DEEPFAKE LIABILITY*

AYELET GORDON-TAPIERO,** YOTAM KAPLAN,*** GIDEON PARCHOMOVSKY****

Deepfake technology is spinning out of control. Recent advancements allow users to quickly, easily, and anonymously create fake yet highly realistic images and videos featuring real people. While this technology has potential benefits, it is widely used nefariously to create pornographic images and videos of young girls. Ninety-eight percent of all deepfake videos online are pornographic deepfakes. The number of deepfake pornography videos has gone up 464% since 2019 and is expected to continue increasing as technology develops and becomes even more accessible.

Scholars have taken note of this frightening trend, seeking to establish some form of liability for the harms generated by sexual deepfakes. This research has focused on two key concepts: user accountability and platform accountability. Yet scholars also show that these two theories of liability are bound to fail. Identifying users who create deepfakes is extremely challenging at best, and usually unhelpful, as they are typically judgment proof. And platforms, where deepfakes are widely shared, are protected from liability by Section 230 of the Communications Decency Act.

Against this backdrop, this Article offers a third solution by proposing a novel approach to deepfake liability hitherto not discussed in the literature: common-law manufacturer accountability. This proposal focuses on the companies developing generative Artificial Intelligence ("AI") tools and offers a natural response to the deepfake crisis with a high likelihood of success. The companies developing generative AI models are manufacturing dangerous and unsafe tools,

** Research fellow, Benin School of Computer Science and Engineering. The authors wish to thank Brenda Dvoskin, Katrina Ligett, Paul Ohm, Roy Shapira, Kobi Kastiel, Noam Kolt, Kobbi Nissim, Catherine Sharkey, and Yonatan Belinkov for helpful comments and discussions. For excellent research assistance, we thank Roy Ashkenazi and Shay Hay. Co-funded by the Fritz Family Fellowship and the European Union (ERC, UEPP, 101077050). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work was carried out with support by the ERC Grant, by a gift to the McCourt School of Public Policy and Georgetown University, Simons Foundation Collaboration 733792, and the Dieter Schwarz Foundation's TUM-HUJI Joint AI Research Hub.

*** Frieda & Solomon Rosenzweig Professor of Law, HU Law School, Visiting Professor, Washington University Law School.

**** Robert G. Fuller, Jr. Professor of Law, University of Pennsylvania Law School.

with no effective safeguards to minimize associated harms. These companies can therefore be found liable for the resulting harm based on familiar concepts of products liability law and design defect. These companies are not protected under Section 230, are incredibly profitable and powerful, and have a crucial impact on the AI ecosystem.

We show that manufacturer liability will provide much-needed protection to the victims of sexual deepfakes. We also show that, by inducing AI companies to include basic safety measures in their products, our proposal will have positive spillover effects on political deepfakes, where the misuse of AI technology currently threatens to destabilize democratic order and to expose democratic processes to dangerous foreign manipulations.

Our proposal addresses a critical oversight in current research and regulatory discourse aimed at tackling the deepfake crisis. The widespread integration of generative AI into user-friendly interfaces has made this powerful technology accessible even to young children, who can now use it to inflict significant harm with minimal technical skills. This urgent reality necessitates a new approach—one that reassigns liability and responsibility to the companies that develop and deploy the technology enabling deepfakes.

INTRODUCTION 379
I. THE STATE OF DEEPFAKES 383
A. *Deepfake Technology* 385
1. Image Generation 386
2. Video Generation 388
B. *Generative-AI Manufacturers* 390
C. *Deepfake Harm* 394
1. Sexually Explicit Deepfakes 396
2. Political Disinformation 399
II. THE STATE OF THE LAW 402
A. *Legislation & Regulation* 403
B. *Platform Liability* 407
C. *User Liability* 408
III. THE PROPOSAL: MANUFACTURER LIABILITY 409
A. *Products Liability Law* 411
B. *Design Defect* 412
1. Traceability 415
2. Nonidentifiability 417
3. Context Matters 417
C. *Duty of Care* 418
D. *Remedies* 421

1. Compensation for Harm 421
2. Disgorgement of Profits 422
IV. CHALLENGES AND IMPLICATIONS 424
A. *Comparing Modalities of Liability* 424
B. *Standing & Spillover* 426
C. *Friction* 427
D. *Watermarking* 428
E. *Open-Source Release of Code* 430
CONCLUSION 431

INTRODUCTION

The use of deepfakes, which reached an unprecedented high last year, is among the foremost challenges confronting our society. Recent technological advancements have made the creation of deepfakes incredibly easy.[1] With the aid of generative AI technology, any user can swiftly and easily create highly realistic, though fake, images and videos of other individuals. While generative AI technology no doubt has many beneficial applications, it is increasingly being misused to produce deepfake pornography targeting innocent high schoolers.[2] The phenomenon is becoming alarmingly common, and deepfake creators are subjecting their victims to bullying and humiliation.[3] The number of deepfake pornography videos went up a staggering 464% between 2022 and 2023.[4] Ninety-six percent of deepfakes posted online reportedly depict women in nonconsensual pornography.[5] Adolescents are particularly vulnerable to this

1. Kevin Roose, *An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy.*, N.Y. TIMES (Sep. 2, 2022), https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html [https://perma.cc/DTK5-7NNZ (staff-uploaded, dark archive)] ("A.I.-generated art has been around for years. But tools released this year—with names like DALL-E 2, Midjourney and Stable Diffusion—have made it possible for rank amateurs to create complex, abstract or photorealistic works simply by typing a few words into a text box.").

2. Danielle Keats Citron, *Sexual Privacy*, 128 YALE L.J. 1870, 1870 (2019) [hereinafter Citron, *Sexual Privacy*] (describing the harms caused by sexual deepfakes).

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

4. *2023 State of Deepfakes: Realities, Threats and Impact*, SEC. HERO, https://www.securityhero.io/state-of-deepfakes/#overview-of-current-state [https://perma.cc/K4FQ-TTK5].

5. Vandinika Shukla, *Deepfakes and Elections: The Risk to Women's Political Participation*, TECH POL'Y PRESS (Feb. 29, 2024), https://www.techpolicy.press/deepfakes-and-elections-the-risk-to-womens-political-participation/ [https://perma.cc/A788-SXH4].

abuse, and victims have been reported to struggle with lasting anxiety[6] and distress[7] and even develop suicidal tendencies.[8]

In their article, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, law professors Bobby Chesney and Danielle Citron analyze this threat, as well as potential legal responses.[9] They discuss two legal approaches to deepfakes: *user accountability*,[10] focusing on the individuals who create deepfakes,[11] and *platform accountability*,[12] focusing on the social media platforms on which deepfakes are shared.[13] As Chesney and Citron show, both approaches lead to a dead end. User accountability is an impractical solution, since policing the vast number of users is an administrative and regulatory nightmare.[14] For the same reason, systematically identifying deepfake creators is an unrealistic option.[15] Platform accountability is even less attainable, as it is explicitly blocked by Section 230 of the Communications Decency Act,[16] which shelters platforms from liability for content created by third parties.[17]

Against this backdrop, this Article offers a third solution: *manufacturer accountability*, focusing on the companies developing new AI tools.[18] We argue that this approach is a much more natural response to the crisis and much more likely to succeed. The crisis is driven by new technological changes and by the

6. Danielle K. Citron, *Why Sexual Privacy Matters for Trust*, 96 WASH. U. L. REV. 1189, 1208 (2019).

7. *See* DANIELLE KEATS CITRON, THE FIGHT FOR PRIVACY: PROTECTING DIGNITY, IDENTITY, AND LOVE IN THE DIGITAL AGE 10–11 (2022) [hereinafter CITRON, THE FIGHT FOR PRIVACY]; Michael P. Goodyear, *Deepfakes and Dignity*, ARIZ. ST. L.J. (forthcoming 2025) (manuscript at 6) (on file with the North Carolina Law Review).

8. Citron, *Sexual Privacy*, *supra* note 2, at 1926 (reporting that some victims of nonconsensual pornography contemplate suicide).

9. Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1792 (2019) [hereinafter Chesney & Citron, *Deep Fakes*].

10. *Id.* at 1793.

11. *Id.*

12. *Id.* at 1795.

13. *Id.*

14. *Id.* at 1792–93.

15. *Id.* at 1792 ("Civil liability cannot ameliorate harms caused by deep fakes if plaintiffs cannot tie them to their creators.").

16. Communications Decency Act of 1996, Pub. L. No. 104-104, § 230, 110 Stat. 133 (1996) (codified as amended at 47 U.S.C. § 230(c)(1)–(2)).

17. *Id.* § 509; *see* Danielle Keats Citron & Mary Anne Franks, *The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform*, 2020 U. CHI. LEGAL F. 45, 45–48 (2020) (detailing the protection afforded by Section 230 and the urgent need for its reform).

18. Ayelet Gordon-Tapiero, *A Liability Framework for AI Companions*, 1 GEO. WASH. J.L. & TECH. (forthcoming 2025) (on file with the North Carolina Law Review) (suggesting applying products liability to the developers of AI companions); Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1793–95. Chesney and Citron analyze two options for deepfake liability. First, they study the possibly of holding users who created deepfakes liable. *Id.* Second, they study the possibility of liability for platforms where deepfakes are posted and shared. *Id.* at 1795–1801. The authors do not suggest nor analyze the option of attributing liability to the companies developing and deploying the generative AI models used to generate deepfakes. *See generally id.*

emergence of easy-to-use and immensely powerful generative AI tools. The companies developing generative AI tools enable the production of deepfakes. Without their technologies, the scope and magnitude of the problem would have been much smaller. Furthermore, AI producers reap the profits of their technology but are presently completely oblivious to the cost. Finally, they are in the best position to make critical changes that would limit the harms stemming from deepfakes.[19] Hence, we propose that liability should be imposed on AI technology producers on both fairness and efficiency grounds. AI technology is a product, and the time has come to apply the law of product liability to AI producers.

In this vein, we show that producer accountability can be operationalized through the established framework of products liability law. Surprisingly, previous scholarship has not considered this option. In recent years, generative AI technology has changed in ways that allow the models to be more powerful, more accessible, simpler to use, and able to produce more believable deepfakes. The developing companies are keenly aware of the harm their product is causing and are best positioned to introduce safeguards that would limit such harm. This reality requires adopting a new approach—one that will shift liability and responsibility to the companies developing and deploying the technology underlying deepfakes.

Our proposal to implement products liability to the companies developing deepfake technology comes with two suggestions. First, courts can use the familiar concept of *design defect* to establish the liability of companies that fail to integrate necessary safeguards into their products. Companies that wish to avoid such liability will need to introduce safeguards ensuring their products do not facilitate easy, anonymous, and untraceable creation of sexually explicit deepfakes of identifiable individuals. Second, we suggest that implementing such basic safeguards will have positive spillover effects on other, closely related issues. For example, in addition to sexual deepfakes, political deepfakes are becoming a major concern, threatening to destabilize democratic processes. As we illustrate in more detail below, products liability law may not enable liability in the context of political deepfakes.[20] However, the inclusion of some basic safeguards in generative AI tools may also indirectly mitigate some of the harms associated with political deepfakes.[21]

19. *See infra* Section I.B.

20. *See infra* Section III.C. This is closely related to issues of political free speech and First Amendment jurisprudence. David A. Logan, *Rescuing Our Democracy by Rethinking* New York Times v. Sullivan, 81 OHIO ST. L.J. 759, 772 (2020) ("[P]olitical speech is at the heart of the First Amendment."); Jessica Ice, Note, *Defamatory Political Deepfakes and the First Amendment*, 70 CASE W. RSRV. L. REV. 417, 418 (2019).

21. *See infra* Section IV.B.

Our proposal does not reflect a desire to prevent technological development of generative AI tools. Quite the opposite. As AI technology is becoming an integral part of our daily lives, it is important to ensure that it is safe, and especially that it is not harmful to vulnerable members of society. To the extent that it is not safe, the law should induce producers of AI technologies to improve design defects, and at the very least, to compensate the victims. We submit that our proposal can facilitate safe and responsible development of generative AI in a way that will allow individuals and society to enjoy the promises it holds while limiting some of the most egregious harms stemming from its misuse.

In making these arguments, this Article offers three novel contributions. The first contribution is conceptual, identifying a critical omission in existing literature and policy debates. Scholars and policymakers alike have sought to address the challenge of deepfakes by targeting either users or platforms, overlooking the option of imposing liability on the companies developing and deploying generative AI models. The second contribution is normative. We show that imposing liability on AI producers outperforms the alternatives of user and platform liability and thus marks the best way forward. Our third and final contribution is doctrinal. We highlight the legal elements necessary to establish a product liability doctrine in the context of generative AI tools. Specifically, we construct a harm-based and a gain-based liability regime for ensuring full compensation of victims of deepfakes created by generative AI technologies.

The Article proceeds as follows. Part I describes the deepfake crisis, its origins, and its consequences. It describes the generative AI technology facilitating the creation of deepfakes[22] and provides an in-depth analysis of the two domains in which the use of deepfakes has reached pandemic-like prevalence: sexually explicit deepfakes[23] and political deepfakes.[24] Part II studies existing and proposed legal responses to the deepfake crisis and highlights their limitations. We describe existing legislative and regulatory efforts,[25] as well as the scholarly focus on the concepts of platform liability[26] and user liability.[27] The analysis in Part II shows that existing efforts virtually ignore the responsibility of the companies developing and deploying the generative AI models used to create deepfakes. Part III is the conceptual core of the Article and offers a reorientation of the legal response to the deepfake crisis by introducing the concept of manufacturer liability into this ongoing

22. *See infra* Section I.A.
23. *See infra* Subsection I.C.1.
24. *See infra* Subsection I.C.2.
25. *See infra* Section II.A.
26. *See infra* Section II.B.
27. *See infra* Section II.C.

conversation. We discuss the basic doctrinal elements necessary to establish such a claim[28] and apply them in the context of the development of generative AI tools.[29] We analyze the doctrinal elements of design defect[30] and the duty of care,[31] and we discuss the possibility of both harm-based[32] and gain-based[33] remedies. Part IV offers a broader discussion of the implications of our proposal and the practical aspects of its implementation. This part compares our proposal for manufacturer liability with the familiar concepts of platform liability and user liability;[34] highlights the possible positive third-party effects of our proposal;[35] and considers technological challenges relating to the concepts of friction,[36] watermarking technology,[37] and open-source code.[38] A short conclusion follows.

I. THE STATE OF DEEPFAKES

Deepfakes[39] are a form of digital impersonation;[40] they replace the face, body, or voice of one person with another person's likeness, enabling the creator to generate a perception of images or events that never took place.[41] The term "deepfake" stems from a combination of the terms "deep learning" and "fake

28. *See infra* Section III.A.
29. *See infra* Sections III.B–C.
30. *See infra* Section III.B.
31. *See infra* Section III.C.
32. *See infra* Subsection III.D.1.
33. *See infra* Subsection III.D.2.
34. *See infra* Section IV.A.
35. *See infra* Section IV.B.
36. *See infra* Section IV.C.
37. *See infra* Section IV.D.
38. *See infra* Section IV.E.
39. Professor Richard Hasen defines deepfakes as "audio and video clips [that] can be manipulated using machine learning and artificial intelligence and can make a politician, celebrity, or anyone else appear to say or do anything that the manipulator wants." Richard L. Hasen, *Deep Fakes, Bots, and Siloed Justices: American Election Law in a "Post-Truth" World*, 64 ST. LOUIS U. L.J. 535, 542 (2020); *see also* Lucas Whittaker, Kate Letheren & Rory Mulcahy, *The Rise of Deepfakes: A Conceptual Framework and Research Agenda for Marketing*, 29 AUSTRALASIAN MKTG. J. 204, 204 (2021).
40. Ice, *supra* note 20, at 418.
41. *Id.*; *see also* Regina Rini & Leah Cohen, *Deepfakes, Deep Harms*, 22 J. ETHICS & SOC. PHIL. 143, 143 (2022).

content."[42] The ability to manipulate content can be helpful in certain settings.[43] In the entertainment and movie industries, AI tools can help alter a person's appearance, generate new characters, age a person, and safely create special effects and stunts.[44] For educators, these tools can help bring past events to life, making them more accessible to students.[45] Algorithmically generated images can serve as a helpful learning and diagnostic tool for medical practitioners,[46] and the list goes on.[47] As with other forms of "modern science[, they] are not in themselves good or bad; it is the way they are used that determines their value."[48]

This part will proceed in three sections: Section A offers an overview of how deepfake technology works; Section B describes the actors in the AI market and how valuable they have become; and Section C discusses the intense harms that have begun to arise from deepfake technology, including political

42. Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham & Cuong M. Nguyen, *Deep Learning for Deepfakes Creation and Detection: A Survey*, 223 COMPUT. VISION & IMAGE UNDERSTANDING 1, 1 (2022); Robert Chesney & Danielle Citron, *Deepfakes: A Looming Crisis for National Security, Democracy and Privacy?*, LAWFARE (Feb. 21, 2018, at 10:00 ET), https://www.lawfaremedia.org/article/deepfakes-looming-crisis-national-security-democracy-and-privacy [https://perma.cc/5Q5F-U8LC]; Alan Zucconi, *An Introduction to Neural Networks and Autoencoders*, ALAN ZUCCONI (Mar. 14, 2018), https://www.alanzucconi.com/2018/03/14/an-introduction-to-autoencoders/ [https://perma.cc/MYA3-MJ7T]; Morgan Meaker, *Deepfake Audio is a Political Nightmare*, WIRED (Oct. 9, 2023, at 11:30 ET), https://www.wired.com/story/deepfake-audio-keir-starmer/ [https://perma.cc/ER94-T5B4 (dark archive)].

43. Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1768–69.

44. Jace Dela Cruz, *AI Anxiety: Hollywood Stunt Workers Fear Being Replaced by 'Digital Replicas*,' TECH TIMES (Aug. 12, 2023, at 04:07 ET), https://www.techtimes.com/articles/295032/20230812/ai-anxiety-hollywood-stunt-workers-fear-replaced-digital-replicas.htm [https://perma.cc/E2U6-XZ3U]; Stewart Townsend, *Exploring the Impact of AI on Film Production in 2024*, MEDIUM (Mar. 2, 2024), https://medium.com/@channelasaservice/exploring-the-impact-of-ai-on-film-production-in-2024-f02da745af00 [https://perma.cc/TY56-MEKF]; Rick Spair, *The Rise of AI in Hollywood*, MEDIUM, https://medium.com/@dxtoday/the-rise-of-ai-in-hollywood-how-technology-is-changing-the-movie-industry-innovation-technology-0677faa67886 [https://perma.cc/2P4Q-YBDS] (last updated May 22, 2024).

45. Dan Patterson, *Deepfakes for Good? How Synthetic Media Is Transforming Business*, TECH INFORMED (Oct. 5, 2023), https://techinformed.com/deepfakes-for-good-how-synthetic-media-is-transforming-business/ [https://perma.cc/FHA5-3J8V] ("Deepfake algorithms can animate historical photos and footage, allowing influential figures to give speeches and presentations as if they were in the classroom. The resulting videos are far more engaging and interactive than lectures or textbooks.").

46. Nagendra Rao, *Deepfakes: Healthcare's Future Is Here, and It's Not What You Expect*, TRIGENT SOFTWARE, INC.: BLOG (Feb. 29, 2024), https://trigent.com/blog/deepfakes-healthcares-future-is-here-and-its-not-what-you-expect/ [https://perma.cc/J2DK-WN8T]; ORLY LOBEL, THE EQUALITY MACHINE 140 (1st ed. 2022) (describing the contribution of AI to the medical field); Kim Martineau, *Generative AI Could Offer a Faster Way To Test Theories of How the Universe Works*, IBM: Blog (Mar. 14, 2024), https://research.ibm.com/blog/time-series-AI-transformers [https://perma.cc/CKN4-SSZS].

47. Orly Lobel, *The Law of AI for Good*, 75 FLA. L. REV. 1073, 1080 (2023) [hereinafter Lobel, *The Law of AI for Good*] (detailing the benefits of generative AI).

48. This quote is widely attributed to General David Sarnoff. *See, e.g.*, MARSHALL MCLUHAN, UNDERSTANDING MEDIA: THE EXTENSIONS OF MAN 26 (1964).

disinformation and the dissemination of sexually explicit images that appear real. With such potential harm to society and democracy, it is important to begin this Article by exploring the current state of generative AI tools and the manufacturers behind them.

