prohibit falsely representing that one is speaking on behalf of the Government, or that prohibit impersonating a Government officer, also protect the integrity of Government processes, quite apart from merely restricting false speech").

[234] Deepfakes cannot be banned merely because they depict upsetting content. *See, e.g.*, Snyder v. Phelps, 562 U.S. 443, 458 (2011) (holding speech on a matter of public concern "cannot be restricted simply because it is upsetting or arouses contempt"); Texas v. Johnson, 491 U.S. 397, 414 (1989) ("If there is a bedrock principle underlying the First Amendment, it is that the government may not prohibit the expression of an idea simply because society finds the idea itself offensive or disagreeable."); Hustler Magazine, Inc. v. Falwell, 485 U.S. 46, 50 (1988) (declining to hold that "a State's interest in protecting public figures from emotional distress is sufficient to deny First Amendment protection to speech that is patently offensive and is intended to inflict emotional injury, even when that speech could not reasonably have been interpreted as stating actual facts about the public figure involved").

electioneering near polling places,[235] but that would provide far more narrowly tailored protection than most participants are seeking.

One type of falsity-related deepfake regulation that is on firmer constitutional ground is a labeling requirement for any deepfake video that is defamatory in nature. Since participants were somewhat less concerned about labeled deepfakes in the nonpornographic context, such a policy would be consistent with public views. Given that defamation is one of the categories excluded from First Amendment protection, this would likely survive scrutiny. Though such videos would violate existing defamation law—arguably making such a provision superfluous—the added emotional impact of a defamatory deepfake video may be reason to grant enhanced protection against deepfake defamation.

2. Nonconsensual Pornography

Though a labeling requirement might deal with some of the harms from attitudinal deepfakes, our study shows that the harm of pornographic deepfakes is unmitigated by such an intervention. Further, participants in our final study treated deepfake pornography as on par with traditional nonconsensual pornography, which is already widely prohibited. These findings raise the question of whether it is possible to ban even labeled nonconsensual pornographic deepfakes. No court has directly addressed this issue, but there is parallel case law on nonconsensual pornography and doctored videos that depict child pornography. Based on this case law and the survey responses, we believe there is a strong case for viewing the regulation of deepfake pornography as a compelling state interest.

Nonconsensual pornography, sometimes called revenge pornography, refers to sexually graphic images and videos that are generally made with consent by the depicted subjects and then nonconsensually made public.[236] Unlike deepfake pornography, nonconsensual pornography is not altered and depicts no falsity. As of November 2021, forty-eight jurisdictions have criminalized nonconsensual pornography,[237] and those statutes have been challenged in state courts on First Amendment grounds in seven states.[238] The highest courts of only four states, those in Vermont (*State v. VanBuren*),

---

[235] Burson v. Freeman, 504 U.S. 191, 207–08 (1992) (upholding a law creating a campaign-free zone within 100 feet of the entrance to a polling place).

[236] Citron, *supra* note 6, at 1917–18.

[237] Sales & Magaldi, *supra* note 113, at 1500; *48 States + DC + One Territory Now Have Revenge Porn Laws*, CYBER C.R. INITIATIVE, https://www.cybercivilrights.org/revenge-porn-laws/ [https://perma.cc/C5EH-GK5W].

[238] Nonconsensual-pornography statutes have been challenged in Arizona, Texas, Wisconsin, Vermont, Illinois, Indiana, and Minnesota. *See* Sales & Magaldi, *supra* note 113, at 1533–34; State v. Casillas, 952 N.W.2d 629, 634 (Minn. 2020); Order Dismissing Charging Information, ¶¶ 12, 28, Indiana v. Katz, No. 76C01-2005-CM-000421 (Ind. Cir. Ct. Oct. 2, 2020).