A. *Deepfake Technology*

The field of generative AI has gained prominence in recent years.[49] This development has several driving forces. First, the technology driving generative AI has experienced profound developments.[50] Improvements in neural network architecture have significantly enhanced the capabilities of generative AI models and applications.[51] Companies have also been able to secure more computational power and graphics processing units ("GPUs") necessary for the development of these sophisticated models.[52] In addition, companies have also increased their data use, constructing gargantuan datasets comprised of large

49. *See* Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1753; Ayelet Gordon-Tapiero & Yotam Kaplan, *Generative AI Training as Unjust Enrichment*, 86 OHIO ST. L.J. 285, 287 (2025) ("Generative AI has made a wild leap forward and is disrupting the way we work, learn, teach, research, and think."); Ajay Bandi, Pydi Venkata Satya Ramesh Adapa & Yudu Eswar Vinay Pratap Kumar Kuchi, *The Power of Generative AI: A Review of Requirements, Models, Input-Output Formats, Evaluation Metrics, and Challenges*, FUTURE INTERNET (SPECIAL ISSUE) 1–2 (2023) ("[T]he worldwide market for generative AI . . . is projected to reach . . . a compound annual growth rate . . . of 27.02% . . . from 2023 to 2032."); Patrick Parsons, *AI in 2024: A Year of Crossroads and Decisions*, 40 GA. ST. U. L. REV. ix, ix (2024).

50. *See* Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mizra, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville & Yoshua Bengio, *Generative Adversarial Nets* 1 (arXiv, Working Paper No. 1406.2661, 2014), https://arxiv.org/pdf/1406.2661 [https://perma.cc/G8A3-B996]; Ryan Abbott & Elizabeth Rothman, *Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence*, 75 FLA. L. REV. 1141, 1141 (2023).

51. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin, *Attention Is All You Need* 1 (arXiv, Working Paper No. 1706.03762, 2023), https://arxiv.org/pdf/1706.03762 [https://perma.cc/UD2M-AM7R] (introducing the Transformer architecture, a neural network at the base of leading generative AI models); Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville & Yoshua Bengio, *Generative Adversarial Networks,* 11 COMM. ACM 139, 139 (2020); Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova, *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding* 2 (arXiv, Working Paper No. 1810.04805, 2019), https://arxiv.org/pdf/1810.04805 [https://perma.cc/F2DM-NG89].

52. *See* Hongyang Du, Dusit Niyato, Jiawen Kang, Zehui Xiong, Ping Zhang, Shuguang Cui, Xuemin (Sherman) Shen, Shiwen Mao, Zhu Han, Abbas Jamalipour, H. Vincent Poor & Dong In Kim, *The Age of Generative AI and AI-Generated Everything* 5 (arXiv, Working Paper No. 2311.00947, 2023), https://arxiv.org/pdf/2311.00947 [https://perma.cc/Q7GT-4HNA] ("Transformer-based decoders[] have brought efficiency . . . accuracy and computational speed."); Arvind Narayanan, *I Set Up a ChatGPT Voice Interface for My 3-Year Old. Here's How It Went.*, AI SNAKE OIL (Apr. 14, 2023), https://www.aisnakeoil.com/p/i-set-up-a-chatgpt-voice-interface [https://perma.cc/LJD2-9V5U]; *Generative AI: Risks and Opportunities for Children*, UNICEF, https://www.unicef.org/innocenti /generative-ai-risks-and-opportunities-children [https://perma.cc/DGY3-Y435 (staff-uploaded archive)].

parts of the internet, which some argue are the secret sauce of today's generative AI models.[53]

One of the salient features of recent generative AI applications is their intuitiveness and ease of use.[54] ChatGPT was a breakthrough in this respect. For the first time, people with no coding expertise or technical experience could directly interact with a sophisticated form of generative AI through a simple, accessible, and intuitive interface.[55] Children with minimal digital literacy can easily use generative AI tools—not always in positive ways.[56] In particular, recent advancements in generative AI technology have brought about a flurry of applications capable of generating deepfake images and videos.[57]

1. Image Generation

Image generation, the creation of an image by a generative AI model based on input by a human user, typically operates in one of two ways: text-to-image or image-to-image. Text-to-image models allow users to convert text prompts

53. Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving & Iason Gabriel, *Ethical and Social Risks of Harm from Language Models* 8 (arXiv, Working Paper No. 2112.04359, 2021) https://arxiv.org/pdf/2112.04359 [https://perma.cc/M8RK-8NSD] ("Training models on extremely large datasets such as the Colossal Clean Crawl Corpus (C4) and WebText resulted in sequence prediction systems with much more general applicability compared to the prior state-of-the-art." (citations omitted)); Dario Amodei, Alec Radford, Tom Brown, Sam McCandlish, Nick Ryder, Jared Kaplan, Sandhini Agarwal, Amanda Askell, Girish Sastry & Jack Clark, *Language Models Are Few-Shot Learners* 8 (arXiv, Working Paper No. 2005.14165, 2020), https://arxiv.org/pdf/2005.14165 [https://perma.cc/BM6F-YSPQ] ("Datasets for language models have rapidly expanded.").

54. John Werner, *Kids Can Use AI, Too—Look What They're Coming Up With . . .*, FORBES (Nov. 22, 2023, at 09:00 ET), https://www.forbes.com/sites/johnwerner/2023/11/22/kids-can-use-ai-too-look-what-theyre-coming-up-with/ [https://perma.cc/X9A9-3GXK (dark archive)] ("The simplicity of the interface . . . is going to provide more access to kids.").

55. *See* Gary Marcus, *AI Platforms Like ChatGPT Are Easy To Use But Also Potentially Dangerous,* SCIAM (Dec. 19, 2022), https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous/ [https://perma.cc/2GXQ-KMBW (dark archive)].

56. Kevin Kelly, *Picture Limitless Creativity at Your Fingertips*, WIRED (Nov. 17, 2022, at 06:00 ET), https://www.wired.com/story/picture-limitless-creativity-ai-image-generators/ [https://perma.cc/SX7F-CURK] ("Not everyone can write, direct, and edit an Oscar winner like *Toy Story 3* or *Coco*, but everyone *can* launch an AI image generator and type in an idea."); *see, e.g.*, Nikolas Lanum, *Snapchat AI Chatbot Allegedly Gave Advice to 13-Year-Old Girl on Relationship with 31-Year-Old Man, Having Sex*, FOX NEWS (Apr. 13, 2023, at 14:30 ET), https://www.foxnews.com/media/snapchat-ai-chatbot-gave-advice-13-year-old-girl-relationship-31-year-old-man-having-sex [https://perma.cc/B9JD-3FPM].

57. Matt Burgess, *Deepfake Creators Are Revictimizing GirlsDoPorn Sex Trafficking Survivors*, WIRED (June 25, 2024, at 06:00 ET), https://www.wired.com/story/girlsdoporn-deepfake-victim-videos/ [https://perma.cc/8ZJX-AQLC]. The developments in generative adversarial network ("GAN") technologies have been particularly meaningful in pushing this field forward. *See* Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1760; *see also* Goodfellow et al., *supra* note 50, at 144.

into visual representations.[58] Datasets used for text-to-image generation usually include image-text pairs in which the text describes the content of the image.[59] Some such datasets include billions of image-text pairs.[60] The neural networks used in the training of text-to-image models analyze the connections between images and their accompanying text, learning to predict what text is most statistically likely to describe which images.[61] Once these pairings have been learned using supervised deep learning techniques, the neural network trains to maximize the similarity between text and images as learned from its training data.[62] When a user inputs a text prompt describing the image they want generated, the model interprets the semantic meaning of the text to know what objects to include in the output and the style, context, or atmosphere in which to set it.[63]

Image-to-image models are a subset of image generation models that can receive images as input and generate image outputs after they have undergone modification or enhancements. These models can generate a wide array of alterations including changing the background of a picture (transporting people from an office setting to a rock concert); switching out an element in an image (switching out an ex to your current partner); changing style (editing a sky to Van Gogh style); and combining elements from different images (using the head of one person and the body of another).[64]

58. 3D text models are another relatively recent development in the field of generative AI. "In a single week this past September, three novel text-to-3D/video image generators were announced: GET3D (NVIDIA), Make-A-Video (Meta), and DreamFusion (Google)." Kelly, *supra* note 56.

59. Edward Lee, *Prompting Progress: Authorship in the Age of AI*, 76 FLA. L. REV. 1445, 1458 (2024) (describing the dataset used by Stable Diffusion to train its text-to-image model).

60. Anthony Alford, *LAION Releases Five Billion Image-Text Pair Dataset LAION-5B*, INFOQ (May 17, 2022), https://www.infoq.com/news/2022/05/laion-5b-image-text-dataset/ [https://perma.cc/LE5A-SQ5C].

61. *See* Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger & Ilya Sutskever, *Learning Transferable Visual Models from Natural Language Supervision* 7–16 (arXiv, Working Paper No. 2103.00020, 2021), https://arxiv.org/pdf/2103.00020 [https://perma.cc/2D29-EZEK]; *CLIP: Connecting Text and Images*, OPENAI (Jan. 5, 2021), https://openai.com/blog/clip/ [https://perma.cc/CS7D-TKMG (staff-uploaded archive)].

62. Priynashu Prasad, *Artificial Faces: The Encoder-Decoder and GAN Guide to Deepfakes*, MEDIUM (Apr. 2, 2024), https://medium.com/@priyanshuprasad1718/artificial-faces-the-encoder-decoder-and-gan-guide-to-deepfakes-75a1eed0e265 [https://perma.cc/JS3A-5MW7].

63. *See generally* Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim & Jinwook Seo, *Large-Scale Text-to-Image Generation Models for Visual Artists' Creative Works* (arXiv, Working Paper No. 2210.08477, 2023), https://arxiv.org/pdf/2210.08477 [https://perma.cc/PY2L-5WU8] (describing the operation of text-to-image models); Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang & In So Kweon, *Text-to-Image Diffusion Models in Generative AI: A Survey* 1–7 (arXiv, Working Paper No. 2303.07909, 2024), https://arxiv.org/pdf/2303.07909 [https://perma.cc/9YBP-SED3] (detailing how text-to-image models operate).

64. *See* Narek Tumanyan, Michal Geyer, Shai Bagon & Tali Dekel, *Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation* 7–10 (arXiv, Working Paper No. 2211.12572, 2022),

2. Video Generation

Video generation models have made significant advancements in recent years.[65] High-quality videos require a high level of temporal coherence,[66] meaning that the frames comprising the video flow naturally into each other, rather than transition abruptly.[67] Another challenge in video generation involves generating spatial consistency within each frame and across frames. Lack of spatial consistency can result in unrealistic textures and blurry elements.[68] Recent developments in spatial consistency have significantly improved the ability to overcome these challenges.[69]

Generating video from text requires a strong understanding of language; the ability to translate semantic input such as objects, spaces, actions, and scenes into video output; and a large dataset to train on.[70] Text-to-video generators are trained on large numbers of pairs of videos and corresponding text

https://arxiv.org/pdf/2211.12572 [https://perma.cc/JV83-BKFP]; Hyunsoo Lee, Minsoo Kang & Bohyung Han, *Conditional Score Guidance for Text-Driven Image-to-Image Translation* 1 (arXiv, Working Paper No. 2305.18007, 2023), https://arxiv.org/pdf/2305.18007 [https://perma.cc/W597-8GBC] (describing text driven image-to-image generation models).

65. *See, e.g.*, Andreas Blattman, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Domink Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Kampani & Robin Rombach, *Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Dataset* 1–2 (arXiv Working Paper No. 2311.15127, 2023), https://arxiv.org/pdf/2311.15127 [https://perma.cc/THN7-HKC7].

66. *See* Mengyu Chu, You Xie, Jonas Mayer, Laural Teal-Taixé & Nils Thuerey, *Learning Temporal Coherence via Self-Supervision for GAN-Based Video Generation* 1 (arXiv, Working Paper No. 1811.09393, 2020), https://arxiv.org/pdf/1811.09393 [https://perma.cc/S8N5-3XF6].

67. *Id.*

68. *But see* Sabrina Ortiz, *The Best AI Image Generators Are Getting Scary Good at Things They Used to Be Terrible At*, ZDNET, https://www.zdnet.com/article/best-ai-image-generator/ [https://perma.cc/63PP-497J (staff-uploaded archive)] (last updated May 9, 2025, at 14:29 PT); Ema Lukan, *The 13 Best AI Video Generators (Free & Paid) to Try in 2025*, SYNTHESIA: THE SYNTHESIA BLOG (June 11, 2025), https://www.synthesia.io/post/best-ai-video-generators [https://perma.cc/V57Z-CLEW].

69. *See* Barry Solaiman, *Addressing Access with Artificial Intelligence: Overcoming the Limitations of Deep Learning to Broaden Remote Care Today*, 51 U. MEMPHIS L. REV. 1103, 1109 (2021) (providing a review of AI, machine learning and deep learning); Niels Justesen, Philip Bontrager, Julian Togelius & Sebastian Risi, *Deep Learning for Video Game Playing*, 12 IEEE TRANSACTIONS ON GAMES 1, 5–6 (2020); Arun Rai, *Explainable AI: From Black Box to Glass Box*, 48 J. ACAD. MKTG. SCI. 137, 138 (2020) ("[D]eep learning algorithms are a class of ML algorithms which sacrifice transparency and interpretability for prediction accuracy."). *See generally* Satyam Kumar, Dayima Musharaf, Seerat Musharaf & Anil Kumar Sagar, *A Comprehensive Review of the Latest Advancements in Large Generative AI Models*, *in* 1 ADVANCED COMM. & INTEL. SYS. 90, 91–101 (Rabindra Nath Shaw, Marcin Paprzycki & Ankush Ghosh eds.) (2023) (describing recent developments in the field of generative AI models).

70. Ganchao Tan, Daqing Liu, Meng Wang & Zheng-Jun Zha, *Learning To Discretely Compose Reasoning Module Networks for Video Captioning*, 29 INT'L JOINT CONF. ON A.I. 745, 745 (2020) https://www.ijcai.org/proceedings/2020/0104.pdf [https://perma.cc/4P5M-FJR5] ("For example, to generate the sentence 'a man is shooting a basketball,' we need to first locate and describe the subject 'man,' next reason out the man is 'shooting,' then describe the object 'basketball' of shooting.").

descriptions.[71] Some datasets can include over 100 million such pairs,[72] depicting tens of thousands of activities from different categories.[73]

Generative AI also enables the creation of videos based on a single image input.[74] To produce a high-quality video, an image-to-video model must achieve temporal consistency and smoothness of the video while maintaining visual fidelity to the input image.[75] Image-to-video generation models are trained on large datasets comprised of pairs of videos along with a key frame from them or image sequences.[76] This allows them to create a full video from a single image or frame. Such datasets can include hundreds of thousands of video clips.[77]

Some image and video generation models include restrictions on the type of outputs the model can generate, depending on the policy of the developing

71. *See, e.g.*, Jun Xu, Tao Mei, Ting Yao & Yong Rui, *MSR-VTT: A Large Video Description Dataset for Bridging Video and Language*, 2016 IEEE CONF. ON COMP. VISION & PATTERN RECOGNITION 5288, 5290–91 (2016), https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7780940 [https://perma.cc/9J45-8VWV (staff-uploaded archive)] (describing one such dataset, MSR-VTT, which includes 10,000 web video clips from twenty categories, with each video clip is annotated with twenty English sentences).

72. Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev & Josef Sivic, *HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips* 1–4 (arXiv, Working Paper No. 1906.03327, 2019), https://arxiv.org/pdf/1906.03327 [https://perma.cc/2CN7-2V48] (describing HowTo100M, which includes 136 million captioned clips sourced from fifteen years of YouTube videos).

73. Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev & Josef Sivic, *What is HowTo100M?*, HOWTO100M, https://www.di.ens.fr/willow/research/howto100m/ [https://perma.cc/EW7Q-K3YK] (describing 23,611 types of activities from different "domains such as cooking, hand crafting, personal care, gardening or fitness").

74. *See generally* Yaniv Nikankin, Niv Chaim & Michal Irani, *SinFusion: Training Diffusion Models on a Single Image or Video* (arXiv, Working Paper No. 2211.11743, 2023), https://arxiv.org/pdf/2211.11743 [https://perma.cc/3VWN-UHEG] (describing training AI models using one image or video).

75. *See* Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang & Wenhu Chen, *ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation* 2, 8 (arXiv, Working Paper No. 2402.04324, 2024), https://arxiv.org/pdf/2402.04324 [https://perma.cc/BEU2-WQ5F].

76. *See* Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman & Andrew Zisserman, *The Kinetics Human Action Video Dataset* 1–2 (arXiv, Working Paper No. 1705.06950, 2017), https://arxiv.org/pdf/1705.06950 [https://perma.cc/S778-NMJX] (describing the Kinetics dataset, a dataset of image-video pairs). *See generally* Khurram Soomro, Amir Roshan Zamir & Mubarak Shah, *UCF101: A Dataset of 101 Human Actions Classes From Videos in the Wild* (arXiv Working Paper No. 1212.0402, 2012) https://www.crcv.ucf.edu/papers/UCF101_CRCV-TR-12-01.pdf [https://perma.cc/EXB2-7LHL] (describing the UCF101 dataset, a dataset of image-video pairs).

77. Kay et al., *supra* note 76, at 2.

company.[78] Unfortunately, these are sometimes quite easily circumventable, particularly in open-source models, or by using the correct prompts.[79]

B. *Generative-AI Manufacturers*

The market for companies developing image and video generation models is led by several large companies. These include OpenAI, Google, Stability AI, NVIDIA, and Adobe.[80] OpenAI is considered a leader in the field of generative AI, particularly since it released its user-friendly ChatGPT in November 2022.[81] Since then, it has fought to maintain its lead in the generative AI market. Dall-E, the company's text-to-image generator, was trained on twelve billion text-image pairs.[82] The company's most recent version, Dall-E3, can be used and accessed through ChatGPT, making it easily accessible to users who have become acquainted with the chatbot.[83] This is especially helpful since the release of GPT-4, which can receive images as inputs.[84] The company announced its

78. Will Knight, *This Uncensored AI Art Tool Can Generate Fantasies—and Nightmares*, WIRED (Sep. 21, 2022, at 07:00 ET), https://www.wired.com/story/the-joy-and-dread-of-ai-image-generators-without-limits/ [https://perma.cc/5A85-USX4 (staff-uploaded, dark archive)] ("[M]ost of these image generators are designed to restrict what users can depict, banning pornography, violence, and pictures showing the faces of real people."); Benjamin Sobel, *Elements of Style: Copyright, Similarity, and Generative AI*, 38 HARV. J.L. & TECH. 49, 60 (2024) [hereinafter Sobel, *Elements of Style*] ("[T]he major image-generating services . . . place contractual and technological limits on functionality . . . set[ting] certain categories of imagery—such as violence and pornography—off-limits.").

79. Knight, *supra* note 78 ("[B]ecause the full code of the AI model has been released, it has been possible for others to remove those [output-restricting] limits."). The fact that such restrictions are in place serves as a form of friction. *Infra* Section IV.C.

80. *See* Ortiz, *supra* note 68; Lukan, *supra* note 68.

81. Brian X. Chen, Nico Grant & Karen Weise, *How Siri, Alexa and Google Assistant Lost the A.I. Race*, N.Y. TIMES (Mar. 15, 2023), https://www.nytimes.com/2023/03/15/technology/siri-alexa-google-assistant-artificial-intelligence.html [https://perma.cc/59NB-L36R (staff-uploaded, dark archive)]; Richard Nieva, Alex Konrad & Kenrick Cai, *'AI First' to Last: How Google Fell Behind in the AI Boom,* FORBES, https://www.forbes.com/sites/richardnieva/2023/02/08/google-openai-chatgpt-microsoft-bing-ai/ [https://perma.cc/C6X4-8HMH (dark archive)] (last updated Feb. 9, 2023, at 11:20 ET); Chris Stokel-Walker, *OpenAI Overtakes Google in Race to Build the Future, but Who Wants It?*, NEWSCIENTIST (May 15, 2024), https://www.newscientist.com/article/2431326-openai-overtakes-google-in-race-to-build-the-future-but-who-wants-it/ [https://perma.cc/NES8-DRB4 (staff-uploaded, dark archive)].

82. Vivek Kumar Upadhyay, *Dall-E: A Detailed Guide to the Ultimate Image Generator*, MEDIUM (Feb. 20, 2024), https://vivekupadhyay1.medium.com/dall-e-a-detailed-guide-to-the-ultimate-image-generator-0a1ef931f7dd [https://perma.cc/K8EB-E7L4].

83. Reece Rogers, *How To Create Images with ChatGPT's New Dall-E Integration*, WIRED (Oct. 20, 2023, at 07:00 ET) https://www.wired.com/story/how-to-use-chatgpt-dalle-3-create-images/ [https://perma.cc/49SY-RNJM (staff-uploaded, dark archive)]; Cade Metz & Tiffany Hsu, *ChatGPT Can Now Generate Images, Too*, N.Y. TIMES (Sep. 20, 2023), https://www.nytimes.com/2023/09/20/technology/chatgpt-dalle3-images-openai.html [https://perma.cc/M9XQ-5KBK (staff-uploaded, dark archive)].