Illinois (*People v. Austen*), Minnesota (*State v. Casillas*), and Texas (*Ex parte Jones*), have reviewed the constitutionality of their respective state's nonconsensual pornography statutes.[239]

Although much of the First Amendment analysis in these cases focuses on the language of the statutes, all of the state supreme courts specifically note the harm associated with nonconsensual pornography and find the state's interest in protecting victims of nonconsensual pornography compelling, substantial, or important.[240] The opinions variously cited privacy, reputational, and psychological harms; the perpetration of domestic violence; and the subsequent harassment and threats victims experience after the dissemination of the images or videos.[241] For example, the Vermont court wrote that prior U.S. Supreme Court statements suggest that "the government may regulate speech about purely private matters that implicates privacy and reputational interests."[242] The courts further acknowledged that victims have been fired and have difficulty finding employment.[243] The Vermont Supreme Court specifically underscored the emotional and reputational harms of nonconsensual pornography, stating, "The personal consequences of such profound personal violation and humiliation generally include, at a minimum, extreme emotional distress."[244] And the Texas court also recognized that "[v]ictims of revenge porn cannot counterspeak their way out of a violation of their most private affairs and bodily autonomy nor the serious harms that may accompany that violation."[245] It noted that this lack

---

[239] *See* State v. VanBuren, 214 A.3d 791, 794 (Vt. 2019); People v. Austin, 155 N.E.3d 439, 448 (Ill. 2019); *Casillas*, 952 N.W.2d at 629; *Ex parte* Jones, No. PD-0552-18, 2021 WL 2126172, at *1 (Tex. Crim. App. May 26, 2021), *reh'g denied*, (July 28, 2021). The Texas Court of Criminal Appeals is Texas's highest court. It did not publish its decision in this case, possibly because the statute had since been materially amended.

[240] *See VanBuren*, 214 A.3d at 810–11; *Austin*, 155 N.E.3d at 461–62; *Casillas*, 952 N.W.2d at 641–42; *Jones*, 2021 WL 2126172, at *7 ("We agree with the State that the privacy interest in the statute is a compelling government interest . . . [and] particularly, the interest in sexual privacy is substantial."). A lower court in Wisconsin also used similar language. State v. Culver, 918 N.W.2d 103, 110 (Wis. Ct. App. 2018) ("In prohibiting the knowing publication of intentionally private depictions of another person who is either nude, partially nude, or engaged in sexually explicit conduct, the statute serves to protect an important state interest—individual privacy. No one can challenge a state's interest in protecting the privacy of personal images of one's body that are intended to be private—and specifically, protecting individuals from the nonconsensual publication on websites accessible by the public.").

[241] *See VanBuren*, 214 A.3d at 810–11 (privacy, reputational, and psychological harm; harassment; threats of violence); *Austin*, 155 N.E.3d at 461–62 (psychological harm; threats of violence; harassment; facilitation of domestic violence, human trafficking, and sexual assault); *Casillas*, 952 N.W.2d at 641–42 (privacy, psychological, and reputational harm); *Jones*, 2021 WL 2126172, at *7 (privacy, reputational, and psychological harm; harassment).

[242] *VanBuren*, 214 A.3d at 802.

[243] *See id.* at 810–11; *Austin*, 155 N.E.3d at 461.

[244] *VanBuren*, 214 A.3d at 810.

[245] *Jones*, 2021 WL 2126172, at *7.

of a counterspeech remedy makes nonconsensual pornography different than other categories of harmful expression.[246]

There are substantial similarities between the privacy-related harms contained within deepfake pornography and nonconsensual pornography. As with nonconsensual pornography, victims of deepfake pornography report various harms, including harassment and threats.[247] The survey responses are also consistent with the notion that deepfake pornography, both labeled and unlabeled, is extremely harmful and an affront to the dignity of the person depicted. In *VanBuren*, the court relied heavily on prior case law that determined the state has a compelling interest in the regulation of purely private matters such as intimate images of a person.[248] The court in *Austin* relied on a similar privacy rationale, at times borrowing from *VanBuren*.[249] Deepfake pornography, like nonconsensual pornography generally, concerns the dignitary privacy one has over her likeness. Nonconsensual pornography and deepfake pornography both involve a type of dignitary harm that stems from one's ability to control information about oneself.[250] Nonconsensual pornography involves disclosure of personal information, which "can severely inhibit a person's autonomy and self-development."[251] Deepfake pornography creates similar harm as a "distortion" that manipulates "the way a person is perceived and judged by others, and involves the victim being inaccurately exposed to the public."[252] Much like the painful accuracy of nonconsensually disclosed pornography, the misrepresentation of deepfake pornography impacts one's ability to control their sexual identity.[253] As noted by the court in *VanBuren*, "In the constellation of privacy interests, it is difficult to imagine something more private than images depicting an individual engaging in sexual conduct . . . ."[254]

---

[246] *Id.* at *7 n.79 (suggesting that counterspeech may serve "as a remedy for lies and 'speech we do not like'" (quoting United States v. Alvarez, 567 U.S. 709, 726–28 (2012))).