84. *ChatGPT Image Inputs FAQ*, OPENAI, https://help.openai.com/en/articles/8400551-image-inputs-for-chatgpt-faq [https://perma.cc/6AMW-8YEQ (staff-uploaded archive)].

text-to-video generator, Sora, in February 2024[85] and publicly launched it in December 2024.[86] The public launch of Sora was delayed due to concerns over potential dangers stemming from the wide availability of this type of technology.[87] In February 2024, OpenAI was valued at $80 billion.[88]

Google has long been a pioneer in the data ecosystem, though some argue the company has fallen behind in the AI race.[89] Google developed text-to-image models several years ago, but concerns about their potential applications prevented Google from releasing them publicly.[90] In 2022, Google publicly released its first version of Imagen, a text-to-image model.[91] Google's video

85. *Creating Video from Text*, OPENAI, https://openai.com/index/sora/ [https://perma.cc/DZ7H-PMTC] (staff-uploaded archive) (last modified Oct. 12, 2025); Steven Levy, *OpenAI's Sora Turns AI Prompts into Photorealistic Videos*, WIRED (Feb. 15, 2024, at 13:15 ET), https://www.wired.com/story/openai-sora-generative-ai-video/ [https://perma.cc/728D-DM27 (staff-uploaded, dark archive)].

86. *See* Wendy Lee, *Open AI's Controversial Sora Is Finally Launching Today. Will It Truly Disrupt Hollywood?*, L.A. TIMES (Dec. 9, 2024, at 10:02 PT), https://www.latimes.com/entertainment-arts/business/story/2024-12-09/openais-controversial-text-to-video-tool-sora-is-widely-released [https://perma.cc/7XU7-ZH5A (dark archive)].

87. Cade Metz, *OpenAI Unveils AI that Instantly Generates Eye-Popping Videos*, N.Y. TIMES (Feb. 15, 2024), https://www.nytimes.com/2024/02/15/technology/openai-sora-videos.html [https://perma.cc/J2BB-F3LY (staff-uploaded, dark archive)]; Shirin Ghaffary & Rachel Metz, *OpenAI's Sora Video Generator Is Impressive, but Not Ready for Prime Time*, BLOOMBERG (Feb. 22, 2024, at 17:06 ET), https://www.bloomberg.com/news/newsletters/2024-02-22/openai-s-sora-video-generator-is-impressive-but-not-ready-for-prime-time [https://perma.cc/6E4B-28CZ (staff-uploaded, dark archive)].

88. Cade Metz & Tripp Mickle, *OpenAI Completes Deal that Values the Company at $80 Billion*, N.Y. TIMES (Feb. 16, 2024), https://www.nytimes.com/2024/02/16/technology/openai-artificial-intelligence-deal-valuation.html [https://perma.cc/4QCP-JGK3 (staff-uploaded, dark archive)].

89. Nieva et al., *supra* note 81.

90. Kelly, *supra* note 56 ("Scientists now at Google invented the diffusion computational models that are at the core of image generators today, but the company has been so concerned about what people might do with them that it still has not opened its own experimental generators, Imagen and Parti, to the public.").

91. *See* Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet & Mohammad Norouzi, *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding* 3–5 (arXiv, Working Paper No. 2205.11487, 2022), https://arxiv.org/pdf/2205.11487 [https://perma.cc/2KNL-GK5B] (explaining Imagen and its training methodology); Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet & Tim Salimans, *Imagen Video: High Definition Video Generation with Diffusion Models* 1–6 (arXiv, Working Paper No. 2210.02303, 2022), https://arxiv.org/pdf/2210.02303 [https://perma.cc/2BZ5-G7KX] (presenting Imagen and its video results); Charley F. Brown & Catherine I. Seibel, *Google Facing New Copyright Suit Over AI-Powered Image Generator*, BALLARD SPAHR (May 3, 2024) https://www.ballardspahr.com/insights/alerts-and-articles/2024/05/google-facing-new-copyright-suit-over-ai-powered-image-generator [https://perma.cc/BC7W-J7TU (staff-uploaded archive)] (detailing that Google is currently facing a copyright lawsuit stemming from Imagen's launch in 2022).

generator, Lumiere, can create videos from text as well as from images.[92] As of August 2025, Google's parent company, Alphabet, is valued at over $2 trillion.[93]

StabilityAI developed Stable Diffusion, an open-source model that allows users to generate images from text and edit existing images through text prompts.[94] The open-source release of its code allows the generative AI research community to experiment with its model in various ways.[95] The initial Stable Diffusion model was trained on over 2.3 billion pairs of images and text.[96] StabilityAI also released an open video model, Stable Video Diffusion, able to create short videos from both text and image prompts.[97] StabilityAI was reportedly valued at approximately $1 billion.[98]

In June 2024, NVIDIA topped Microsoft as the world's most valuable public company, valued at $3.34 trillion.[99] NVIDIA dominates "what analysts call the 'new gold or oil in the tech sector[,]' . . . the chips needed for artificial

92. Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel & Inbar Mosseri, *Lumiere: A Space-Time Diffusion Model for Video Generation* 1 (arXiv, Working Paper No. 2401.12945, 2024), https://arxiv.org/pdf/2401.12945 [https://perma.cc/9G8H-HFJB].

93. *Alphabet Inc. (GOOGL)*, STOCK ANALYSIS, https://stockanalysis.com/stocks/googl/market-cap/ [https://perma.cc/EU2L-2KC2].

94. Sobel, *Elements of Style*, *supra* note 78, at 52–53.

95. *See* Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Rawamswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazr, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhagn, Daniel E. Ho, Percy Liang & Arvind Naranyan, *On the Societal Impact of Open Foundation Models* 1–5 (arXiv, Working Paper No. 2403.07918, 2024), https://arxiv.org/pdf/2403.07918 [https://perma.cc/ZCD6-GQ7Q] (detailing the societal impact of releasing code in an open-source fashion); Kevin Roose, *A Coming-Out Party for Generative A.I., Silicon Valley's New Craze*, N.Y. TIMES (Oct. 21, 2022), https://www.nytimes.com/2022/10/21/technology/generative-ai.html [https://perma.cc/H9XH-WZ75 (staff-uploaded, dark archive)].

96. Damanpreet Kaur Vohra, *How To Train a Stable Diffusion Model*, HYPERSTACK https://www.hyperstack.cloud/technical-resources/tutorials/how-to-train-a-stable-diffusion-model [https://perma.cc/73NU-L886] (last updated Nov. 26, 2025); Kashmir Hill, *This Tool Could Protect Artists from A.I.-Generated Art that Steals Their Style*, N.Y. TIMES, https://www.nytimes.com/2023/02/13/technology/ai-art-generator-lensa-stable-diffusion.html [https://perma.cc/9T85-4RTS (staff-uploaded, dark archive)] (last updated Feb. 17, 2023) (noting the copyright challenges involved in training Stable Diffusion's model).

97. Blattman et al., *supra* note 65, at 1.

98. Deepa Seetharaman, *Tech Investor Sean Parks Leads Rescue of Struggling AI Startup*, WALL ST. J. (June 25, 2024, at 10:00 ET), https://www.wsj.com/tech/ai/tech-investor-sean-parker-leads-rescue-of-struggling-ai-startup-0d2e2b3b [https://perma.cc/JZM4-NSD4 (staff-uploaded, dark archive)] ("Stability raised $101 million in late 2022 in a round valuing it at $1 billion.").

99. Mitchell Labiak, *AI Frenzy Makes NVIDIA the World's Most Valuable Company*, BBC (June 19, 2024), https://www.bbc.com/news/articles/cyrr40x0z2mo [https://perma.cc/SX95-U6YH] (noting that the company's value almost doubled since the beginning of 2024); Tripp Mickle & Joe Rennison, *NVIDIA Becomes Most Valuable Public Company, Topping Microsoft*, N.Y. TIMES (June 18, 2024), https://www.nytimes.com/2024/06/18/technology/nvidia-most-valuable-company.html [https://perma.cc/556B-89G7 (staff-uploaded, dark archive)].

intelligence (AI)."[100] The company is renowned for its developments in GPUs, which are essential components in the gaming industry and the development of AI.[101] NVIDIA has developed several pre-trained video and image creation models.[102] These include the GauGan2,[103] Generative AI by Getty Images,[104] StyleGan3,[105] NeVA,[106] and more. NVIDIA also offers a platform with a wide array of products and services allowing third parties to develop and deploy customized generative AI tools.[107]

Adobe was founded in 1982.[108] In the decades since, it has been a major actor in the fields of digital creativity as well as design and document management. The integration of Firefly, offering a variety of text-to-image services, into Adobe's other services has made it easily accessible for users of these services.[109] Adobe's reported worth in July 2024 was over $233 billion.[110]

There are also numerous smaller companies offering a wide range of options to create images and videos.[111] The companies leading the market of

100. Labiak, *supra* note 99.

101. Press Release, NVIDIA Newsroom, NVIDIA Brings Generative AI to Millions, with Tensor Core GPUs, LLMs, Tools for RTX PCs and Workstations (Jan. 8, 2024) (on file with the North Carolina Law Review).

102. *AI Models*, NVIDIA DEV.: A.I., https://developer.nvidia.com/ai-models [https://perma.cc/6WMX-LESB].

103. Isha Salian, *Stroke of Genius: GauGAN Turns Doodles into Stunning, Photorealistic Landscapes*, NIVIDIA: BLOGS (Mar. 18, 2019), https://blogs.nvidia.com/blog/gaugan-photorealistic-landscapes-nvidia-research/ [https://perma.cc/NR5D-CVKZ].

104. *Generate New AI Images or Modify Our Creative Imagery*, GETTY IMAGES: AI GENERATOR, https://www.gettyimages.com/ai/generation/about [https://perma.cc/H62M-QQD6].

105. Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen & Timo Aila, *Alias-Free Generative Adversarial Network* 2 (arXiv, Working Paper No. 2106.12423, 2021), https://arxiv.org/pdf/2106.12423 [https://perma.cc/V37A-XJQQ] (explaining the "new StyleGAN3" generator as compared to StyleGAN2); StyleGAN3 Pretrained Models, NVIDIA: NGC CATALOG, https://catalog.ngc.nvidia.com/orgs/nvidia/teams/research/models/stylegan3 [https://perma.cc/Z2LA-72FS].

106. *NeVA*, NVIDIA: NEMO FRAMEWORK USER GUIDE, https://docs.nvidia.com/nemo-framework/user-guide/24.12/nemotoolkit/multimodal/mllm/neva.html [https://perma.cc/FY5B-G3WK].

107. *E.g.*, *Agentic AI Solutions: Transform Your Business with NVIDIA*, NVIDIA, https://www.nvidia.com/en-eu/ai-data-science/generative-ai/ [https://perma.cc/9N77-T2J3]; *NVIDIA NeMo*, NVIDIA, https://www.nvidia.com/en-eu/ai-data-science/products/nemo/ [https://perma.cc/PU3E-QSQF]; *Explore Visual Design Models: Try NVIDIA NIM APIs*, NVIDIA, https://build.nvidia.com/explore/visual-design [https://perma.cc/2NSL-SL8H].

108. *Fast Facts*, ADOBE, https://www.adobe.com/content/dam/cc/en/fast-facts/pdfs/fast-facts.pdf [https://perma.cc/942M-U3FM (staff-uploaded archive)].

109. *See* Shelly Putnam Tupper, *Hands On: Adobe Firefly in Photoshop*, PCMAG (Sep. 13, 2023), https://www.pcmag.com/news/hands-on-adobe-firefly-in-photoshop [https://perma.cc/LL4J-6VZR] (last updated Sep. 13, 2023).

110. *Adobe Inc. (ADBE): Adobe Market Cap & Net Worth*, STOCK ANALYSIS (Nov. 21, 2025, at 16:00 ET), https://stockanalysis.com/stocks/adbe/market-cap/ [https://perma.cc/X4ZM-NZGP].

111. Nguyen et al., *supra* note 42, at 4; *e.g.*, *Luma Dream Machine: New Freedoms of Imagination*, LUMA AI, https://lumalabs.ai/dream-machine [https://perma.cc/56DA-6FAS]; MIDJOURNEY,

generative AI models are varied—some have been around for decades, while others are still in their early years; some led the market, while the founders of the others were still children; and some offer a wide variety of services, while others focus primarily on generative AI. What they all have in common is the desire to provide users with the most advanced tools to realize their creative and artistic dreams.

C. *Deepfake Harm*

As with other innovative technologies, image- and video-generating AI tools have intriguing and promising applications,[112] but they can also be incredibly harmful. These tools are commonly used to promote political disinformation[113] and to generate sexually explicit images,[114] child sexual abuse material ("CSAM"),[115] and pornography.[116] In the wrong hands, deepfakes can be weaponized and used for malicious purposes, harming individuals,[117] distorting democratic discourse, manipulating elections, and exacerbating social

https://www.midjourney.com/home [https://perma.cc/F7VD-42AX]; *see also* Caroline Haskins, *A Deepfake Nude Generator Reveals a Chilling Look at Its Victims*, WIRED (Mar. 25, 2024, at 07:00 ET), https://www.wired.com/story/deepfake-nude-generator-chilling-look-at-its-victims/ [https://perma.cc/VT2P-EFXU (staff-uploaded, dark archive)].

112. Alphabet CEO Sundar Pichai even likened the development of new AI technology to humans' discovery of electricity or fire. Lauren Goode, *Google CEO Sundar Pichai Compares Impact of AI to Electricity and Fire*, VERGE (Jan. 19, 2018, at 16:50 ET), https://www.theverge.com/2018/1/19/16911354/google-ceo-sundar-pichai-ai-artificial-intelligence-fire-electricity-jobs-cancer [https://perma.cc/KX3X-HHMD (staff-uploaded, dark archive)] ("AI is 'one of the most important things that humanity is working on. It's more profound than, I don't know, electricity or fire.'"); Prathana Prakash, *Alphabet CEO Sundar Pichai Says that AI Could Be 'More Profound' than Both Fire and Electricity—But He's Been Saying that for Years*, FORTUNE (Apr. 17, 2023, at 13:50 ET), https://fortune.com/2023/04/17/sundar-pichai-a-i-more-profound-than-fire-electricity/ [https://perma.cc/8MM9-X3ZK (staff-uploaded, dark archive)].

113. Catherine Kim, *How Deepfakes Could Upend the 2024 Elections*, POLITICO (July 2, 2024, at 19:00 ET), https://www.politico.com/newsletters/politico-nightly/2024/07/02/how-deepfakes-could-upend-2024s-elections-00166347 [https://perma.cc/5PJZ-NA95 (staff-uploaded, dark archive)]; Sander Van Der Linden, *AI-Generated Fake News Is Coming to an Election Near You*, WIRED (Jan. 22, 2024, at 07:00 ET), https://www.wired.com/story/ai-generated-fake-news-is-coming-to-an-election-near-you/ [https://perma.cc/K2AL-9K3F (staff-uploaded, dark archive)].

114. Matt Burgess, *Deepfake Porn Is Out of Control*, WIRED (Oct. 16, 2023, at 07:00 ET), https://www.wired.com/story/deepfake-porn-is-out-of-control/ [https://perma.cc/B2JM-HM8W (staff-uploaded, dark archive)].

115. Press Release, Dep't of Just. Off. of Pub. Affs., Man Arrested for Producing, Distributing, & Possessing AI-Generated Images of Minors Engaged in Sexually Explicit Conduct (May 20, 2024), https://www.justice.gov/opa/pr/man-arrested-producing-distributing-and-possessing-ai-generated-images-minors-engaged [https://perma.cc/QPK2-XGU3].

116. Issie Lapowsky, *The Race To Prevent 'the Worst Case Scenario for Machine Learning*,' N.Y. TIMES (June 24, 2023), https://www.nytimes.com/2023/06/24/business/ai-generated-explicit-images.html [https://perma.cc/KB53-3EFQ (staff-uploaded, dark archive)]; Roose, *supra* note 95.

117. Benjamin L.W. Sobel, *A Real Account of Deep Fakes*, 124 MICH. L. REV. (forthcoming 2026) (manuscript at 22), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4829598 [https://perma.cc/N3PN-3EQ9 (staff-uploaded archive)] [hereinafter Sobel, *A Real Account of Deep Fakes*].

divisions.[118] Danielle Citron, a professor of law and the leading expert on the harms of sexual deepfakes, has been a steadfast voice in warning of the potential harms of deepfakes in sexual contexts and in intimate relationships.[119]

Research shows that while people believe they can detect deepfakes,[120] some deepfakes can be virtually indistinguishable from authentic content,[121] and people's actual ability to distinguish the two is only slightly better than a random guess.[122] Detection of deepfakes is challenging, and even AI experts and tools developed to perform such identification have limited success.[123] Raising people's awareness of the problem of deepfakes does not meaningfully improve detection rates.[124] The inability to effectively distinguish authentic content from fake content negatively impacts people's perception of reality, their ability to identify truth, and their very notion of truth.[125]

118. Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1771–85.

119. Citron, *Sexual Privacy*, *supra* note 2, at 1922–23; CITRON, THE FIGHT FOR PRIVACY, *supra* note 7, at 38.

120. Nils C. Köbis, Barbora Doležalová & Ivan Soraperra, *Fooled Twice: People Cannot Detect Deepfakes but Think They Can*, 24 ISCIENCE, Nov. 19, 2021, at 1, 4 [hereinafter Köbis et al., *Fooled Twice*] ("Results reveal that participants have exaggerated beliefs in their detection abilities when such beliefs are elicited in an unincentivized way.").

121. CITRON, THE FIGHT FOR PRIVACY, *supra* note 7, at 38 ("In January 2019, deep fakes were buggy and flickery. Nine months later, I've never seen anything like how fast they're going.") (quoting Rob Toews, *Deepfakes Are Going to Wreak Havoc on Society*, FORBES (May 25, 2020, at 23:54 ET), https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/ [https://perma.cc/93RE-R7TC]); Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza & Hafiz Malik, *Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Way Forward*, 53 APPLIED INTEL. 3974, 3975 (2023); Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano & Hao Li, *Protecting World Leaders Against Deep Fakes*, PROCS. IEEE/CVF CONF. ON COMPUT. VISION & PATTERN RECOGNITION (CVPR) WORKSHOPS 38, 38 (2019); Marissa Koopman, Andrea Macarulla Rodriguez & Zeno Geradts, *Detection of Deepfake Video Manipulation*, PROCS. 20TH IRISH MACH. VISION & IMAGE PROCESSING CONF. 133, 133 (2018); Yuezun Li, Ming-Ching Chang & Siwei Lyu, *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking* 1 (arXiv, Working Paper No. 1806.02877, 2018), https://arxiv.org/pdf/1806.02877 [https://perma.cc/GM3P-A2FX].

122. Köbis et al., *Fooled Twice*, *supra* note 120, at 6–7; Kimberly T. Mai, Sergi Bray, Toby Davies & Lewis D. Griffin, *Warning: Humans Cannot Reliably Detect Speech Deepfakes*, 18 PLOS ONE, Aug. 2, 2023, at 1, 8 (placing detection rates of audio deepfakes at approximately seventy-three percent).

123. CITRON, THE FIGHT FOR PRIVACY, *supra* note 7, at 38 ("Deepfakes are so sophisticated that even experts struggle to distinguish [them]."); Nguyen et al., *supra* note 42, at 1; Cade Metz & Tiffany Hsu, *OpenAI Releases 'Deepfake' Detector to Disinformation Researchers*, N.Y. TIMES (May 7, 2024), https://www.nytimes.com/2024/05/07/technology/openai-deepfake-detector.html [https://perma.cc/JY37-UYYH (staff-uploaded, dark archive)].

124. Köbis et al., *supra* note 120, at 10 ("[A]ppeals aiming to increase people's awareness of the problem do not suffice to improve people's detection abilities.").

125. Spencer McKay & Chris Tenove, *Disinformation as a Threat to Deliberative Democracy*, 74 POL. RSCH. Q. 703, 708 (2020) (suggesting we call this effect "epistemic cynicism"); Melissa Heikkilä, *An AI Startup Made a Hyperrealistic Deepfake of Me That's So Good It's Scary*, MIT TECH. REV. (Apr. 25, 2024), https://www.technologyreview.com/2024/04/25/1091772/new-generative-ai-avatar-deepfake-synthesia/ [https://perma.cc/6MYA-EQJT].

People are becoming increasingly skeptical about the existence of an objective truth and about the ability of technology or human experts to discover such a truth.[126] Such a reality is detrimental to "science and other evidence-based social and political institutions (such as courts)" because individuals are "skeptical about the possibility of objectively verifiable truth, or distrustful of experts and procedures that enable the discovery of such truth."[127] This can further fragment society and undermine trust in institutions, such as courts and scientists, who are based on a belief that a truth is attainable and are important for the functioning and flourishing of democratic societies.

1. Sexually Explicit Deepfakes

In late January 2024, the internet caught fire. Fake, sexually explicit images of Taylor Swift went viral.[128] One of the images, shared on X (previously Twitter), garnered forty-seven million views before the user's account was suspended.[129] X also suspended the ability to search for any Taylor Swift-related content on the platform.[130] But the images had already been shared widely on other social media platforms. As with other online viral content, attempts to remove the image from the internet's collective memory are likely to be unsuccessful.

This is far from an isolated incident. Recently, the *New York Times* noted that "[t]een girls confront an epidemic of deepfake nudes in school."[131] The scope of the phenomenon is alarming. Nearly two-thirds of women report being worried about becoming the victim of deepfake pornography.[132] These fears are justified: ninety to ninety-five percent of deepfakes are nonconsensual

126. Tomer Shadmy, *Content Traffic Regulation: A Democratic Framework for Addressing Misinformation*, 63 JURIMETRICS 1, 11 (2022); Shreeharsh Kelkar, *Post-Truth and the Search for Objectivity: Political Polarization and the Remaking of Knowledge Production*, 5 ENGAGING SCI. TECH. & SOC'Y 86, 89–90 (2019) (discussing how misinformation hinders the belief in a shared truth).

127. Shadmy, *supra* note 126, at 11.

128. Kate Conger & John Yoon, *Explicit Deepfake Images of Taylor Swift Elude Safeguards and Swamp Social Media*, N.Y. TIMES (Jan. 26, 2024), https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html [https://perma.cc/X9GZ-G9EE (staff-uploaded, dark archive)].

129. *Id.*

130. Luc Cohen, *Taylor Swift Searches Blocked on X After Fake Explicit Images Spread*, REUTERS, https://www.reuters.com/technology/taylor-swift-searches-blocked-x-after-fake-explicit-images-spread-2024-01-28/ [https://perma.cc/2NVX-TTZ8 (staff-uploaded, dark archive)] (last updated Jan 29, 2024).

131. Natasha Singer, *Teen Girls Confront an Epidemic of Deepfake Nudes in Schools*, N.Y. TIMES (Apr. 8, 2024), https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html [https://perma.cc/V7E8-4ZBT (staff-uploaded, dark archive)].

132. Press Release, ESET UK, Nearly Two-Thirds of Women Worry about Being a Victim of Deepfake Pornography, ESET UK Research Reveals (Mar. 20, 2024) https://www.eset.com/uk/about/newsroom/press-releases/nearly-two-thirds-of-women-worry-about-being-a-victim-of-deepfake-pornography-eset-uk-research-reveals/?srsltid=AfmBOorzUvkRb-XrOm_RWDicq26blCNu3eRrbqMCXPRsO2yijoLAhNir [https://perma.cc/JK7B-TD8A].

pornography; of those, ninety percent are of girls and women.[133] A 2019 survey found that fourteen percent of respondents "had experienced someone creating, distributing or threatening to distribute a digitally altered image representing them in a sexualized way."[134] From a gender perspective, sexually explicit deepfake images and videos of girls and women can be framed within the broader feminist debate about the objectification of women.[135] Professor Catharine MacKinnon, a pioneering feminist known particularly for her research on sexual harassment and gender-based violence, has argued that pornography functions as a means to objectify women, framing them as playthings meant to satisfy male viewers, perpetuating their dominance over women in society.[136] In the online context, sexually explicit deepfakes "turn real people into digital toys."[137]

A variety of harms have been noted to stem from sexual deepfakes and appropriation of another's image, including "interference with freedom and self-development;"[138] denial of agency;[139] difficulty in establishing intimate relationships;[140] a perpetual state of fear;[141] and overwhelming psychological distress, which may include anxiety, depression, and feelings of helplessness.[142] Professor Citron explains that

> [b]eing able to reveal one's naked body, gender identity, or sexual orientation at the pace and in the way of one's choosing is crucial to identity formation. When the revelation of people's sexuality or gender is out of their hands at pivotal moments, it can shatter their sense of self.[143]

133. Anastasia Powell, Adrian J. Scott, Asher Flynn & Asia A. Eaton, *Whether of Politicians, Pop Stars or Teenage Girls, Sexualised Deepfakes Are on the Rise. They Hold a Mirror to Our Sexist World*, CONVERSATION (Feb. 7, 2024, at 14:16 ET), https://theconversation.com/whether-of-politicians-pop-stars-or-teenage-girls-sexualised-deepfakes-are-on-the-rise-they-hold-a-mirror-to-our-sexist-world-222491 [https://perma.cc/UAB9-XNTD]; CITRON, THE FIGHT FOR PRIVACY, *supra* note 7, at 39 ("For intimate privacy violations, women and minors are more likely to be the victims.").

134. Powell et al., *supra* note 133.

135. Citron, *Sexual Privacy*, *supra* note 2, at 1925–26; *see* ROBIN WEST, CARING FOR JUSTICE 103 (1997) (explaining "exposure of the sexual body" as a particular type of invasion that is "gender specific").

136. Rini & Cohen, *supra* note 41, at 145; CATHARINE A. MACKINNON, FEMINISM UNMODIFIED 130 (1987) ("Pornography not only teaches the reality of male dominance. It is one way its reality is imposed as well as experienced. It is a way of seeing and using women.").

137. Rini & Cohen, *supra* note 41, at 147.

138. Solove, *supra* note 3, at 548.

139. Citron, *Sexual Privacy*, *supra* note 2, at 1924.

140. *Id.*

141. Welsh v. Martinez, 114 A.3d 1231, 1242 (Conn. App. Ct. 2015).

142. CITRON, THE FIGHT FOR PRIVACY, *supra* note 7, at 41; Citron & Franks, *supra* note 17, at 55.

143. Citron, *Sexual Privacy*, *supra* note 2, at 1884.

Even when it is clear that a particular image or video has been algorithmically generated and is widely disbelieved, the harms from this form of abuse can manifest similarly to the harms from other forms of sexual violence.[144] "Victims report[] experiencing psychological, social, physical, economic, and existential trauma."[145] Some girls refuse to leave their house for weeks after an incident; in a horrific outcome, a fourteen-year-old girl—who had been bullied by boys who were creating and sharing sexually explicit deepfakes at her school—took her life.[146] Teenagers are at a stage where they are exploring their sense of self,[147] placing them in a particularly vulnerable state, in which they can be deeply harmed by sexual deepfakes.[148]

Law enforcement, schools, platforms, parents, and girls are at a loss as to how to battle the deepfake pandemic.[149] A group of brave tenth-grade girls in Westfield, New Jersey, told school administrators that boys in their grade were using generative AI software to generate sexually explicit deepfakes of them.[150] Despite reporting that an investigation had been opened, the school did nothing for months.[151] Parents in Seattle faced similar heartbreak and anger when their daughters became the victims of sexually explicit deepfakes.[152] Initially, school officials did not even report the incident to the police, claiming they did not know what it was they were supposed to report.[153] Often, the creators of

144. *See* Rini & Cohen, *supra* note 41, at 150.

145. Powell et al., *supra* note 133.

146. Nicole Dominique, *Girl, 14, Commits Suicide After Boys Shared Fake Nude Photos of Girls and Called Her Friend Group the "Suicide Squad*," EVIE (Jan. 24, 2024), https://www.eviemagazine.com/post/girl-14-commits-suicide-boys-shared-fake-nude-photo-suicide-squad [https://perma.cc/ET6W-H4NL]. Citron reports that some victims of nonconsensual pornography contemplate suicide. Citron, *Sexual Privacy*, *supra* note 2, at 1926.

147. Peter Weinreich, *Identity Exploration in Adolescence*, INT'L J. ADOLESCENT MED. & HEALTH 51, 57 (1985) (describing the process of identity development in adolescents).

148. Citron, *Sexual Privacy*, *supra* note 2, at 1927.

149. *See* Matt Burgess, *Google Is Getting Thousands of Deepfake Porn Complaints*, WIRED (Mar. 11, 2024, at 03:00 ET), https://www.wired.com/story/google-deepfake-porn-dmca-takedowns/ [https://perma.cc/72NA-C6JP (staff-uploaded, dark archive)] (explaining why the Digital Media Copyright Act is "not fit for purpose" to combat sexually explicit deepfakes).

150. Singer, *supra* note 131.

151. *Id.*

152. *Id.*

153. *Id.*

deepfakes are believed to be the victims' classmates,[154] who are seldom identified and held accountable for the harm they created.[155]

2. Political Disinformation

Political deepfakes are used to target politicians and to undermine trust in the democratic process and institutions.[156] Deepfakes targeting politicians can be used to showcase the politician in an embarrassing light or to create the perception of the politician doing or saying something untrue. The war in Ukraine generated one such example. In the early days of the war, Ukrainian President Volodymyr Zelenskyy warned of the possibility of Russia creating a deepfake video or image of him.[157] Sure enough, shortly after this warning, a video emerged of Zelenskyy surrendering to the Russian forces.[158] Zelenskyy's warning, as well as the low quality of the video, quickly exposed its fake nature.[159] A similar scenario in the context of national elections could manipulate their outcomes.

For example, imagine Alice and Bob are two presidential candidates. The night before Election Day, a video of Alice surfaces. "Dear supporters," she may say,

154. *Id.*; Kat Tenbarge & Liz Kreutz, *A Beverly Hills Middle School Is Investigating Students Sharing AI-Made Nude Photos of Classmates*, NBC NEWS (Feb. 27, 2024, at 18:10 ET), https://www.nbcnews.com/tech/misinformation/beverly-vista-hills-middle-school-ai-images-deepfakes-rcna140775 [https://perma.cc/8C58-79GA (staff-uploaded archive)]; Jessica Grose, *AI Is Making the Sexual Exploitation of Girls Even Worse*, N.Y. TIMES (Mar. 2, 2024), https://www.nytimes.com/2024/03/02/opinion/deepfakes-teenagers.html?pgtype=Article&action=click&module=RelatedLinks [https://perma.cc/E7MV-ALPQ (staff-uploaded, dark archive)]; Caroline Haskins, *Florida Middle Schoolers Arrested for Allegedly Creating Deepfake Nudes of Classmates*, WIRED (Mar. 8, 2024, at 11:35 ET), https://www.wired.com/story/florida-teens-arrested-deepfake-nudes-classmates/ [https://perma.cc/JZ2D-YQQC (staff-uploaded, dark archive)]; Jason Koebler & Emanuel Maiberg, *A High School Deepfake Nightmare*, 404 MEDIA (Feb. 15, 2024, at 13:30 ET), https://www.404media.co/email/547fa08a-a486-4590-8bf5-1a038bc1c5a1/ [https://perma.cc/JD9W-Z345 (staff-uploaded, dark archive)].

155. Sometimes the creators of the deepfakes are in fact held accountable and forced to pay a price for their actions. *See, e.g.*, Haskins, *supra* note 154; Kat Tenbarge, *Beverly Hills Middle School Expels 5 Students After Deepfake Nude Photos Incident*, NBC NEWS (Mar. 8, 2024, at 13:55 ET), https://www.nbcnews.com/tech/tech-news/beverly-hills-school-expels-students-deepfake-nude-photos-rcna142480 [https://perma.cc/K66S-L8T6 (staff-uploaded archive)]; Singer, *supra* note 131.

156. Tiffany Hsu & Steven Lee Myers, *Can We No Longer Believe Anything We See?*, N.Y. TIMES (Apr. 8, 2023), https://www.nytimes.com/2023/04/08/business/media/ai-generated-images.html [https://perma.cc/WQT9-RE4C (staff-uploaded, dark archive)] ("Artificial intelligence allows virtually anyone to create complex artworks."); *see* Cade Metz & Tiffany Hsu, *An AI Researcher Takes on Election Deepfakes*, N.Y. TIMES (Apr. 2, 2024), https://www.nytimes.com/2024/04/02/technology/an-ai-researcher-takes-on-election-deepfakes.html [https://perma.cc/K2Q8-C8KM (staff-uploaded, dark archive)]; Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1778–79.

157. Hany Farid, *Creating, Using, Misusing, and Detecting Deep Fakes*, 1 J. ONLINE TRUST & SAFETY 1, 11 (2022) (describing the incident).

158. *Id.*

159. *Id.*

> Thank you for your trust in me. We have fought a brave battle on the campaign trail these last months. Unfortunately, it has become clear that we do not have a real chance of winning this election. To protect the integrity of elections and the unity of the American people, I have decided to withdraw my candidacy and pull out of the presidential race. For the sake of our country, let us all peacefully accept that Bob has won the election. Bob, we wish you the best of luck—your success is the country's success.

By the time the truth emerges, and Alice attempts to clarify that the video is a deepfake of her image and voice, enough damage may have been done for Bob to win the elections, undermining the process of free, democratic elections.

A deepfake does not have to be extremely sophisticated to cause a high level of disruptiveness to American confidence in their elected officials. Cheapfakes, or shallow fakes, have emerged as a low level of manipulated content.[160] Cheap or shallow fakes only slightly alter or manipulate content yet are believable and damaging.[161] In 2019, a video of then-Democratic House Speaker Nancy Pelosi looking and sounding drunk emerged and was circulated on social media.[162] Upon closer examination, it turned out that the video was a manipulation of a real interview Pelosi gave but had been significantly slowed down, making her sound unwell.[163] In June 2024, a video seemingly showed President Biden in Normandy, France, trying to sit down on a chair that was not there.[164] In a press conference, White House Press Secretary Karine Jean-Pierre called these videos "cheap fakes" and confirmed that they were in fact manipulated videos of the President.[165] In March 2024, the Democratic

160. Hany Farid, *Why the Fake Biden Videos Flooding Social Media Are More Insidious than They Appear*, MSNBC (June 18, 2024, at 16:23 ET), https://www.msnbc.com/opinion/msnbc-opinion/fake-video-joe-biden-g7-italy-rcna157767 [https://perma.cc/4SA8-GR2Q].

161. Vittoria Elliot, *Worried About Political Deepfakes? Beware the Spread of 'Cheapfakes*,' WIRED (Dec. 18, 2023, at 09:10 ET), https://www.wired.com/story/meta-youtube-ai-political-ads/ [https://perma.cc/MDN2-9K6D (staff-uploaded, dark archive)].

162. Brandi Vincent, *Bill to Combat Deepfakes Passes House Committee*, NEXTGOV (Sep. 26, 2019), https://www.nextgov.com/artificial-intelligence/2019/09/bill-combat-deepfakes-passes-house-committee/160189/ [https://perma.cc/UU2L-7ZXE (staff-uploaded archive)].

163. *Fact Check: "Drunk" Nancy Pelosi Video Is Manipulated*, REUTERS, https://www.reuters.com/article/uk-factcheck-nancypelosi-manipulated/fact-check-drunk-nancy-pelosi-video-is-manipulated-idUSKCN24Z2BI/ [https://perma.cc/6GUT-B4TW (staff-uploaded, dark archive)] (last updated Aug. 3, 2020); Hannah Denham, *Another Fake Video of Pelosi Goes Viral on Facebook,* WASH. POST (Aug. 3, 2020), https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/ [https://perma.cc/44RX-Y25K (staff-uploaded, dark archive)].

164. Melissa Goldin, *Video Edited to Make It Appear Biden Tried to Sit Down when There Wasn't a Chair*, ASSOCIATED PRESS (June 6, 2024, at 14:52 ET), https://apnews.com/article/fact-check-biden-invisible-chair-normandy-634944982428 [https://perma.cc/Q4ZJ-S4KC (staff-uploaded archive)].

165. Brett Samuels, *White House Slams 'Bad Faith' Videos that Claim to Show Biden Confused*, HILL (June 17, 2024, at 15:49 ET), https://thehill.com/homenews/administration/4726190-white-house-videos-biden-confused/ [https://perma.cc/WK2H-DY37 (staff-uploaded archive)].

National Committee made a deepfake version of a song recorded by Lara Trump, titled "Party's Fallin' Down."[166]

Just before the Democratic primaries in New Hampshire, voters were surprised to get robocalls from President Biden urging them not to participate in the primary elections, but instead to "save your vote for the November elections."[167] A man suspected of generating the deepfake audio was later charged with several counts of voter suppression.[168] Politicians may also use deepfakes to promote themselves.

In March 2024, an image of President Donald Trump surrounded by a large group of cheerful Black voters started making its way around the internet.[169] A BBC investigation found that the image had been algorithmically generated.[170] The Associated Press reported that the picture was part of the Trump campaign's effort to "win over" Black voters.[171]

False news can generate more engagement than real content because people share sources even without checking their reliability.[172] Thus, for almost a decade, amplification algorithms driving the presentation of content on social media have been disproportionately promoting deepfakes, amplifying their spread and reach.[173] By the time a deepfake is noticed, identified, analyzed, and found to be fake, it may have spread virally already, reaching hundreds of thousands of voters and potentially impacting the election process and outcome.[174]

166. Richard Lawler, *The DNC Made a Weird AI-Generated Parody of a Lara Trump Song*, VERGE (Mar. 29, 2024, at 23:45 ET), https://www.theverge.com/2024/3/29/24116156/ai-generated-dnc-lara-trump-song-parody [https://perma.cc/G884-WLFJ (staff-uploaded, dark archive)].

167. Ali Swenson & Will Weissert, *New Hampshire Investigating Fake Biden Robocall Meant to Discourage Voters Ahead of Primary*, ASSOCIATED PRESS (Jan. 22, 2024, at 23:32 ET), https://apnews.com/article/new-hampshire-primary-biden-ai-deepfake-robocall-f3469ceb6dd613079092287994663db5 [https://perma.cc/UN8C-4RN3 (staff-uploaded archive)].

168. Shannon Bond, *A Political Consultant Faces Charges and Fines for Biden Deepfake Robocalls*, NPR (May 23, 2024, at 14:58 ET), https://www.npr.org/2024/05/23/nx-s1-4977582/fcc-ai-deepfake-robocall-biden-new-hampshire-political-operative [https://perma.cc/NFV5-2G7S].

169. Matt Brown & David Klepper, *Fake Images Made to Show Trump with Black Supporters Highlight Concerns Around AI and Elections*, ASSOCIATED PRESS (Mar. 8, 2024, at 00:09 AM), https://apnews.com/article/deepfake-trump-ai-biden-tiktok-72194f59823037391b3888a1720ba7c2 [https://perma.cc/XWG2-BNBR (staff-uploaded archive)].

170. *Id.*

171. *Id.*

172. *See* Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1766 ("[O]ur natural tendency to propagate negative and novel information may enable viral circulation of deep fakes.").

173. Hagay Perlmutter, *How AI and Deepfakes Dominate Social Media and Spread Fiction*, UNIBEAM: BLOG (Oct. 23, 2024), https://unibeam.com/resource/how-ai-and-deepfake-ai-dominate-social-media-and-spread-fictional-facts/ [https://perma.cc/YTK8-M4TU (staff-uploaded archive)].

174. Hany Farid, *This Month in Generative AI: Election Season*, CONTENT AUTHENTICITY INITIATIVE: BLOG (Feb. 20, 2024), https://contentauthenticity.org/blog/february-2024-this-month-in-generative-ai-election-season [https://perma.cc/NS69-NX7L] (describing examples where

If society is not able to hold the parties generating these deepfakes accountable for the insurmountable harm they may cause to the very basic principles American society is built on, we may find ourselves in what will turn out to be fake democratic elections. Recently, several leading generative AI companies announced a commitment to fight election-related deepfakes.[175] While this is an important initiative, it is far from satisfactory for a variety of reasons. First, when such restrictions are voluntarily adopted, they can also be voluntarily abandoned. Second, the extent to which the companies will act and the steps they will take remain unclear. Finally, if such restrictions are not enforced across the board regarding the activity of all companies, the harms will continue presenting themselves. Thus, external intervention beyond what is already in existence is necessary.

II. THE STATE OF THE LAW

The emergence of new technologies challenges policymakers to provide appropriate responses that ensure safe development of the technology for the benefit of humanity while limiting its potential harms and dangers. As discussed above, the ability to generate deepfake images and videos has harmful implications for both the individuals depicted and, more broadly, for society and democracy.[176] Developers of generative AI have themselves called on regulators to act to limit the potential harms stemming from the technology.[177]

The analysis in this Section highlights the limitations of current legal mechanisms designed to contend with the harms of deepfakes. Regulation often

deepfakes impacted the outcome of elections); *see also* Simon Ellery, *Fake Photos of Pope Francis in a Puffer Jacket Go Viral, Highlighting the Power and Peril of AI*, CBS NEWS (Mar. 28, 2023, at 11:39 ET), https://www.cbsnews.com/news/pope-francis-puffer-jacket-fake-photos-deepfake-power-peril-of-ai/ [https://perma.cc/U65M-WFZP] (describing deepfakes reaching millions of views before being identified as fake).

175. Agatha Cantrill, *OpenAI, Amazon, Google Agree To Fight AI Abuse in 2024 Elections*, BLOOMBERG (Feb. 16, 2024, at 12:15 ET, https://news.bloomberglaw.com/artificial-intelligence/openai-amazon-google-agree-to-fight-ai-abuse-in-2024-elections [https://perma.cc/TH83-S8P8 (staff-uploaded, dark archive)].

176. *Supra* Section I.C.