[247] *See, e.g.*, Citron, *supra* note 6, at 1921–23 (describing a female journalist targeted on social media with sexual violence accompanied with attitudinal and pornographic deepfake videos); Harwell, *supra* note 74 (describing pornographic deepfake videos as being "weaponized disproportionately against women, representing a new and degrading means of humiliation, harassment, and abuse").

[248] *See VanBuren*, 214 A.3d at 808 ("Time and again, the Supreme Court has recognized that speech concerning purely private matters does not carry as much weight in the strict-scrutiny analysis as speech concerning matters of public concern, and may accordingly be subject to more expansive regulation.").

[249] *See Austin*, 155 N.E.3d at 460–62.

[250] *See* ALAN F. WESTIN, PRIVACY AND FREEDOM 7 (1967) ("Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others.").

[251] Daniel J. Solove, *The Virtues of Knowing Less: Justifying Privacy Protections Against Disclosure*, 53 DUKE L.J. 967, 991 (2003).

[252] Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477, 547 (2006).

[253] *See* Citron, *supra* note 6, at 1921.

[254] State v. VanBuren, 214 A.3d 791, 810 (Vt. 2019).

Though each of the four states to rule on these statutes has upheld them, the constitutionality of nonconsensual-pornography laws is disputed.[255] To the extent nonconsensual pornography can be criminalized, however, it follows that deepfake pornography can also be criminalized. Our participants appear to have viewed pornographic deepfakes as a dignitary violation rather than as a defamatory message because they were not substantially reassured by the prospect that the videos could be labeled as fake. This finding makes us comfortable categorizing pornographic deepfakes as speech that implicates sexual privacy, the protection of which has consistently been considered a substantial or compelling government interest.[256]

3. Morphed Pornography

The question of whether deepfake pornographic videos are effectively the same as real pornographic videos has arisen before in the context of child pornography. Child pornography law differentiates between virtual child pornography, which does not depict actual children, and morphed child pornography, which inserts the face of a real child onto the body of an adult in a pornographic picture or video. These are, effectively, deepfakes before deepfakes. Fully virtual child pornography cannot be criminalized under the Supreme Court's decision *Ashcroft v. Free Speech Coalition*,[257] but that case specifically left open the question of morphed child pornography.[258]

All circuits addressing the question of morphed child pornography have held that it is permissible to criminalize morphed pornography that uses the face of a real child.[259] The Fifth Circuit case was the most recent. In agreeing with the Second and Sixth Circuits that morphed child pornography is not

---

[255] *See, e.g.*, Andrew Koppelman, *Revenge Pornography and First Amendment Exceptions*, 65 EMORY L.J. 661, 662 (2016) ("The constitutionality of [revenge-porn] laws is uncertain . . . ."); John A. Humbach, *The Constitution and Revenge Porn*, 35 PACE L. REV. 215, 260 (2014) ("It appears that most of the revenge-porn laws recently proposed and enacted, which simply punish sexually-themed images disseminated without consent of persons depicted, are unconstitutional . . . .").

[256] *See VanBuren*, 214 A.3d at 811; People v. Austin, 155 N.E.3d 439, 462 (Ill. 2019); People v. Iniguez, 202 Cal. Rptr. 3d 237, 243 (2016).

[257] 535 U.S. 234, 256 (2002). More specifically, it cannot be criminalized under the child pornography exception to the First Amendment. It may be possible to criminalize it as obscenity.

[258] *See id.* at 242.