177. Cecilia Kang, *OpenAI's Sam Altman Urges AI Regulation in Senate Hearing*, N.Y. TIMES (May 16, 2023), https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html?smid=url-share [https://perma.cc/5DP5-RGYN (staff-uploaded, dark archive)]; Johana Bhuiyan, *Tech Leaders Agree on AI Regulation but Divided on How in Washington Forum*, GUARDIAN (Sep. 13, 2023, at 20:27 ET), https://www.theguardian.com/technology/2023/sep/13/tech-leaders-washington-ai-saferty-forum-elon-musk-zuckerberg-pichai [https://perma.cc/C4AG-4W23 (staff-uploaded archive)]; Cade Metz & Gregory Schmidt, *Elon Musk and Others Call for Pause on AI, Citing 'Profound Risks to Society*,' N.Y. TIMES (Mar. 29, 2023), https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html [https://perma.cc/TA5W-ZR8Q (staff-uploaded, dark archive)]; Mary Clare Jalonick & Matt O'Brien, *Tech Industry Leaders Endorse Regulating Artificial Intelligence at Rare Summit in Washington*, ASSOCIATED PRESS (Sep. 13, 2023), https://apnews.com/article/schumer-artificial-intelligence-elon-musk-senate-efcfb1067d68ad2f595db7e92167943c [https://perma.cc/Z883-HWL3 (staff-uploaded archive)].

lags behind new developments. For example, the first American-made gas-powered car was created in 1892–93,[178] yet almost a century later, federal regulation requiring passengers to wear seatbelts was still a debated issue.[179] The first state law requiring passengers to wear them only took effect in 1984.[180] The regulatory lag is even more pronounced in the context of AI regulation. The breakneck speed at which AI has developed over the past decade, as well as its black-box nature, makes it especially challenging for regulators to keep up. Even programmers and engineers developing the models are continuously surprised by their emergent abilities.[181]

A. *Legislation & Regulation*

Legislative and regulatory efforts at the state and federal levels aim to address concerns raised by the spread of deepfake-generating technologies. These steps focus primarily on four categories: combatting and identifying deepfakes in political settings,[182] adapting CSAM laws to the age of generative

178. THE COMPLETE HISTORY OF WHEELED TRANSPORTATION: FROM CARS AND TRUCKS TO BUSES AND BIKES 39 (Erik Gregersen ed., Britannica Educ. Publ'g 2012).

179. *See* Kenneth E. Warner, *Bags, Buckles, and Belts: The Debate Over Mandatory Passive Restraints in Automobiles*, 8 J. HEALTH POL. POL'Y & L. 44, 44 (1983).

180. *State Seat Belt Law Takes Effect Today*, N.Y. TIMES (Dec. 1, 1984), https://www.nytimes.com/1984/12/01/nyregion/state-seat-belt-law-takes-effect-today.html [https://perma.cc/AX3Z-9KQN (staff-uploaded, dark archive)].

181. *See, e.g.*, Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean & William Fedus, *Emergent Abilities of Large Language Models* 1 (arXiv, Working Paper No. 2206.07682, 2022), https://arxiv.org/pdf/2206.07682 [https://perma.cc/5DDL-Z24F]; Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le & Denny Zhou, *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* 1 (arXiv, Working Paper No. 2201.11903, 2023), https://arxiv.org/pdf/2201.11903 [https://perma.cc/3F94-7Y3X]; Samuel Bowman details more uncertainties involved in LLMs, a particular type of generative AI model:

> Experts are not yet able to interpret the inner working of LLMs. . . . This means that when a lab invests in training a new LLM that advances the scale frontier, they're buying a mystery box . . . [T]hey can make few confident predictions about what those capabilities will be or what preparations they'll need to make to be able to deploy them responsibly.

Samuel R. Bowman, *Eight Things To Know about Large Language Models* 3, 6 (arXiv, Working Paper No. 2304.00612, 2023), https://arxiv.org/pdf/2304.00612 [https://perma.cc/B5US-G7T3]; *see also* Cade Metz, *Meet GPT-3. It Has Learned to Code (and Blog and Argue)*., N.Y. TIMES (Nov. 24, 2020) https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html [https://perma.cc/Z4BD-JUHC (staff-uploaded, dark archive)].

182. *See, e.g.*, Act of Sep. 29, 2022, ch. 745, 2022 Cal. Stat. 8285, 8285–58 (codified at CAL. ELEC. CODE § 20010); S.B. 6638A, 2023–2024 State Assemb., Reg. Sess. (N.Y. 2023); A.B. 7106A, 2023–2024 State Assemb., Reg. Sess. (N.Y. 2023). For a survey of recent suggested and adopted regulatory initiatives to combat deepfakes, see Thomas E. Kadri & Sonja R. West, *Deepfake Torts: Emerging Tort Frameworks in U.S. Deepfake Regulation*, 18 J. TORT L., at 4–23 (2025).

AI,[183] imposing general requirements regulating the use of another person's likeness in AI-generated content,[184] and requiring the marking of AI-generated or altered content as such.[185]

Several laws aiming to regulate aspects of deepfakes have been introduced in recent years. The most recent of these was proposed in response to the Taylor Swift deepfakes incident. In January 2024, following bipartisan collaboration, legislators introduced the Disrupt Explicit Forged Images and Non-Consensual Edits Act of 2024 (DEFIANCE Act).[186] If passed, the Act would "hold accountable those responsible for the proliferation of nonconsensual, sexually-explicit 'deepfake' images and videos."[187] The Act introduces a civil remedy for victims of deepfakes depicting them engaging in sexually explicit conduct.[188] Another law, the Identifying Outputs of Generative Adversarial Network Act (IOGAN Act)[189] aims to identify deepfakes and directs government agencies "to study and accelerate the creation of technology that can detect the disruptive content."[190] Both these laws focus on the individuals creating and spreading the deepfakes, with no liability attributed to the companies developing and deploying the models used in the process. Several other laws addressing various aspects of deepfakes have already gone into effect.[191] Some state laws criminalize

183. Riana Pfefferkorn, *Addressing Computer-Generated Child Sex Abuse Imagery: Legal Framework and Policy Implications*, *in* THE DIGIT. SOC. CONT.: A LAWFARE PAPER SERIES (2024), https://s3.documentcloud.org/documents/24403088/adressing-cg-csam-pfefferkorn-1.pdf [https://perma.cc/5QXY-JW8T].

184. *See, e.g.*, H.B. 986, 2023–2024 Gen. Assemb., Reg. Sess. (Ga. 2024).

185. *Id*. *See* Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1788, 1791–92 (noting that a deepfake ban is "not desirable" and therefore regulation must find a balance between allowing the ongoing use of the technology and limiting its harms is necessary); *see also* S.B. 8420A, 2025–2026 State Assemb., Reg. Sess. (N.Y. 2025).

186. DEFIANCE Act. S. 3696, 118th Congress (2024); *see also* Kat Tenbarge, *The Defiance Act Passes in the Senate, Potentially Allowing Deepfake Victims To Sue over Nonconsensual Images*, NBC NEWS (July 24, 2024, at 15:28 ET) https://www.nbcnews.com/tech/tech-news/defiance-act-passes-senate-allow-deepfake-victims-sue-rcna163464 [https://perma.cc/X5UD-5J4A (staff-uploaded archive)].

187. *Durbin, Graham, Klobuchar, Hawley Introduce DEFIANCE Act to Hold Accountable Those Responsible for the Proliferation of Nonconsensual, Sexually-Explicit "Deepfake" Images and Videos*, U.S. SENATE COMM. ON THE JUDICIARY (Jan. 30, 2024), https://www.judiciary.senate.gov/press/releases/durbin-graham-klobuchar-hawley-introduce-defiance-act-to-hold-accountable-those-responsible-for-the-proliferation-of-nonconsensual-sexually-explicit-deepfake-images-and-videos [https://perma.cc/Y2S8-BYJW].

188. *Id.*

189. Pub. L. No. 116-258, 134 Stat. 1150 (2020) (codified at 15 U.S.C. §§ 9201–04).

190. Vincent, *supra* note 162.

191. *E.g.*, S.B. 175, 108th Leg., Reg. Sess. (La. 2023); Act of Oct. 1, 2022, ch. 212, 2022 Fla. Laws 1936, 1937–44 (codified as amended at FLA. STAT. § 775.0847 (2022)); Fabricated Intimate or Sexually Explicit Images Act, ch. 88, 2024 Wash. Sess. Laws 474, 475 (codified in scattered sections of WASH. REV. CODE §§ 9.68A, 9A.86, 7.110 (2024)); Act of Mar. 18, 2019, ch. 490, 2019 Va. Acts 868, 868 (codified at VA. CODE § 18.2-386.2 (2019)); *see also* Kate Cox, *Deepfake Revenge Porn Now a Crime in Virginia*, ARS TECHNICA (July 1, 2019, at 15:15 ET), https://arstechnica.com/tech-policy/2019/07/deepfake-revenge-porn-distribution-now-a-crime-in-virginia/ [https://perma.cc/626J-LJ6T (staff-uploaded archive)]; Sobel, *Elements of Style*, *supra* note 78, at 6–27.

sexually explicit deepfakes, in particular those depicting minors.[192] The FBI has noted that algorithmic generation of CSAM is illegal,[193] and several states have legislated explicit prohibitions regarding the creation, possession, and dissemination of algorithmically generated CSAM as a distinct category.[194] Several people have already been arrested for using generative AI tools to create CSAM.[195]

In the political context, recognizing the severity of the threat of deepfakes on the political process, several legislatures have passed laws addressing the use of deepfakes during elections.[196] Congress has authorized the Department of Homeland Security to track and report on the state of digital content forgery technology.[197] The Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2023 ("DEEP FAKES Accountability Act")[198] debuted in September 2023; if it had passed, the Act would have required companies developing generative technologies to include provenance techniques, allowing outputs to be identified as having been algorithmically generated.[199] The Protecting Consumers from Deceptive AI Act was introduced in March 2024 and would require the National Institute of Standards and Technology to develop standards pertaining to the identification of algorithmically generated or modified content.[200] Companies developing and

192. *State Laws Criminalizing AI-Generated or Computer-Edited CSAM*, ENOUGH ABUSE, https://enoughabuse.org/get-vocal/laws-by-state/state-laws-criminalizing-ai-generated-or-computer-edited-child-sexual-abuse-material-csam/ [https://perma.cc/P6S7-Y97G] (last modified Aug. 8, 2025).

193. FED. BUREAU OF INVESTIGATION, ALERT NO. I-032924-PSA, CHILD SEXUAL ABUSE MATERIAL CREATED BY GENERATIVE AI AND SIMILAR ONLINE TOOLS IS ILLEGAL (2024), https://www.ic3.gov/PSA/2024/PSA240329 [https://perma.cc/KD3F-BX7X].

194. *E.g.*, Act of Feb. 12, 2024, ch. 87, § 2, 2024 S.D. Sess. Laws 145, 148 (codified as amended at S.D. CODIFIED LAWS § 22-24A).

195. FED. BUREAU OF INVESTIGATION, *supra* note 193; Press Release, Dep't of Just., Off. of Pub. Affs., Man Arrested for Producing, Distributing, and Possessing AI-Generated Images of Minors Engaged in Sexually Explicit Conduct (May 20, 2024), https://www.justice.gov/opa/pr/man-arrested-producing-distributing-and-possessing-ai-generated-images-minors-engaged [https://perma.cc/68AM-2CXS].

196. Act of Sep. 1, 2023, ch. 355, 2023 Tex. Gen Laws 780, 780–81 (codified at TEX. PENAL CODE § 21.165 (2023)) (criminalizing the fabrication of a deceptive video with intent to injure a candidate or influence the outcome of an election); Act of Mar. 27, 2024, Pub. L. No. 81, 2024 Ind. Acts 1180, 1180–83 (codified at IND. CODE § 3-9-8 (2024)) (requiring the disclosure of algorithmically generated election campaign communications); Act of Mar. 27, 2024, ch. 62, 2024 Or. Laws 2131, 2131 (codified as amended at OR. REV. STAT. Ch. 62, §§ 1, 4 (2024)) (requiring the disclosure of the use of synthetic media in elections campaign communications).

197. *See* William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, Pub. L. 116-283, 134 Stat. 4775, 4774–76 (2021).

198. H.R. 5586, 118th Cong. (2023).

199. *Id. See generally* John Collomosse & Andy Parsons, *To Authenticity and Beyond! Building Safe and Fair Generative AI Upon the Three Pillars of Provenance*, 44 IEEE COMPUT. GRAPHICS & APPLICATIONS 82 (2024) (explaining how provenance techniques teach about the source or history of the outputs of generative AI).

200. Protecting Consumers from Deceptive AI Act, H.R. 7766, 118th Cong. (2024).

deploying generative AI models have largely avoided being targeted and faced with liability for the harms generated using their models.

Adopting a complimentary approach, the Federal Trade Commission ("FTC") recently issued a rule prohibiting government and business impersonation schemes.[201] The commission proposed expanding the rule so that it would provide civil redress not only when fraudsters impersonate government or business entities, but also when they impersonate individuals.[202] The rule would not only expand the group of potential plaintiffs, but also potential defendants. It would allow for any actor that committed an impersonating scam to be held liable for ensuing harms. This would include the companies developing and deploying generative AI models used to generate deepfakes.[203]

While regulating innovative technology is not an easy task, current regulatory trends, as well as public discourse, leave little room to doubt that the use of generative AI in the alteration and generation of deepfakes will be subject to regulation. The leading developers of generative AI are also calling upon regulators to take action.[204] Sam Altman, the CEO of OpenAI, has specifically noted that "societal misalignments" could have dangerous consequences for generative AI and its users.[205] He noted that "these systems [are] out in society and through no particular ill intention, things just go horribly wrong."[206]

While some of these efforts represent positive possible developments, they all share one key limitation: by the time they arrive, it might simply be too late for far too many. Regulation, by its nature, is slow to develop and adjust, and requires new frameworks as new technologies enter the market. But deepfake harms are too great for us to simply wait for new regulatory frameworks to be created and adopted. The lives of young girls are being ruined today, often by their classmates who are using simple and easily accessible technology to humiliate, degrade, and shame their peers. In the political context, recent changes to content moderation policies on social media platforms are likely to

201. Statement of Chair Lina M. Khan, Joined by Comm'r Rebecca Kelly Slaughter and the Comm'r Alvaro M. Bedoya, Regarding the Final Rule on the Trade Regulation Rule on Impersonation of Government and Businesses, Commission File No. R207000, at 1 (Feb. 15, 2024), https://www.ftc.gov/system/files/ftc_gov/pdf/r207000impersonationrulelmkstmt.pdf [https://perma.cc/L895-H49M].

202. *Id.*

203. *Id.* at 2.

204. *See* Ryan Tracy, *ChatGPT's CEO Sam Altman Warns that AI 'Could Go Quite Wrong,'* WALL ST. J. (May 16, 2023, at 13:12 ET), https://www.wsj.com/articles/chatgpts-sam-altman-faces-senate-panel-examining-artificial-intelligence-4bb6942a [https://perma.cc/2WMH-7Y3F (staff-uploaded, dark archive)].

205. Ivana Saric, *"Societal Misalignments" Could Pose AI Dangers, OpenAI CEO Says*, AXIOS (Feb. 13, 2024), https://www.axios.com/2024/02/13/sam-altman-ai-danger-world-governments-summit [https://perma.cc/D6KN-Q5UT (staff-uploaded, dark archive)].

206. *Id.*

result in more fake content appearing in users' feeds.[207] We urgently need tools to ensure that girls and women are protected *now*, and that deepfakes created by political opponents and foreign interference are not those deciding the future of our democracies. The imposition of liability through the court system using established legal doctrine is the only available means to address the problem, but the currently available legal theories are insufficient, as the next two sections will show.

B. *Platform Liability*

Scholars have explored the possibility of keeping deepfakes in check through platform accountability.[208] Under this proposal, social media platforms where deepfakes are posted and shared would be liable for subsequent harms. This could change the deepfake landscape and provide strong protection for victims of sexual deepfakes.

Professors Chesney and Citron, however, note that although holding platforms liable for failing to police deepfake content may be desirable as a matter of policy, it is impossible as a matter of law.[209] Section 230 of the Communications Decency Act provides platforms immunity from liability for hosting content created by third parties.[210] Of course, Section 230 does not mandate that platforms freely host sexually explicit deepfakes, and platforms'

207. Mike Isaac & Theodore Schleifer, *Meta To End Fact-Checking Program in Shift Ahead of Trump Term*, N.Y. TIMES (Jan. 7, 2025), https://www.nytimes.com/2025/01/07/technology/meta-fact-checking-facebook.html?smid=url-share [https://perma.cc/M6J7-H9G7 (staff-uploaded, dark archive)].

208. Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1795 ("Given the key role that content platforms play in enabling the distribution of deep fakes, and the fact that creators of harmful deep fakes in some cases may be difficult to find and deter, the most efficient and effective way to mitigate harm may be to impose liability on platforms."). On the challenge of attributing liability to platforms for deepfakes, see Nicholas O'Donnell, Note, *Have We No Decency? Section 230 and the Liability of Social Media Companies for Deepfake Video*s, 2021 U. ILL. L. REV. 701, 720–31 (2021) (discussing the challenge of attributing liability to platforms for deepfakes); Danielle Keats Citron, *How To Fix Section 230*, 103 B.U. L. REV. 713, 717 (2023) (noting that Section 230 precludes liability for platforms).

209. Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1797–98.

210. Telecommunications Act of 1996, Pub. L. No. 104-104, § 509, 110 Stat. 56, 137–39 (codified as amended at 47 U.S.C. § 230(c)(1)). In recent years, there is a growing sentiment that Section 230 needs to undergo reform. *See, e.g.*, Danielle K. Citron & Benjamin Wittes, *The Problem Isn't Just Backpage: Revising Section 230 Immunity*, 2 GEO. L. TECH. REV. 453, 465 (2018) ("An overbroad reading of the [Communications Decency Act] has given platforms a free pass to ignore destructive activities, to deliberately repost illegal material, and to solicit unlawful activities while ensuring that abusers cannot be identified."); Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61, 118 (2009) ("[B]road immunity for operators of abusive websites would eliminate incentives for better behavior by those in the best position to minimize harm."); Nicholas Conlon, *Freedom to Filter Versus User Control: Limiting the Scope of § 230(c)(2) Immunity*, 105 ILL. J.L. TECH. & POL'Y 105, 105 (2014).

content moderation policies sometimes restrict posting and sharing of such content.[211] Evidently, this is nowhere near a solution to the crisis.

C. *User Liability*

Another alternative discussed by scholars and policymakers is holding users who create or share offensive content liable under tort-related doctrines.[212] Tort law offers several doctrinal avenues that may be relevant in this context. First, users can be sued for defamation when the content they create damages the reputation of another.[213] Many deepfakes easily fall into this category, even when the victims' depiction is not completely believable. State tort doctrine often allows private individuals a cause of action even if the publication was merely negligent, rather than intentional or malicious.[214] When plaintiffs are public figures, they face higher burdens and must show that the publication was motivated by malice, knowledge, or reckless disregard.[215]

Privacy torts also provide avenues for relief.[216] For example, the false-light tort, involving the reckless creation of a harmful and misleading depiction of another in public, is highly relevant for deepfakes.[217] The intentional infliction of emotional distress tort can be used when the plaintiff can demonstrate "extreme and outrageous conduct" by the defendant.[218] Creating and/or disseminating deepfake sex videos, for example, may fall under this category.

Although these doctrinal frameworks may initially seem helpful, further inspection reveals them to be an ineffective legal response to the deepfake crisis. Pursuing individual defendants is an exercise in futility, and "[c]ivil liability cannot ameliorate harms caused by deepfakes if plaintiffs cannot tie them to

211. *See, e.g.*, Emma Higham, *How We're Addressing Explicit Fake Content in Search*, GOOGLE: KEYWORD (July 31, 2024), https://blog.google/products/search/google-search-explicit-deep-fake-content-update/?mc_cid=7064fdd8bd&mc_eid=4727fac622 [https://perma.cc/46QG-34DX]; Monica Bickert, *Our Approach to Labeling AI-Generated Content and Manipulated Media*, META: NEWSROOM (Apr. 5, 2024), https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/ [https://perma.cc/MZ6M-7X2G (staff-uploaded archive)]; Andrew Hutchinson, *LinkedIn Adds Labels for AI Generated Content*, SOC. MEDIA TODAY (May 20, 2024), https://www.socialmediatoday.com/news/linkedin-labels-ai-generated-content/716674/ [https://perma.cc/RV4R-7BT6]; Ian Sherr, *Apple Intelligence Will Label AI-Generated Images in Metadata*, CNET (June 19, 2024, at 11:52 PT), https://www.cnet.com/tech/services-and-software/apple-intelligence-will-label-ai-generated-images-in-metadata/ [https://perma.cc/C2MZ-9XJH].