[259] *See* Doe v. Boland, 698 F.3d 877, 880 (6th Cir. 2012) ("Morphed images are of a piece [with traditional pornography], offering a difference in degree of injury but not in kind."); United States v. Mecham, 950 F.3d 257, 260 (5th Cir. 2020), *cert. denied*, 141 S. Ct. 139; United States v. Hotaling, 634 F.3d 725, 730 (2d Cir. 2011) ("[H]ere we have six identifiable minor females who were at risk of reputational harm and suffered the psychological harm of knowing that their images were exploited and prepared for distribution by a trusted adult."); United States v. Anderson, 759 F.3d 891, 895–96 (8th Cir. 2014) ("Although subjects of morphed images . . . do not suffer the direct physical and psychological effects of sexual abuse that accompany the production of traditional child pornography, the morphed images' 'continued existence causes the child victims continuing harm by haunting the children in years to come.'" (quoting Osborne v. Ohio, 495 U.S. 103, 111 (1990))).

protected speech, the court noted, "By using identifiable images of real children, these courts conclude, morphed child pornography implicates the reputational and emotional harm to children that has long been a justification for excluding real child pornography from the First Amendment."[260] In effect, fake child pornography that appears to feature a real child can be criminalized for a subset of the same reasons that real child pornography featuring that child can be criminalized.

It is tempting to directly apply the same rationale to nonconsensual adult pornography and nonconsensual adult deepfake pornography. In each case, the fact that the video is morphed rather than genuine fails to prevent the harm to dignity and the risk of concrete consequences to employment. Historically, however, child pornography has been treated differently than adult pornography. In *New York v. Ferber*, the Supreme Court upheld a ban on child pornography, holding that the state has a compelling interest in the well-being of minors and that child pornography relates to the sexual abuse of children in two ways.[261] "First, the materials produced are a permanent record of the children's participation and the harm to the child is exacerbated by their circulation."[262] Second, to combat the sexual exploitation necessarily involved in the production of child pornography, the distribution networks must be closed.[263] Almost a decade later, the Court upheld an Ohio statute banning the possession and viewing of child pornography.[264] There, the Court reasoned that the statute encouraged the destruction of child pornography, which otherwise creates a permanent recording of child victims and their abuse and is used to coerce children into engaging in sexual conduct.[265]

The protection of children, therefore, is an especially compelling state interest. Courts may be less willing to grant expansive protection against abuses perpetrated on adults with morphed images and videos than they are in the case of children because, historically, courts have "sustained legislation aimed at protecting the physical and emotional well-being of youth even when the laws have operated in the sensitive area of constitutionally protected rights."[266] This means that courts could justifiably distinguish here between the importance of morphing in the child and adult contexts. Recall that the *Ashcroft* Court extended protection to fully virtual

---

[260] *Mecham*, 950 F.3d at 265.

[261] 458 U.S. 747, 759 (1982).

[262] *Id.* at 759.

[263] *Id.*

[264] *Osborne*, 495 U.S. at 111.

[265] *Id.*

[266] *Ferber*, 458 U.S. at 757. It is a little unclear how this interest in protecting children works in the case of morphed images. If the picture was taken at age ten and the subject is now thirty, should they still get the enhanced protection due children?

child pornography in part because it did not require harming real children to make it.[267] One could imagine a court using similar language regarding deepfake pornography of adults.

Nevertheless, the reputational and emotional harms credited by courts in the context of morphed child pornography are similar to those reported by adults depicted in nonconsensual deepfake pornography. Indeed, our survey respondents acknowledged that those depicted in pornographic deepfakes would experience such harm. The results of our studies, therefore, reinforce the logic of the morphed child pornography cases and their application to deepfake adult pornography.

CONCLUSION

If a person has a supply of good pictures of a target, they can make a video of that target saying or doing almost anything. This revolution in video-morphing technology has caused deepfake videos to explode in prevalence over the last several years. Our studies show that there is a strong moral consensus that the creation of nonconsensual deepfakes is wrongful and causes extensive harm. Further, the studies show that pornographic deepfake videos—which are the majority of deepfake videos on the internet—are considered especially harmful. Though the public has divided views about some attitudinal deepfakes, even sexualized videos lacking nudity were almost universally condemned.

Labeling a deepfake as fake mitigates the harm for attitudinal deepfakes but not for pornographic deepfakes. Though there are sharp constitutional limits on whether it is possible to prohibit the creation of labeled attitudinal deepfakes, it is likely possible to prohibit the creation of pornographic deepfakes given the existing First Amendment case law on nonconsensual pornography. The public attitudes captured here provide strong support for doing so and should be taken seriously by courts and policymakers grappling with this new technology.

The case of deepfake technology further points to an emerging problem in the privacy landscape. Privacy in this context is about dignity, autonomy, and identity expression—about people losing control of their public identities. To appropriately understand the dangers associated with deepfakes and the unauthorized use of one's likeness, courts and policymakers must take seriously the kinds of dignitary harms associated with these new kinds of privacy invasions.

---

[267] *See* Ashcroft v. Free Speech Coalition, 535 U.S. 234, 236 (2002) ("*Ferber's* judgment about child pornography was based upon how it was made, not on what it communicated. The case reaffirmed that where the speech is neither obscene nor the product of sexual abuse, it does not fall outside the First Amendment's protection.").

APPENDIX A: DEMOGRAPHICS OF THE SAMPLES

The sample for Study 1 was recruited by Dynata. The samples for Studies 2 and 3, reported in Sections II.D and II.E, respectively, were recruited by CloudResearch.

TABLE A1: DEMOGRAPHIC DATA FOR EACH SURVEY

	Study 1	**Study 2**	**Study 3**	**Census**[268]
Gender				
Female	52.1%	50.9%	55.2%	50.8%
Male	47.9%	49.1%	44.4%	49.2%
Other	0.0%	0.0%	0.5%	
Age (Years)				
Median	48	47	45[269]	
Mean	47.81 (17.50)	49.18 (15.55)	44.81 (15.80)	
Political Orientation (1–7)[270]	4.12 (1.80)	4.10 (1.79)	3.97 (1.78)	
Race and Ethnicity				
White	79.1%	84.1%	76.5%	76.3%
Black or African American	11.1%	9.6%	16.1%	13.4%
American Indian or Native American	0.7%	1.3%	1.4%	1.3%
Asian American	5.6%	4.1%	1.2%	5.9%
Hawaiian or Pacific Islander	0.3%	0.0%	1.9%	0.2%
Multiracial or Other	3.2%	1.1%	2.9%	2.8%
Hispanic (of Any Race)	16.0%	8.1%	10.6%	18.5%
Educational Attainment				
Less Than High School Diploma	7.8%	1.8%	4.3%	10.9%
High School Diploma or GED	31.2%	18.7%	30.9%	28.6%
Two-Year or Some College	28.8%	38.2%	38.1%	28.2%
Four-Year College	20.8%	25.1%	18.9%	20.6%
Graduate Degree	11.4%	16.2%	7.7%	11.6%

Note. For age and political orientation: means (standard deviations in parentheses). Hispanic identity was assessed in a separate question than racial identity.

[268] Ethnicity and gender statistics are from the U.S. Census website. *See QuickFacts*, U.S. CENSUS BUREAU, https://www.census.gov/quickfacts//fact//table//US//PST045217 [https://perma.cc/S5BR-9P3J]. Educational attainment was calculated from data in table 1 in *Educational Attainment in the United States: 2018*, U.S. CENSUS BUREAU (Apr. 17, 2020), https://www.census.gov/data/tables/2018/demo/education-attainment/cps-detailed-tables.html [https://perma.cc/Q458-PS5U].

[269] Two participants in Study 3 entered what appears to have been their birth year. Their ages were estimated based off that information. One participant entered an out-of-range number, so their response to the age question was disregarded.

[270] Political orientation was assessed on a scale ranging from 1, very liberal, to 7, very conservative.

APPENDIX B: UNLABELED VARIANTS OF ALL SCENARIOS FROM PRIMARY STUDY

These are the unlabeled scenario variants used in the studies. The labeled variants were adapted from these by replacing the final sentences as described on page 637.

A. Pornographic Scenarios

Written Pornographic Story, Friend

Imagine Jane is a friend of Will. Will has written a story about Jane. In Will's story, he describes what Jane really looks like and depicts her having graphic sex with a man. The story is very detailed. Will posts his story online publicly, and he includes Jane's first and last name. Though this story is made up, a reader cannot easily tell. Will does not indicate that it is fake when he posts it.