212. *See, e.g.*, Citron, *Sexual Privacy*, *supra* note 2, at 1933–35 (suggesting that the torts of intrusion, disclosure, false light, defamation, appropriation, and intentional infliction of emotional distress "could provide redress for sexual-privacy invasions"); Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1793–95 (discussing the possibility of suing creators of deepfakes under tort law).

213. Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1793.

214. *Id.*

215. *Id.* at 1793–94.

216. *Id.* at 1794.

217. *Id.*

218. *Id.*

their creators."[219] Pursuing individual defendants seems a highly ineffective legal solution to the deepfake problem. The number of potential infringers is vast, and most of them are unreachable (if they are foreign bots operating to destabilize Western democracies), judgment proof (if they are teenagers), and nearly impossible to identify, protected by the internet's anonymity.[220] Thus, user liability, much like platform liability, is not a promising legal answer to deepfakes.

* * *

Existing analysis of the legal response to deepfakes—focusing on regulation, platform liability, and user liability—leads to a dead end. Regulation might eventually help victims, but in the meantime, too many are suffering.[221] As the adage goes, "in the long run we are all dead."[222] Platform liability is blocked by Section 230,[223] and user liability is a nonstarter.[224] We suggest that this familiar picture misses the most obvious answer: manufacturer liability. Existing doctrine, perhaps with some minor adjustments by the courts, can be used to hold the creators of generative AI tools accountable for deepfake harms.

III. THE PROPOSAL: MANUFACTURER LIABILITY

In this Part, we propose a different approach to the governance of deepfakes through common-law manufacturer liability by considering the AI software as the product. Our proposal is unique in two critical respects. First, unlike most contemporary scholars who put their faith in the regulatory state and its machinery, we focus on private law—the law of torts. Second, by contrast to the orthodoxy of commentators who debate whether to impose liability on users or platforms, we concentrate on the companies developing and deploying the AI tools used to produce deepfakes.

Regulators have proven to be unable to keep up with the rapid pace at which AI technology has been developing over recent years. By the time regulators take action, if they ever do, their response will be obsolete. In the meantime, AI reshapes our world and our expectations. Therefore, private law claims mark a much more promising path for dealing with the dangers of AI in the immediate future.

The rapid advancement of AI tools renders *producer* liability a more relevant response than *user* or *platform* liability. These traditional frameworks may have seemed like the only relevant frameworks back in 2019, when Chesney

219. *Id.* at 1792.
220. *Id.* at 1792–93.
221. *See supra* Section II.A.
222. JOHN MAYNARD KEYNES, A TRACT ON MONETARY REFORM 80 (1923).
223. *See supra* Section II.B.
224. *See supra* Section II.C.

and Citron proposed their analysis. Today, when new AI technology has driven an unprecedented deepfake crisis, AI manufacturer liability emerges as the most natural response.

Until recently, a certain level of skill and expertise was needed to generate deepfakes, and the outcomes were not always seamless or believable.[225] The technology was not readily accessible to teenagers and children. Today, all that it takes to create a high-quality deepfake is a computer with an internet connection[226]—scholars warn that "[l]ess and less effort is required to produce a stunningly convincing tempered footage. . . . [A]lmost anyone can create deepfakes these days using existing deepfake tools."[227] The ease with which deepfakes can be created is inversely correlated with the extent of the harm that they can cause.

We suggest that manufacturer liability is the natural legal response to this type of development. The reason is simple: The current deepfake crisis is driven by the availability of powerful new AI tools. The manufacturers of these tools reap enormous profits from their distribution while remaining indifferent, though not oblivious, to the harm their products cause. In fact, they enjoy the lack of regulation in the AI space by incorporating no safety measures into their technologies. This state of affairs should not continue. Producers of harmful products are held liable for the harms their products cause in a myriad of other contexts. If a gun manufacturer were to produce a handgun with no safety, designed to be marketed to children,[228] any court would consider the liability of a manufacturer based on a design defect.[229] To take a more benign example, automobile companies, too, are liable for design defects and mechanical failures that cause harm to consumers—even if the harm is purely economic. The same

225. Ice, *supra* note 20, at 425–26 (describing the 2019 technology used to generate deepfakes as "too cumbersome for the average computer user" and stating that "the process requires above average computer literacy, including understanding torrenting, path configuration, file structures, and application versioning").

226. Heikkilä, *supra* note 125 ("Until now, all AI-generated videos of people have tended to have some stiffness, glitchiness, or other unnatural elements that make them pretty easy to differentiate from reality. Because they're so close to the real thing but *not quite it*, these videos can make people feel annoyed or uneasy or icky—a phenomenon commonly known as the uncanny valley.").

227. Nguyen et al., *supra* note 42, at 2.

228. For a discussion of the robust market of AI tools marketed directly to children, see, for example, A.A.A – AllAboutAI, *Best AI Friends for Children: Complete AI Companions 2025 Guide & Future Trends [Market Report]*, SMART AI DAILY NEWS (Mar. 18, 2025), https://smartaidaily.com/ai-hardware/best-ai-friends-for-children-complete-ai-companions-2025-guide-future-trends-market-report/ [https://perma.cc/Y787-5EV4].

229. Patrick Luff, *Regulating Firearms Through Litigation*, 46 CONN. L. REV. 1581, 1590 (2014). For an example of such an attempt, see RICHARD C. MILLER, 4 LITIGATING TORT CASES § 51:40, Westlaw LITGTORT (updated Oct. 2024) (describing a lawsuit against a major handgun manufacturer for design defect and failure to warn); John C.P. Goldberg & Benjamin C. Zipursky, *The Easy Case for Products Liability Law: A Response to Professors Polinsky and Shavell*, 123 HARV. L. REV. 1919, 1924 (2010) (explaining the concept of a design defect).

ought to be true for a company that develops an AI tool so dangerous that it could instantly ruin the lives of women and young girls. Indeed, it is an anomaly that no form of liability has been imposed on them so far.

In this Part, we develop the main proposal set forth in this paper and describe a blueprint for assigning liability to producers of AI technologies by declaring the software has been defectively designed and the duty of care ignored. We first describe the basic premise of products liability law, including the elements of a claim in the field of tort law. We then move on to apply the basic elements of the claim—a defect and a duty of care towards the plaintiff—to the case of deepfakes. We end by discussing the remedies that may be appropriate in this context.

A. *Products Liability Law*

The concept of products liability usually refers to a common-law cause of action developed by American courts in the 1960s and 1970s.[230] The fact that it is controlled by common law is important—it is developed by the courts and thus can be further finetuned to contend with developing technological threats,[231] such as algorithmically generated deepfakes. This avenue requires no regulatory change, but merely the application of existing legal tools to new fact patterns, and the necessary and natural development of these categories in light of changing technological and social conditions.[232]

A products liability claim allows recovery against a commercial seller for "sending into the stream of commerce a product containing a dangerous defect irrespective, at least to some degree, of how the defect arose."[233] The core requirement of such a cause of action is the existence of a *defect*—that the product is "sub-standard in one way or another."[234]

230. Goldberg & Zipursky, *supra* note 229, at 1923.

231. Yotam Kaplan, Adi Libson & Gideon Parchomovsky, *The Renaissance of Private Law*, 119 NW. U. L. REV. 1427, 1446 (2025).

232. *Id.*

233. Goldberg & Zipursky, *supra* note 229, at 1924.

234. *Id.*; *see* Greenman v. Yuba Power Prods., Inc., 377 P.2d 897, 901 (Cal. 1963) (explaining that products liability pertains only to products containing a "defect in design and manufacture"); RESTATEMENT (SECOND) OF TORTS § 402A(1) (A.L.I. 1965) (noting that product liability applies to products "in a defective condition unreasonably dangerous to the user or consumer"); Order at 36, Garcia v. Character Techs. Inc., No. 6:24-cv-1903-ACC-DCI (M.D. Fla. July 15, 2025) ("Character A.I. is a product for the purposes of Plaintiff's product liability claims so far as Plaintiff's claims arise from defects in the Character A.I. app . . . .").

Products liability law recognizes three core categories of defects: design defects,[235] manufacturing defects,[236] and failures to warn.[237] The concept of a design defect is the most interesting for our discussion as we argue that the product was faulty to begin with: Nothing went wrong in the manufacturing process, and the issue is not of a failure to warn consumers.[238] The product is simply dangerous, and should not have been marketed as it was without some added safety measures.[239] The concept of a design defect is usually operationalized through a risk-utility test: a product is considered defective if it lacks cost-effective safety measures.[240]

In addition to the defect requirement, the *duty of care* is another core concept of products liability law. Thus, recovery is typically limited to a specific class of plaintiffs, to whom the manufacturer owes a duty of care.[241] This concept is closely tied to the idea of warranty, meaning that the duty of care originates from an explicit or implicit obligation between a seller and a buyer of a product.[242] Courts also offer recovery to bystanders, recognizing a duty of care applies toward all foreseeable victims, not just consumers who bought the product from the defendant directly.[243]

B. *Design Defect*

This Section applies the familiar concept of design defect to the software driving generative AI tools. While the EU has clarified that software is considered a product and therefore subject to products liability law,[244] the legal status of software in the United States is less clear.[245] The Restatement (Third)

235. *See generally* Sheila L. Birnbaum, *Unmasking the Test for Design Defect: From Negligence [to Warranty] to Strict Liability to Negligence*, 33 VAND. L. REV. 593 (1980) (detailing the requirement for a design defect).

236. *See* Goldberg & Zipursky, *supra* note 229, at 1944–45 (describing circumstances where a manufacturer would be held liable for a defective product).

237. *See generally* Beshada v. Johns-Mansville Prods. Corp., 447 A.2d 539 (N.J. 1982) (finding liability for failure to warn as a products liability defect).

238. Birnbaum, *supra* note 235, at 603.

239. *Id.*

240. Soule v. Gen. Motors Corp., 882 P.2d 298, 308 (Cal. 1994).

241. RESTATEMENT (SECOND) OF TORTS § 402A cmt. o (A.L.I. 1965).

242. Commonwealth v. Johnson Insulation, 682 N.E.2d 1323, 1326 (Mass. 1997); RESTATEMENT (SECOND) OF TORTS § 402A cmt. m (A.L.I. 1965); *see* MARK A. GEISTFELD, PRINCIPLES OF PRODUCTS LIABILITY 10–21 (3d ed. 2020) (detailing the evolution of strict Products liability from the implied warranty).

243. DAVID G. OWEN & MARY J. DAVIS, OWEN & DAVIS ON PRODUCTS LIABILITY § 5:5, at 393–95 (4th ed. 2014) (describing liability to "foreseeable bystanders").

244. Council Directive 2024/2853, art. 4, 2024 O.J. (L 2853) 12 (EC).

245. Gordon-Tapiero, *supra* note 18, at 56–58 (analyzing the legal classification of software in the United States); *see* Asaf Lubin, *On Software Bugs and Legal Bugs: Product Liability in the Age of Code*, 100 IND. L.J. 1891, 1891 (2025) (exploring the unsettled classification of software as a product in U.S. courts); *id.* at 1900 ("[T]he Third Restatement hasn't rejected the possibility of treating software as a product.").

of Torts: Products Liability, defines a product as "tangible personal property distributed commercially for use or consumption."[246] Deepfake technology does not fall neatly into this definition. At the same time, in recent cases, software *was* classified as a product and *was* therefore subject to products liability.[247] Most recently, in *Garcia v. Character Technologies, Inc.*,[248] the court ruled that the software driving AI companions was a product for the purpose of a product liability claim.[249]

We argue that software that allows any individual to harm others by creating sexual deepfakes of an identifiable individual while providing no tools to track the history of the deepfake and identify its creator should be considered as having been defectively designed.

In this Section, we detail the characteristics of what we believe a nondefective design of software should look like. The analysis of what a nondefective design is must be guided by the duty of care that companies developing deepfake technology have towards users and foreseeable victims. These standards must strike a delicate balance between limiting the harm stemming from the wide spread of these models and allowing for their future use in positive and constructive contexts.[250] These minimal standards will take time for regulators to adopt. Yet, they can already serve as a basis for litigating general residual common-law claims such as products liability.[251] A company that fulfills these standards will be effectively shielded from liability for harms created by its model's output. Companies that do not adhere to these standards may find themselves liable for the harmful outputs of their models that they failed to account for and protect against.[252]

246. RESTATEMENT (THIRD) OF TORTS § 19(a) (A.L.I. 1998).

247. Order Denying Defendant Lyft's Motion for Partial Summary Judgment as to Plaintiff's Product Liabiltiy [sic] Claims (Counts VI–VIII) at 3, Brookes v. Lyft Inc., No. 50-2019-CA-004782, 2022 WL 19799628, at *2 (Fla. Cir. Ct. Sep. 30, 2022) (concluding the Lyft App is a product for purposes of Florida product liability law); Report and Recommendation on Motion to Dismiss, T.V. v. Grindr, LLC, No. 3:22-cv-864-MMH-PDB, 2024 WL 4128796, at *26 (M.D. Fla. Aug. 13, 2024) (recommending the rejection of the defendant's argument that the Grindr app is not a product).

248. No. 6:24-cv-1903-ACC-DCI (M.D. Fla. July 15, 2025).

249. Order, *supra* note 234, at 36.

250. Chesney & Citron, *Deep Fakes*, *supra* note 9 ("Crafting a law prohibiting destructive applications of deep-fake technology while excluding beneficial ones would be difficult, but perhaps not impossible.").

251. Isabel Gottlieb, *AI Firms Face Broad Liability Under FTC Deepfake Proposal (1)*, BLOOMBERG L. https://news.bloomberglaw.com/artificial-intelligence/ai-providers-face-broad-liability-under-ftcs-deepfakes-proposal [https://perma.cc/556F-WF8Y (staff-uploaded, dark archive)] (Feb. 20, 2024, at 22:15 ET) ("A more targeted approach could offer a safe harbor for companies that complied with certain guidelines—like watermarking content, promptly taking down harmful material, or conducting know-your-customer checks.").

252. Leaving decisions regarding AI regulatory frameworks to be developed by courts on a case-by-case basis is not without challenges. *See* Alicia Solow-Niederman, *Do Cases Generate Bad AI Law?* 25 COLUM. SCI. & TECH. L. REV. 261, 266 (2024) ("Generative AI litigation and adjudication may

Researchers have identified several high-level principles that might be translated into safety requirements in AI tools. These include notions such as transparency,[253] explainability,[254] traceability,[255] accountability,[256] bias mitigation, and privacy and data protection. With these principles in mind, we suggest that a safe generative AI application must be one that cannot be used to anonymously, immediately, and untraceably create deepfakes of real people in sexually explicit contexts. An AI tool capable of doing so simply does not include the necessary basic safety requirements for such a dangerous tool and is therefore a defective product. Practically, this requires companies developing deepfake technology to ensure that when a user creates a sexual deepfake, the software includes tools that enable tracing the history of the deepfake as well as identifying its creator. Similarly, technology that enables the creation of sexual deepfakes must include safeguards that limit the ability to use the likeness of an identifiable individual. This can be done by allowing users to create sexual

produce low-quality decisions, as well as fail to address and redress collective harms. In addition to generally applicable concerns about unrepresentative facts producing bad legal rules and inequities in litigation, the facts presented in emerging generative AI lawsuits may limit available causes of action under entrenched precedents."); Gordon-Tapiero, *supra* note 18, at 63 ("[M]aking decisions on a case-by-case basis . . . risks missing broader or systemic challenges stemming from AI companions.").

253. *See generally* Gianclaudio Malgieri & Frank Pasquale, *From Transparency to Justification: Toward Ex Ante Accountability for AI* (Brussels Priv. Hub Working Paper No. 33, 2022), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4099657 [https://perma.cc/3QY6-N9HT] (click "Download This Paper" or "Open PDF in Browser") (proposing a model where developers of AI have the burden of proof in showing that their technology is not discriminatory, unfair, or inaccurate); John Zerilli, Alistair Knott, James Maclaurin & Colin Gavaghan, *Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?*, 32 PHIL. & TECH. 661 (2019) (arguing AI decision-making may be being held to unrealistically high standards when it comes to transparency); Axel Walz & Kay Firth-Butterfield, *Implementing Ethics into Artificial Intelligence: A Contribution, from a Legal Perspective, to the Development of an AI Governance Regime*, 18 DUKE L. & TECH. REV. 176, 196 (2019); John Zerilli, Umung Bhatt & Adrian Weller, *How Transparency Modulates Trust in Artificial Intelligence*, 3 PATTERNS 1, Apr. 8, 2022 (detailing that a high degree of transparency encourages trust in the AI system).

254. *See* Lobel, *The Law of AI for Good*, *supra* note 47, at 1133–34 ("The right to explainable AI is part of EU and American regulatory reforms. But behavioral research indicates that whether explanations improve human decision-making in relation to AI is situationally dependent."). *See generally* Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COLUM. L. REV. 1829 (2019) (arguing that "judges should demand explanations" for how "black box" machine learning algorithms "reach particular decisions, recommendations, or predictions"); Daniel Ben David, Yehezkel S. Resheff & Talia Tron, *Explainable AI and Adoption of Financial Algorithmic Advisors: An Experimental Study* (arXiv, Working Paper No. 2101.02555, 2021), https://arxiv.org/pdf/2101.02555 [https://perma.cc/CZ5Z-TWVA] (studying if there is an effect on someone's readiness to adopt, willingness to pay, and ability to trust an AI financial consultant).

255. *See* Walz & Firth-Butterfield, *supra* note 253, at 196 ("How can it be guaranteed that the considerations taken into account by an AI system can be traced back for the purpose of allocation of liability?").

256. *See id.* at 187–88 (recognizing the principle of accountability as important in the context of AI); Maranke Wieringa, *What To Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability*, 2020 CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 1, 5–6.

deepfakes based only on text prompts, and not, for example, by creating an interface that allows the user to upload an existing image or video which will serve as the basis for the deepfake output.

We argue that when a generative AI company designs its tools in accordance with the principles detailed above, its AI products should not be considered defective, and liability should not ensue. If—despite implementing safeguards as suggested above—a deepfake still harms a victim, the company should not be liable. A company that did not reasonably implement appropriate safeguards, however, will be viewed as having manufactured and distributed a dangerous and defective product and thus may be held liable for the outcomes of doing so.

1. Traceability

One of the principles often discussed in the context of AI governance is the principle of traceability.[257] One of the features that makes generative AI attractive for those who misuse it is the anonymity that it provides. Currently, tracing a specific deepfake back to its creator and holding the creator accountable for the resulting harm is extremely difficult. The lack of traceability is one of the major obstacles to any significant effort to reduce the harms of deepfakes.[258]

Identifying those who *spread* an image or a video may be less complicated, as this involves a social interaction, whether it was shared on social media, in group chats, or in the school's lunchroom.[259] It is much easier to hide under the cloak of anonymity in *creating* the deepfake.[260] Regulation that mandates the

257. David Oniani, Jordan Hilsman, Yifan Peng, Ronald K. Poropatich, Jeremy C. Pamplin, Gary L. Legault & Yanshan Wang, *Adopting and Expanding Ethical Principles for Generative Artificial Intelligence from Military to Healthcare*, 6 NATURE PARTNER J. DIG. MED. 225, 225 (2023); Xukang Wang & Ying Cheng Wu, *Balancing Innovation and Regulation in the Age of Generative Artificial Intelligence*, 14 J. INFO. POL'Y 385, 398 (2024); Qinghua Lu, Liming Zhu, Xiwei Xu, Zhenchang Xing & Jon Whittle, *Toward Responsible AI in the Era of Generative AI: A Reference Architecture for Designing Foundation Model-Based Systems*, 41 IEEE SOFTWARE 91, 91 (2024); Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma & Francisco Herrera, *Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation*, 99 INFO. FUSION 1, 11 (2023); 2024 O.J. (L 1689) 27, 53, 71.

258. Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1792.

259. Singer, *supra* note 131.

260. Sometimes circumstantial evidence may make it easier to identify the creator of a deepfake, such as in the case of a boy who sent a classmate a friend request for her private account on Instagram. The student subsequently copied pictures of her and used them as the basis to generate sexually explicit images of her and share them in a Snapchat group. The boy was subsequently sued for his actions. Singer, *supra* note 131. *See* Issue Primer, *Synthetic Media & Deepfakes*, CTR. NEWS TECH. & INNOVATION (July 22, 2024), https://innovating.news/article/synthetic-media-deepfakes/ [https://perma.cc/2MNJ-9H86] (last update Oct. 11, 2024); Amitkumar Shrivastava, *Deepfake Phenomenon, Impact and Challenges*, MEDIUM (Jan. 24, 2024), https://medium.com/@amit.ai.mldl/deepfake-phenomenon-impact-and-challenges-f60dd37afec7 [https://perma.cc/X5UC-HT3Y].

ability to trace and identify the creator of deepfakes is likely to be highly effective in limiting them.[261] Traceability of the creator of a deepfake will increase social accountability, which is likely to deter many creators. Thus, the knowledge that a deepfake can be traced back to its creator can suffice for many to avoid creating them, even if no other legal sanction follows. Of course, for those who are not deterred by such identifiability, traceability can also support legal sanctions against creators of deepfakes.