Deepfake Pornographic Video, Friend

Imagine Jane is a friend of Will. Will finds a series of photos of Jane online. Will takes the photos and uses an app to merge her face onto a pornographic video. The final video shows Jane's face on the body of a naked woman having sex with a man. The video shows the entirety of the naked woman's body. Jane's face is clearly identifiable in the video. Will posts the video online publicly, and he includes Jane's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Pornographic Video, Celebrity

Imagine Will finds a series of photos of a famous female celebrity online. Will finds a series of photos of the celebrity online. Will takes the photos and uses an app to merge her face onto a pornographic video. The final video shows the celebrity's face on the body of a naked woman having sex with a man. The video shows the entirety of the naked woman's body. The celebrity's face is clearly identifiable in the video. Will posts the video online publicly, and he includes the celebrity's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Pornographic Video, Sexualized Voice

Imagine Jane is a friend of Will. Will finds a series of photos of Jane online. Will takes the photos and uses an app to merge her face onto a video. The final video shows Jane's face on the body of a woman who is wearing revealing clothing. The woman is not nude. The video depicts Jane speaking

seductively about having sex. Jane's face is clearly identifiable in the video. Will has also used software to simulate Jane's voice, so the voice in the video sounds exactly like Jane's real voice. Will posts the video online publicly, and he includes Jane's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Pornographic Video, No Nudity, BDSM

Imagine Jane is a friend of Will. Will finds a series of photos of Jane online. Will takes the photos and uses an app to merge her face onto a video. The final video shows Jane's face on the body of a woman who is spanking a man. The woman is dressed in a revealing leather outfit. Jane's face is clearly identifiable in the video. Will posts the video online publicly, and he includes Jane's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Pornographic Video, Personal Use, No Consent

Imagine Jenny is a friend of Will. Will has created a video of Jenny. Will finds a series of photos of Jenny online. Will takes the photos and uses an app to merge her face onto a pornographic video. The final video shows Jenny's face on the body of a naked woman having sex with a man. The video shows the entirety of the nude woman's body. Jenny's face is clearly identifiable in the video. Though this video is made up, a viewer cannot easily tell that it has been altered. Will keeps the video for himself and never shares it with anyone.

Deepfake Pornographic Video, Personal Use, Consent

Imagine Jenny is a friend of Will. Will asks Jenny if he can edit her face into a pornographic video that he will not show to anyone else. Jenny says yes. Will finds a series of photos of Jenny online. Will takes the photos and uses an app to merge her face onto a pornographic video. The final video shows Jenny's face on the body of a naked woman having sex with a man. The video shows the entirety of the nude woman's body. Jenny's face is clearly identifiable in the video. Though this video is made up, a viewer cannot easily tell that it has been altered. Will keeps the video for himself and never shares it with anyone.

B. Private Attitudinal Scenarios

Written Cocaine-Use Story

Imagine Jane is a friend of Will. Will has written a story about Jane. In Will's story, he describes what Jane really looks like and depicts Jane using cocaine. The story is very detailed. Will posts his story online publicly, and he includes Jane's first and last name. Though this story is made up, a reader cannot easily tell. Will does not indicate that it is fake when he posts it.

Deepfake Cocaine-Use Video

Imagine Jane is a friend of Will. Will finds a series of photos of Jane online. Will takes the photos and uses an app to merge Jane's face onto a video of someone else. The final video shows Jane's face on the body of a woman who is using cocaine. Jane's face is clearly identifiable in the video. Will decides to post the video online, and he includes Jane's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Self-Insult

Imagine Jane is a friend of Will. Will finds a series of photos of Jane online. Will takes the photos and uses an app to merge Jane's face onto a video of someone else. The final video depicts Jane calling herself a jerk. Jane's face is clearly identifiable in the video. Will has also used software to simulate Jane's voice, so the voice in the video sounds exactly like Jane's real voice. Will decides to post the video online, and he includes Jane's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Scientist Biography, Dead

Imagine Will runs an enthusiast's website about science. Will finds a series of photos of a famous scientist online. The scientist died ten years ago. Will takes the photos and uses an app to merge the scientist's face onto a video of someone else. The final video depicts the scientist talking about their life and accomplishments. The scientist's face is clearly identifiable in the video. Will has also used software to simulate the scientist's voice, so the voice in the video sounds exactly like the scientist's real voice. Will decides to post the video online, and he includes the scientist's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Scientist Biography, Living

Imagine Will runs an enthusiast's website about science. Will finds a series of photos of a famous scientist online. The scientist has just recently retired. Will takes the photos and uses an app to merge the scientist's face onto a video of someone else. The final video depicts the scientist talking about their life and accomplishments. The scientist's face is clearly identifiable in the video. Will has also used software to simulate the scientist's voice, so the voice in the video sounds exactly like the scientist's real voice. Will decides to post the video online, and he includes the scientist's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

C. Politician Attitudinal Scenarios

Written Handshake-with-Child-Molester Story

Imagine Will has written a story about a politician. In Will's story, he states that the politician is friends with a convicted child molester. The story is very detailed. Will posts his story online publicly, and he includes the politician's first and last name. Though this story is made up, a reader cannot easily tell. Will does not indicate that it is fake when he posts it.