Traceability also comes with a cost to creators and may chill innovation and creativity by AI users. Firms might therefore only implement traceability when this is necessary: when users create images and videos depicting *real people*, as opposed to any other type of content.

Traceability can be achieved through several technologies. Watermarking, a mechanism that allows embedding of metadata into digital outputs, is perhaps the leading technology in this context.[262] Watermarks are used in varied contexts: to assert ownership of copyrighted content, to authenticate content and its sources, to verify that content has not been altered, and to track the movement of content throughout the web.[263] Even when watermarking is imperceptible to users it can be algorithmically detectable[264] and can "be integrated into devices that people use to make digital contents to create immutable metadata for storing originality details such as time and location of multimedia."[265] Watermarking technology has seen a rapid development, often spurred by the need to mark AI-generated content as such.[266]

Due to the deterrence created by the ability to trace the creator of a deepfake, we advocate not only for a watermark connoting that a particular video or image has been algorithmically generated, but also allowing the

261. *See* Kaylee Williams, *Exploring Legal Approaches to Regulating Nonconsensual Deepfake Pornography*, TECH POLICY PRESS (May 15, 2023), https://www.techpolicy.press/exploring-legal-approaches-to-regulating-nonconsensual-deepfake-pornography/ [https://perma.cc/C6W7-6Y6P].

262. *See supra* Section IV.D.

263. Jie Ren, Han Xu, Pengfei He, Yingqian Cui, Shenglai Zeng, Jiankun Zhang, Hongzhi Wen, Jiayuan Ding, Pei Huang, Lingjuan Lyu, Hui Liu, Yi Chang & Jiliang Tang, *Copyright Protection in Generative AI: A Technical Perspective* 3, 10 (arXiv, Working Paper No. 2402.02333, 2024), https://arxiv.org/pdf/2402.02333 [https://perma.cc/CE8A-Q7RB].

264. John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers & Tom Goldstein, *A Watermark for Large Language Models* 1 (arXiv, Working Paper No. 2301.10226, 2024), https://arxiv.org/pdf/2301.10226 [https://perma.cc/V8LY-RKWR].

265. Nguyen et al., *supra* note 42, at 8; Robert Chesney & Danielle K. Citron, *Disinformation on Steroids*, COUNCIL FOREIGN RELS. (Oct. 2018), https://www.cfr.org/report/deep-fake-disinformation-steroids [https://perma.cc/SY6J-7SKN (staff-uploaded, dark archive)]; Tiffany Hsu, *Google Joins Effort to Help Spot Content Made With A.I.*, N.Y. TIMES (Feb. 8, 2024), https://www.nytimes.com/2024/02/08/business/media/google-ai.html [https://perma.cc/86ZQ-DR6W (staff-uploaded, dark archive)]. Regarding the general ability to watermark the output of Generative AI, see JaeYoung Hwang & SangHoon Oh, *A Brief Survey of Watermarks in Generative AI*, 14th INT'L CONF. INFO. & COMM. TECH. CONVERGENCE 1157, 1158 (2023).

266. *See infra* Section IV.D.

identification of the creator of the file. This can be done by registering several types of metadata such as the IP from which the creator accessed the model, requiring users to access models through preregistered accounts, or other identification techniques.[267] Such measures are entirely obtainable and have already been included in AI tools. Synthesia is an AI video startup operating a video generator and a text-to-video model.[268] After Synthesia's technology was used to create misinformation, the company increased its protection measures, applying "a watermark with information on where and how the AI avatar videos were created."[269] Synthesia also keeps a record of each and every video created through its system.[270]

2. Nonidentifiability

The harms to individuals who are the victims of deepfake outputs stem largely from their identifiability. The FTC recognizes that individuals who are the subject of a deepfake are likely to "suffer direct harm to their identities . . . and are less likely to be able to repair the reputational injuries."[271]

Companies developing and deploying generative AI build various restrictions into their models. Google prohibits the use of its generative AI tools to generate "[s]exually explicit content."[272] Companies may be able to avoid liability for the harms generated using their models if they have strong policy and technology protections prohibiting the use of identifiable individuals in their models' output.

3. Context Matters

The context in which a model operates should influence the extent to which companies should be expected to implement safeguards. If a model allows users to generate image and video outputs in problematic and potentially extremely harmful contexts, such as sexually explicit images, the output should

267. *C2PA in ChatGPT Images*, OPENAI, https://help.openai.com/en/articles/8912793-c2pa-in-chatgpt-images [https://perma.cc/T3AL-AXAQ (staff-uploaded archive)] (last modified Sep. 6, 2025). Another way to increase traceability of creators is to require registration with a government-issued ID so that each deepfake file can be traced back to a natural person. The need to balance privacy and accountability has been discussed in depth in the context of blockchain technology. *See, e.g.*, Ivan Damgård, Chaya Ganesh, Hamidreza Khoshakhlagh, Claudio Orlandi & Luisa Siniscalchi, *Balancing Privacy and Accountability in Blockchain Identity Management*, CRYPTOGRAPHER'S TRACK RSA CONF. 552, 552–54 (2021).

268. SYNTHESIA, https://www.synthesia.io/ [https://perma.cc/L8NK-UYBD] (last modified Oct. 10, 2025).

269. Heikkilä, *supra* note 125.

270. *Id.*

271. Trade Regulation Rule on Impersonation of Government and Businesses, 89 Fed. Reg. 115017, 115024 (Mar. 1, 2023) (to be codified at 16 C.F.R. pt. 461).

272. *Privacy & Terms*, GOOGLE, https://policies.google.com/terms/generative-ai/use-policy [https://perma.cc/AZE3-VYJJ] (last modified Dec. 17, 2024).

be traceable to its creator, should not include an identifiable individual, or possibly both. A model that creates a benign output could be held to a lower legal standard. There is no reason to restrict the activity or desired anonymity of a user prompting a model to generate an avocado armchair, for example.[273]

In practice, the three design recommendations we discussed in this Section should result in a reality whereby sexual deepfakes are traceable to their creator and do not depict identifiable individuals as their subject.

Ultimately, when generative-AI legislation is adopted, it will establish guidelines similar to those described here. As it could take a very long time for regulation of generative AI to become law, these guidelines can already serve as a basis for courts to develop the tests for liability under existing common law doctrines. As acknowledged by an FTC official, "rulemakings—while not a substitute for a legislative fix—can help ensure that lawbreakers do not profit from their lawbreaking and that wronged consumers can be made whole."[274] The same is true for a judicial response. While not a substitute for regulation, it can offer a temporary solution.

C. *Duty of Care*

In addition to a product defect, a successful products liability claim necessitates showing a duty of care towards the plaintiff. Thus, the plaintiff harmed by the product must be able to show that the defendant-manufacturer was under a duty to design the product in such a way that it would not be harmful to persons of the plaintiff's class or type.[275] Under the traditionally narrow concept of products liability, courts operationalized the duty of care requirement using privity tests, usually limiting the duty of care only to consumers who bought the product from the defendant, with courts discussing the possibility of expanding privity further down the supply chain. Later on, most courts adopted a more liberal approach, recognizing a duty of care towards all reasonably foreseeable victims.[276] Under such a test, the manufacturer is under a duty to design the product in such a way that considers the possibility of harms to all those foreseeably coming in contact with the product, including buyers, users, and, in appropriate cases, bystanders and third parties.[277]

273. Travis Tang, *Have You Seen this AI Avocado Chair?*, MEDIUM (Jan. 7, 2021), https://medium.com/data-science/have-you-seen-this-ai-avocado-chair-b8ee36b8aea [https://perma.cc/D7YN-NDZU].

274. Trade Regulation Rule on Impersonation of Government and Businesses, 89 Fed. Reg. 15030, 15031 (Mar. 1, 2024) (to be codified at 16 C.F.R. pt. 461).

275. Mark A. Geistfeld, *Situating Bystanders Within Strict Products Liability*, 18 BROOK. J. CORP. FIN. & COM. L. 1, 9–10 (2023) (discussing the differences between consumers and bystanders as plaintiffs in products liability law).

276. OWEN & DAVIS, *supra* note 243, at 393–99.

277. *Id.* at 293–95.

In the context of deepfakes, it seems reasonable to find that manufacturers of generative AI tools hold a duty of care towards girls and women whose image has been manipulated to create sexual deepfakes. The use of image-and-video-generating tools to create sexual deepfakes is now so pervasive that it would be strange to argue that girls and women victimized by the technology are not foreseeable plaintiffs. Even if the companies initially could not have reasonably imagined that their product would be used to harm these victims, they can no longer make that argument. The use of generative AI tools to algorithmically generate explicit sexual deepfakes has turned into a full-fledged epidemic[278] and is now not only foreseeable, but impossible to ignore.

Adopting a duty of care, especially towards women and girls who can foreseeably be harmed by sexually explicit deepfakes is in line with current regulatory trends. Indeed, as FTC commissioners have publicly stated, "Ensuring that the upstream actors best positioned to halt unlawful use of their tools are not shielded from liability will help align responsibility with capability and control."[279] Similarly, the Kids Online Safety Act (KOSA)[280] approved by the Senate on July 30, 2024, is a federal bill designed to protect children from online harms.[281] The Act establishes a duty of care for platforms covered by the Act towards minors.[282] In particular, the duty of care requires covered platforms to "exercise reasonable care in the creation and implementation of any design feature to prevent and mitigate . . . harms to minors."[283] The harms include bullying, eating disorders, substance abuse, and sexual exploitation and abuse.[284] The duty of care towards minors established in KOSA may signal a trend in the regulation of technology platforms. It is not possible for regulators to prohibit every single harmful type of behavior the platforms adopt. Addressing such harms through the establishment of a duty of care is therefore the best course of action—in the case of social media platforms as done by KOSA as well as in the context of generative AI models as we propose here.

Companies may argue that women and girls harmed by deepfakes are not directly connected to the company, and that the companies' actions (or lack thereof) only had an indirect role in the harmful outcome. Yet while deepfake victims have had no direct contact with the companies, they are foreseeable

278. Singer, *supra* note 131.

279. Trade Regulation Rule on Impersonation of Government and Businesses, 89 Fed. Reg. 15030, 15031 (Mar. 1, 2024) (to be codified at 16 C.F.R. pt. 461).

280. S. 1748, 119th Cong. (2025).

281. Barabara Ortutay, *What to Know About the Kids Online Safety Act that Just Passed the Senate*, ASSOCIATED PRESS (July 31, 2024), https://apnews.com/article/congress-social-media-kosa-kids-online-safety-act-parents-ead646422cf84cef0d0573c3c841eb6d [https://perma.cc/64F9-6DGV] (last updated July 31, 2024, at 4:02 ET).

282. Kids Online Safety Act, S. 1748, 119th Cong. § 102(a) (2025).

283. *Id.*

284. *Id.*

victims of the companies' negligent behavior. If an individual leaves a loaded gun in a playground, it is reasonably foreseeable that someone might fire it and harm another. Whoever left the gun there can be liable for negligently creating a risk towards foreseeable plaintiffs, even if the trigger was pulled by someone else. By the same token, an AI company can be made liable if the unsafe product it released is used to harm others. Companies must anticipate harm to foreseeable victims and act to install cost-effective measures to minimize it. Applying products liability to companies developing deepfake technology would change their incentives in a way that is likely to induce them to install into their products measures that would minimize the harms stemming from them.

A court wishing to adopt an expansive view of the companies' liability could identify another class of possible plaintiffs: AI users who unwittingly expose themselves to legal and nonlegal risks. In the playground example, the individual who left the gun in the park might owe a duty of care also to the child who shot the gun, thus hurting another. Such a child could experience ensuing distress or trauma and face risks of both social and legal sanctions. Whoever left the gun in the park can be made liable for this harm. In the case of generative AI, companies may owe a duty of care not only to the potential victims, but also to users who have been given a product that does not include sufficient safety measures.

Holding companies developing deepfake technology accountable for harms stemming from political deepfakes may prove to be more challenging. While generating deepfakes in the political context can be embarrassing to politicians and harmful to democracy, it involves complex questions relating to freedom of speech and the First Amendment.[285] Indeed, political deepfakes raise conflicting interests: on one hand, political speech enjoys the highest level of First Amendment protection, while at the same time, such broad protection may undermine trust in democratic elections and institutions.[286] The policy considerations in regulating deepfakes in the political context are wildly different than those governing the thinking about sexually explicit deepfakes. In this context, it is unclear who the direct victim is and to whom a duty of care might be owed. "Harm to democracy" is not suffered by any specific individual,

285. Helen Norton, *Lies to Manipulate, Misappropriate, and Acquire Governmental Power*, *in* LAW AND LIES: DECEPTION AND TRUTH-TELLING IN THE AMERICAN LEGAL SYSTEM 143, 170–71 (Austin Sarat ed., 2015); Ice, *supra* note 20; *see* Sobel, *A Real Account of Deep Fakes*, *supra* note 117, at 13, 55–66 (discussing First Amendment aspects of legal prohibitions on deepfakes); Daxton R. "Chip" Stewart & Jeremy Littau, *The Right to Lie with AI? First Amendment Challenges for State Efforts to Curb False Political Speech Using Deepfakes and Synthetic Media*, 24 COLO. TECH. L.J. (forthcoming 2025) (manuscript at 13); Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1790 (noting the free expression challenges involved in regulating deepfakes, particularly in the political context).

286. Ice, *supra* note 20, at 418.

and it is therefore likely that courts will not recognize a duty of care in this context.

D. *Remedies*

This Section details the possible remedies that plaintiffs may seek in their manufacturer liability claims against companies developing and deploying generative AI. These include compensation for harm and disgorgement of profits. We detail the strengths and challenges involved with each remedy and evaluate the conditions for its availability.

1. Compensation for Harm

Victims of deepfake pornography can sue companies developing AI for damages to compensate for harm caused. For example, a young girl who was the subject of an algorithmically generated deepfake created using technology developed by one of the generative AI companies has suffered significant harm and can sue the company that developed the AI used to harm her.

However, plaintiffs like her may need to overcome some challenges in making such claims. First, the company may point to the fact that the harmful deepfake was created by a third party, thus cutting off the chain of causation connecting their actions to the harmful result.[287] Second, courts might find it difficult to offer compensation based on products liability theories when the plaintiff did not suffer a physical injury.[288] Although products liability doctrine recognizes the possibility of compensation for emotional harms,[289] this is usually limited to cases in which the emotional harm is the result of some physical harm.[290]

Regarding the chain of causality—the fact that a third party was involved does not negate the liability that the company has for creating the risk. Similarly, if a person loads a gun and puts it in the middle of a playground, they are liable for the harm if a child picks up the gun and shoots someone. Second, regarding the nature of emotional harms, the response might be similar to the discussion of the duty of care above. Thus, limitations on the compensation of pure emotional harms are usually mandated to establish a clear limit on the number of some plaintiffs, similar to a duty of care.[291] If all emotional harms are

287. The but-for test for causation is a significant limitation on tort liability and involves significant difficulties. Maytal Gilboa, *Multiple Reasonable Behaviors Cases: The Problem of Causal Underdetermination in Tort Law*, 25 LEGAL THEORY 77, 85–89 (2019); ARIEL PORAT & ALEX STEIN, TORT LIABILITY UNDER UNCERTAINTY 58 (2001).

288. Jane B. Silverman, *Recovery for Emotional Distress in Strict Products Liability*, 61 CHI.-KENT L. REV. 545, 545–46 (1985) (discussing the possibility of compensation for emotional harm in products liability law).

289. *Id.*

290. *Id.*

291. *See id.* at 563 (discussing the need to "draw very fine distinctions between . . . users").

compensable, too many plaintiffs can approach courts, and it is not clear what distinguishes plaintiffs from others.[292] This does not seem to be a major concern in the context of deepfake pornography. Even if the harm to the victim is primarily emotional, it is accompanied by a specific injury special to the plaintiff—the use of her digital image. Therefore, there is no danger that allowing compensation for emotional harm will allow claims by too many indirect plaintiffs, since the class of plaintiffs is well defined.

2. Disgorgement of Profits

A plaintiff who is a victim of a tort may "waive the tort"[293] and sue for the defendant's profits instead of seeking compensation for harms.[294] Plaintiffs could use this ability when the defendant's gains were higher than the plaintiff's losses or easier to prove.[295] In the context of deepfake manufacturer liability, this alternative remedy may be attractive, considering the challenges described above. This type of remedy is meant to reflect the general principle of the law of unjust enrichment, that one should not be allowed to benefit through a wrong they perform.[296] This is also considered necessary to induce deterrence: As long as a socially harmful activity remains privately profitable for the defendant, it should be expected to persist. Only once ill-gotten gains are removed will defendants have an incentive to change their conduct.[297]

In the context of deepfakes, if AI manufacturers provide inherently dangerous products to children and adolescents, their profits from sales may be considered unjust enrichment and subject to disgorgement. This is an extreme solution, usually reserved for the profits obtained through sales of items such as

292. *See id.*

293. Daniel Friedmann, *Restitution of Benefits Obtained Through the Appropriation of Property or the Commission of a Wrong*, 80 COLUM. L. REV. 504, 505 (1980) (explaining the concept of "waiver of tort" and the remedy of disgorgement of profits as alternatives to tort damages).

294. *Id.*

295. RESTATEMENT (THIRD) OF RESTITUTION AND UNJUST ENRICHMENT § 13 (A.L.I. 2011) ("Remedies in restitution may offer important advantages over a remedy in damages . . . . [I]t is easier for the claimant to show the price that was paid (and to reverse the transaction) than to prove the amount of the resulting injury. . . . When a fraudulent transaction has been profitable to the defendant, restitution allows the claimant to obtain disgorgement of the defendant's consequential gains—and thereby to recover more than the claimant's loss.").

296. James J. Park, *Rule 10b-5 and the Rise of the Unjust Enrichment Principle*, 60 DUKE L.J. 345, 345 (2010) ("In numerous areas, the courts have applied Rule 10b-5 to deceptive conduct that is not directed at the market or investors but unjustly enriches some individual. . . . Securities regulation is guided by an evolving principle that sets some limits on the ability to extract wrongful gains from the securities markets.").

297. *See* Ayelet Gordon-Tapiero & Yotam Kaplan, *Unjust Enrichment by Algorithm*, 92 GEO. WASH. L. REV. 305, 309–10 (2024) ("[The law of unjust enrichment] is meant to ensure that misconduct does not pay and to remove the incentive to act in ways that are harmful to others."); Ofer Grosskopf, *Protection of Competition Rules Via the Law of Restitution*, 79 TEX. L. REV. 1981, 1997–98 (2001) (noting that disgorgement of wrongdoers' gains will disincentivize wrongdoings).

drugs[298] or, in some situations, guns.[299] Considering the scope and harms of the deepfake crisis, and especially the vulnerability of children involved, this remedy might well be justified in the present context. Of course, the intention here is not to categorically deprive AI manufacturers of their profits. Even the sale of dangerous items such as guns or highly addictive drugs is legitimate, if done legally and with appropriate precautionary measures. But if a dealer sells dangerous drugs to children, or a manufacturer provides teenagers with guns produced with no basic safety measures, then their profits may constitute unjust enrichment.[300]

By the same token, to avoid this form of liability, all that AI manufacturers need to do is ensure their products are designed in a safe way and marketed responsibly. As detailed above,[301] the large companies leading the market for generative AI are incredibly valuable. There are two ways to think about the enrichment generated by the company's wrongful actions. The first is through analyzing the expenses that the company would have paid to fulfill its duty of care. Ensuring traceability and nonidentifiability of people in problematic outputs costs money. Companies face high costs for research, development, implementation, and appropriate training for their workers to ensure the upholding of their duty of care. Any enrichment that the company enjoyed because it refrained from ensuring it complied with a reasonable duty of care should be viewed as wrongful and should be disgorged.

Another view is also possible. Under this alternative, the profits of a company that refrained from taking actions that would fulfill its duty of care are fully tainted. This framing of the enrichment element would allow plaintiffs to view all profits obtained through sales of the unsafe product as part of the enrichment stemming from the company's wrongful actions.

To establish entitlement for the disgorgement remedy, plaintiffs will also need to show that the enrichment by deepfake manufacturers has been generated *at their expense*. This important doctrinal element can be established if the companies violated their duties of care towards the plaintiffs—in the case at hand, the girls and women who have been humiliated and shamed. Enrichment in this case is not at the immediate financial expense of the plaintiff, but rather stems from the fact that the company violated its duty of care towards them.

298. *E.g.*, United States v. Lane Labs-USA, Inc., 427 F.3d 219, 236 (3d Cir. 2005) (upholding the lower courts order of restitution of profits obtained through the sales of diagnostic tests that did not comply with FDA regulations).

299. *E.g.*, City of Boston v. Smith & Wesson Corp., No. 1999-02590, 2000 Mass. Super. LEXIS 352, at *77–79 (July 13, 2000).