Deepfake Handshake-with-Child-Molester Video

Imagine Will finds a series of photos of a politician online. Will takes the photos and uses an app to merge the politician's face onto a video of someone else. The final video shows the politician's face on the body of a person who is shaking hands with a convicted child molester. The politician's face is clearly identifiable in the video. Will decides to post the video online, and he includes the politician's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Terror Endorsement

Imagine Will finds a series of photos of a politician online. Will takes the photos and uses an app to merge the politician's face onto a video of someone else. The final video shows the politician saying they support a known terrorist organization. The politician's face is clearly identifiable in the video. Will has also used software to simulate the politician's voice, so the voice in the video sounds exactly like the politician's real voice. Will decides to post the video online, and he includes the politician's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Silly Song, No Consent

Imagine Will finds a series of photos of a state-level politician online. Will takes the photos and uses an app to merge the politician's face onto a video of someone else. The final video shows the politician singing a silly song. The politician's face is clearly identifiable in the video. Will has also used software to simulate the politician's voice, so the voice in the video sounds exactly like the politician's real voice. Will decides to post the video online, and he includes the politician's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Silly Song, Consent

Imagine a state-level politician has invited her constituents to make and share silly videos of her for her campaign. This politician represents Will. Will finds a series of photos of the politician online. Will takes the photos and uses an app to merge the politician's face onto a video of someone else. The final video shows the politician singing a silly song. The politician's face is clearly identifiable in the video. Will has also used software to simulate the politician's voice, so the voice in the video sounds exactly like her real voice. The politician has consented to Will making the video. Will decides to post the video online, and he includes the politician's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Polling Place, No Consent

Imagine Will finds a series of photos of a politician online. Will takes the photos and uses an app to merge the politician's face onto a video of someone else. The final video shows the politician telling people where their local polling places are. The politician's face is clearly identifiable in the video. Will has also used software to simulate the politician's voice, so the voice in the video sounds exactly like the politician's real voice. Will decides to post the video online, and he includes the politician's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Deepfake Polling Place, Consent

Imagine a state-level politician has invited her constituents to make and share videos of her telling people the location of their local polling place. This politician represents Will. Will finds a series of photos of the politician online. Will takes the photos and uses an app to merge the politician's face onto a video of someone else. The final video depicts the politician telling

people where their local polling places are. The politician's face is clearly identifiable in the video. Will has also used software to simulate the politician's voice, so the voice in the video sounds exactly like her real voice. The politician has consented to Will making the video. Will decides to post the video online, and he includes the politician's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

APPENDIX C: VARIANTS CONTRASTING DEEPFAKES WITH TRADITIONAL NONCONSENSUAL PORNOGRAPHY

The purpose of this study was to compare nonconsensual deepfake pornography with traditional nonconsensual pornography. The deepfake video scenario below was therefore modified from that used in the prior studies to better mirror the newly created traditional nonconsensual-pornography scenario.

Deepfake Pornographic Video, Ex-Romantic Partner

Imagine Jane used to date her friend Will. After they break-up, Will finds a series of photos of Jane online. Will takes the photos and uses an app to merge her face onto a pornographic video. The final video shows Jane's face on the body of a naked woman masturbating. Jane's face is clearly identifiable in the video, and the video shows the entirety of the naked woman's body. Will posts the video online publicly, and he includes Jane's first and last name. Though this video is made up, a viewer cannot easily tell that it has been altered. Will does not indicate that it is fake when he posts it.

Traditional Nonconsensual Pornography, Ex-Romantic Partner

Imagine Mary used to date her friend James. While they were dating, Mary sent James a video of herself masturbating. James had asked for the video and had promised to keep it private. Mary's face is clearly identifiable in the video, and the video shows the entirety of her naked body. After they break-up, James posts the video online publicly, and he includes Mary's first and last name.