300. *Id.*

301. *Supra* Section I.B.

Viewing companies' enrichment from sexually explicit deepfakes as wrongful does not stem from a disdain for technology or for generative AI. Quite the opposite. Leaders in the AI industry are actively calling on lawmakers to take action and regulate various aspects of their activity, noting the potential dangers of inaction.[302] Our proposal strikes a delicate balance, allowing the ongoing development and deployment of generative AI technology, while pushing companies to recognize the duty of care they have to victims of foreseeable harms. Our hope is that practitioners begin to pursue deepfake claims under theories of products liability and that courts will begin to develop these theories in the context of AI tools. Such a development would motivate companies developing deepfake technology to implement the precautionary measures we discuss in this Article. Such a result would not only be financially beneficial for the companies but would also prevent much harm, pain, and suffering to users and to third parties.

IV. CHALLENGES AND IMPLICATIONS

This part offers a detailed discussion of our proposal introduced in Part III. First, we compare our proposal to existing concepts of liability described in the literature, highlighting its advantages.[303] Second, we study the issue of standing and point out the possibility of positive spillover effects, which might provide indirect responses to issues related to political deepfakes, through the application of manufacturer liability for sexually explicit deepfakes.[304] Third, we explain our proposal in relation to the concept of "friction," now increasingly used in legal scholarship. We explain that our proposal is not meant to completely prevent the creation of deepfakes; instead, it is enough that creating deepfakes becomes even slightly more burdensome for users to turn the tide of the crisis.[305] Fourth, we discuss the technological aspects of our proposal, studying watermarking tools necessary for its implementation.[306] Finally, we study the issue of open-source code release and the challenges it might involve for our proposal.[307]

A. *Comparing Modalities of Liability*

This Section compares deepfake manufacturer liability, as proposed above, with the concepts of user liability and platform liability as studied in prior

302. Kang, *supra* note 176; Bhuiyan, *supra* note 177; Metz & Schmidt, *supra* note 177.
303. *Infra* Section IV.A.
304. *Infra* Section IV.B.
305. *Infra* Section IV.C.
306. *Infra* Section IV.D.
307. *Infra* Section IV.E.

literature.[308] Our argument in this context is simple: while manufacturer liability is not a panacea, it is by far the best solution out of the three options.

Currently, platform liability is a legal impossibility, pending fundamental reconceptualization of Section 230, or its repeal. And while this may be a desirable course of action, it does not seem politically or legally feasible. Such a legal change probably requires legislative intervention and seems highly unlikely in the near future.[309] By comparison, manufacturer liability, even if it requires some subtle modifications in common law doctrine, does not require any new legislation. Products liability law, unlike Section 230, is fully developed by courts,[310] who have authority to responsibly develop the doctrine,[311] within its natural limits.[312]

User liability is also a highly limited solution. Attributing liability to individuals, whether users who create deepfakes or those who share them, is unrealistic and unscalable.[313] More important, focusing on users seems a deeply flawed strategy in terms of identifying cheap cost avoiders. Users have little control over the reality of AI tools, and educating users about AI safety would be a never-ending effort. Focusing on companies developing AI tools, those that can easily change and control the reality of AI technology makes for far better policy.

While our proposal is likely to face challenges, it is the most feasible path forward. Shortcomings and limitations of this proposed regime, though they may exist, should not serve as an excuse to abandon it completely, leaving past and future victims vulnerable and defenseless. We have little doubt that the field of generative AI will eventually be regulated along very similar lines to those minimal standards of care we detail above.[314] Until that day, the concept of manufacturer liability can serve as an important intermediate measure.

Importantly, our proposal is in line with the main trends in contemporary scholarship. Although existing literature has not studied products liability in the context of deep fake technology, scholars have explored products liability as a legal response to AI technology more generally. In particular, Professor Catherine Sharkey argues that products liability is an appropriate liability

308. *See supra* Sections II.B–C.

309. Aaron Mackey & Joe Mullin, *Sunsetting Section 230 Will Hurt Internet Users, Not Big Tech*, ELEC. FRONTIER FOUND. (May 20, 2024), https://www.eff.org/deeplinks/2024/05/sunsetting-section-230-will-hurt-internet-users-not-big-tech [https://perma.cc/X4EN-TKG3 (staff-uploaded archive)].

310. Goldberg & Zipursky, *supra* note 229, at 1923.

311. Kaplan et al., *supra* note 231, at 1446.

312. *Id.*

313. Chesney & Citron, *Deep Fakes*, *supra* note 9, at 1792.

314. *Supra* Section III.B.

framework for AI.[315] She notes that the companies developing AI models are in the best position to be aware of the defective aspects of their products, making them cheapest cost avoiders and therefore natural defendants.[316] One of the advantages of improving the design of the technology is the broad implications it will have on the generation of all deepfakes.

B. *Standing & Spillover*

Manufacturer liability, as described above, seems plausibly applicable in the context of sexually explicit deepfakes, but less clearly relevant in the context of political deepfakes. The reason for this difference is related to the concept of duty of care as described above[317] and to the issue of standing.[318]

Thus, in cases of explicitly sexual deepfakes, recognizing an injured party is straightforward. The injured party is the individual depicted in the deepfake, who therefore has standing to sue based on their injury.[319] Standing is more complicated when political deepfakes are concerned. While political deepfakes may be harmful to the individuals they depict, they are more harmful to society at large or to democratic order. And while these are important and real harms, it is not easy to identify an injured party; therefore, standing might be difficult to establish in such cases.

However, we believe that our proposal to apply manufacturer liability to deepfakes can also have positive indirect implications for political deepfakes.[320] To avoid manufacturer liability in cases of deepfake pornography, AI manufacturers will implement basic safeguards in their product, as described above.[321] For example, they will ensure that users can create sexual deepfakes only in a text-to-image or text-to-video setting and not based on an image or video of an identifiable individual. These safety measures will have positive spillover effects on the realm of political deepfakes. Thus, if political deepfakes

315. *See generally* Catherine M. Sharkey, *A Products Liability Framework for AI*, 25 COLUM. SCI. & TECH L. REV. 240 (2024) (arguing that a products liability framework is a promising avenue to tackle AI-driven harms).

316. Catherine M. Sharkey, *Products Liability in the Digital Age: Online Platforms as "Cheapest Cost Avoiders*," 73 HASTINGS L.J. 1327, 1333–34 (2022) (noting the economic benefit of holding developing companies liable for harms stemming from use of their product).

317. *Supra* Section III.C.

318. *E.g.*, McCrory v. Adm'r of Fed. Emergency Mgmt. Agency of U.S. Dep't of Homeland Sec., 600 F. App'x 807, 808 (2d Cir. 2015) ("As a threshold inquiry, a federal court must determine that the plaintiff has constitutional Article III standing prior to determining . . . the subsequent merits of the case.").

319. Conservation L. Found., Inc. v. Shell Oil Co., 628 F. Supp. 3d 416, 433 (D. Conn. 2022) (discussing injury as a basis for standing).

320. *Supra* Subsection I.C.2.

321. *Supra* Section III.B.

are traceable, the fact they are not real may become more easily apparent to viewers.[322]

C. *Friction*

In recent years, scholars have increasingly recognized the role of "friction" as a governance tool.[323] Under this theory, to prevent undesirable phenomena and problematic manners of conduct, it is often unnecessary to prohibit them, and it may instead be more efficient to introduce "friction," that is any minor obstacle that anyone wishing to engage in the unwanted activity would have to overcome.[324]

The safeguards we discuss above, such as the integration of watermarks that would enable tracing the creator of a deepfake and restricting the ability to create harmful deepfakes of identifiable individuals,[325] can serve as a form of friction. Even if AI-generated content depicting real people is traceable to its creators, it is unlikely to completely eliminate deepfakes. Users who are determined or sophisticated enough will find ways to circumvent such basic safeguards. Yet these safeguards are a form of friction, ensuring that the creation of harmful deepfakes is not as seamless as it is today. This extra effort will be sufficient to deter the vast majority of users, those who casually create harmful content without giving the issue too much thought and without fully considering the implications of their actions. The goal of our proposal is not to eradicate deepfakes completely. This would probably be impossible, and hopefully unnecessary. Rather, our aim is to contend with the deepfake *pandemic* and with the incredibly alarming surge of harmful content we are currently witnessing. Hopefully, once sexual deepfakes are not as common, their creation might again be considered as a clear violation of conventional moral norms, rather than as normalized.

The notion of friction is also very much in line with our proposal also since friction is typically considered a temporary measure: "Friction can sometimes be a good intermediate solution without necessarily being the end goal. Friction can provide policymakers with time they may need to address the underlying

322. Margot Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1529, 1577 (2019) (discussing the combining of individual and collaborative governance regimes).

323. Ayelet Gordon-Tapiero, Paul Ohm & Ashwin Ramaswami, *Fact and Friction: A Case Study in the Fight Against False News*, 57 U.C. DAVIS L. REV. 171, 174 (2023) ("Often the friction is aimed at promoting an important human value beyond simple efficiency or profit maximization, such as to advance fairness, trust, or consensus.").

324. Ellen P. Goodman, *Digital Fidelity and Friction*, 21 NEV. L.J. 623, 648 (2021); William McGeveran, *The Law of Friction*, 2013 U. CHI. LEGAL F. 15, 51 (2013) (referring to friction as "any sort of irritating obstacle" that might divert people from unwanted courses of action).

325. *Supra* Section III.B.

cause of the problem."[326] Our proposal is to have manufacturer liability provide the necessary friction for users, holding the deepfake crisis in check, until such time that legislators and regulators can provide a comprehensive legal response to the problem.

D. *Watermarking*

Traceability of AI-generated content, when appropriate and required as described above,[327] will likely be achieved through some form of watermarking technology.[328]

When traceability of AI-generated content is warranted, several types of watermarking mechanisms might be used.[329] First, a watermark can indicate that a particular file has been generated by AI. The ability to detect that certain content has been algorithmically generated will become increasingly important. Already, leading platforms are attempting to identify and label such content.[330] Second, a watermark may identify the user who created the deepfake.

Integrating watermarks into the outputs of generative AI models is not a simple task.[331] It is, however, a viable option, one that is already being

326. Gordon-Tapiero et al., *supra* note 323, at 249.

327. *Supra* Section III.B.

328. Maurice Schellekens, *Digital Watermarks as Legal Evidence*, 8 DIGIT. EVIDENCE & ELEC. SIGNATURE L. REV. 152, 152 (2011) ("A digital watermark is meta-information that can be added to a work such as a picture or a movie . . . that [can] identify the author or rights holder."); Rosemarie F. Jones, *Wet Footprints? Digital Watermarks: A Trail to the Copyright Infringer on the Internet*, 26 PEPP. L. REV. 559, 569 (1999) ("Watermarks could also deter counterfeiters from making illegal copies because an imitation would be easily identifiable from the original."). *See supra* Section III.B.1.

329. Chloe Wittenberg, Ziv Epstein, Adam J. Berinsky & David G. Rand, *Labeling AI-Generated Content: Promises, Perils, and Future Directions*, PUBPUB (Mar. 27, 2024), https://mit-genai.pubpub.org/pub/hu71se89/release/1?readingCollection=9410b119 [https://perma.cc/TY5P-DWR5]; Yunato Wang, Yanghe Pan, Miao Yan, Zhou Su & Tom H. Luan, *A Survey on ChatGPT: AI-Generated Contents, Challenges, and Solutions* 1 (arXiv, Working Paper No. 2305.18339, 2023), https://arxiv.org/pdf/2305.18339 [https://perma.cc/Q3ST-SBX6].

330. Bickert, *supra* note 210; Kat Tenbarge, *YouTube Says It Will Require Creators to Label 'Realistic' AI Content*, NBC NEWS (Mar. 18, 2024, at 16:55 ET), https://www.nbcnews.com/tech/tech-news/youtube-says-will-require-creators-label-ai-content-rcna143937 [https://perma.cc/86M5-42MR]; Wittenberg et al., *supra* note 329, at 6 (explaining that one of the possible results of marking content as AI-generated could be downranking it by the platform's algorithm). *See generally* Gonzalo J. Aniano Porcile, Jack Gindi, Shivansh Mundra, James R. Verbus & Hany Farid, *Finding AI-Generated Faces in the Wild* (arXiv, Working Paper No. 2311.08577, 2024), https://arxiv.org/pdf/2311.08577 [https://perma.cc/7P5U-EL7G] (attempting to distinguish real faces from AI-generated ones).

331. Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang & Lei Lie, *Invisible Image Watermarks Are Provably Removable Using Generative AI* 1 (arXiv, Working Paper No. 2306.01953, 2024), https://arxiv.org/pdf/2306.01953 [https://perma.cc/657J-HBBG]; Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese & Boaz Barak, *Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models* 11–16 (arXiv, Working Paper No. 2311.04378, 2025), https://arxiv.org/pdf/2311.04378 [https://perma.cc/JT4J-9GMG]; Hwang & Oh, *supra* note 265, at 1158.

researched and experimented with.[332] Companies developing and deploying generative AI models are already attempting integration of watermarking techniques into the outputs of their models.[333] Researchers offer ways to efficiently mark the outputs of diffusion models, which are driving some of the leading generative AI models.[334] There is also extensive research regarding watermarking the output of generative AI models described above.[335] This even includes the ability to watermark text, long considered a difficult challenge.[336]

While these efforts are important first steps, particularly in terms of the technical feasibility of achieving such watermarking, it would be dangerous to rely on the goodwill of companies to voluntarily implement such techniques as a form of self-regulation. While some companies may find it important to develop watermarking techniques and may even use it as a marketing technique

332. Tanja Šarčević, Alicja Karlowicz, Rudolf Mayer, Ricardo Baeza-Yates & Andreas Rauber, *U Can't Gen This? A Survey of Intellectual Property Protection Methods for Data in Generative AI* 18–20 (arXiv, Working Paper No. 2406.15386, 2024), https://arxiv.org/pdf/2406.15386 [https://perma.cc/4TMU-HVG9]; Chris Stokel-Walker & Richard Van Noorden, *The Promise and Peril of Generative AI*, 614 NATURE 214, 214–16 (2023); Scott Aaronson, *My AI Safety Lecture for UT Effective Altruism*, SHTEL-OPTIMIZED (Nov. 28, 2022), https://scottaaronson.blog/?p=6823 [https://perma.cc/4A4K-KKPR]. *See generally* Xun Xian, Ganghua Wang, Xuan Bi, Jayanth Srinivasa, Ashish Kundu, Mingyi Hong, Jie Ding, *RAW: A Robust and Agile Plug-and-Play Watermark Framework for AI-Generated Images with Provable Guarantees* (arXiv, Working Paper No. 2403.18774v1, 2024), https://arxiv.org/pdf/2403.18774v1 [https://perma.cc/WE5N-UKLT] (proposing the introduction of watermarks into original image data); Tate Ryan-Mosley, *Cryptography May Offer a Solution to the Massive AI-Labeling Problem*, MIT TECH. REV. (July 28, 2023), https://www.technologyreview.com/2023/07/28/1076843/cryptography-ai-labeling-problem-c2pa-provenance/ [https://perma.cc/Y4FV-949C].

333. Emilia David, *OpenAI is Adding New Watermarks to Dall-E 3*, THE VERGE (Feb. 6, 2024, at 17:32 ET), https://www.theverge.com/2024/2/6/24063954/ai-watermarks-dalle3-openai-content-credentials [https://perma.cc/9KHN-2UPT (dark archive)]; Umar Shakir, *Google's Invisible AI Watermark Will Help Identify Generative AI Text and Video*, THE VERGE (May 14, 2024, at 14:05 ET), https://www.theverge.com/2024/5/14/24155927/google-ai-synthid-watermark-text-video-io [https://perma.cc/EMC9-SWQ8 (dark archive)].

334. Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung & Min Lin, *A Recipe for Watermarking Diffusion Models* 4–5 (arXiv, Working Paper No. 2303.10137, 2023), https://arxiv.org/pdf/2303.10137 [https://perma.cc/3ZB3-RVWT] (describing the watermarking method that can be applied to diffusion models such as those use by Stable Diffusion).

335. Ting Luo, Jun Wu, Zhouyan He, Haiyong Xu, Gangyi Jiang & Chin-Chen Chang, *WFormer: A Transformer-Based Soft Fusion Model for Robust Image Watermarking*, 8 IEEE TRANSACTIONS EMERGING TOPICS IN COMPUT. INTELL. 4179, 4182–84 (2024); Pierre Fernandez, Guillaume Couairon, Teddy Furon & Matthijs Douze, *Functional Invariants to Watermark Large Transformers*, INST. ELEC. & ELECS. ENG'RS INT'L CONF. ACOUSTICS, SPEECH & SIGNAL PROCESSING 4815, 4818 (2024); Jinmin Li, Kuofeng Gao, Yang Bai, Jingyun Zhang & Shu-Tao Xia, *Video Watermarking: Safeguarding Your Video from (Unauthorized) Annotations by Video-based LLMs* 2–6 (arXiv, Working Paper No. 2407.02411, 2024), https://arxiv.org/pdf/2407.02411 [https://perma.cc/KST2-8QMN]; Xiaodan Xing, Huiyu Zhou, Yingying Fang & Guang Yang, *Assessing the Efficacy of Invisible Watermarks in AI-Generated Medical Images* 2–4 (arXiv, Working Paper No. 2402.03473, 2024), https://arxiv.org/pdf/2402.03473 [https://perma.cc/RH3Y-T9SZ]; Hengyuan Xu, Liyao Xiang, Xingjun Ma, Borui Yang & Baochun Li, *Hufu: A Modality-Agnositc Watermarking System for Pre-Trained Transformers via Permutation Equivariance* 2–3 (arXiv, Working Paper No. 2403.05842v1, 2024), https://arxiv.org/pdf/2403.05842v1 [https://perma.cc/AL6B-W5W6].

336. Kirchenbauer et al., *supra* note 264, at 9–11.

to demonstrate the reliability of their models, other companies may *avoid* watermarking for marketing purposes. We would not want to witness the creation of a subsection of the AI industry specifically designed to appeal to users who wish to create harmful deepfakes. Mandating some form of watermarking, as discussed above, is therefore recommended.

Of course, watermarking is not airtight. A determined and savvy user might be able to tamper with watermarks, undermining the traceability of AI-generated content. As with other types of friction, this does not undermine the overall goal of watermarking. While removing or altering the watermark will be a viable option for some users, it will suffice to deter many others. Since sexually explicit deepfakes are often created by teens, it is reasonable to assume that many users would be deterred from creating deepfakes if such content is watermarked.

E. *Open-Source Release of Code*

Some AI companies make their source code open and publicly accessible,[337] meaning it can be inspected and modified by others.[338] Open-source software promotes collaboration,[339] and can enhance software quality in multiple ways.[340] It can also present challenges in the context of deepfake content.[341]

Thus, even if companies developing AI tools include reasonable safeguards in their products, downstream programmers may remove these safeguards if the company releases its source code.[342] Several responses to this problem might be

337. These include StabilityAI; OpenAI, which released the code for its LLMs up to GPT-2, although from GPT-3 and on the models are not open sourced; Google, which open-sourced the software for BERT; and others.

338. *What is Open Source?* OPENSOURCE.COM, https://opensource.com/resources/what-open-source [https://perma.cc/LV89-CLBQ]; Peter P. Swire, *A Model for when Disclosure Helps Security: What Is Different About Computer and Network Security?*, 3 J. TELECOMM. & HIGH TECH L. 163, 165 (2004) ("For proponents of Open-Source software, revealing the details of the system will actually tend to improve security . . . trying to hide the details of the system will tend to harm security because attackers will learn about vulnerabilities, but defenders will not know where to patch the vulnerabilities.").

339. ERIC S. RAYMOND, THE CATHEDRAL AND THE BAZAAR: MUSINGS ON LINUX AND OPEN SOURCE BY AN ACCIDENTAL REVOLUTIONARY 129 (Tim O'Reilly ed., rev. ed. 2001) (explaining that open-source code that has been peer reviewed is more secure than closing code); Prattay Sanyal, Shubham Sharma, Deepa Bura & Prasenjit Banerjee, *On the Security of Open Source Software*, 11 INT'L J. ADVANCED RSCH. 1338, 1339 (2002) ("Open Source Software is much better than closed software in terms of vulnerability concerning security.").

340. JONATHAN ZITTRAIN, THE FUTURE OF THE INTERNET—AND HOW TO STOP IT 70 (2008) (celebrating "unfiltered contributions from broad and varied audiences").

341. Chinmayi Sharma, *Tragedy of the Digital Commons*, 101 N.C. L. REV. 1129, 1178–85 (2023) (describing some of the challenges stemming from open sourcing of code). *See generally* NADIA EGHBAL, WORKING IN PUBLIC: THE MAKING AND MAINTENANCE OF OPEN SOURCE SOFTWARE (2020) (describing the security challenges involved in open source software).

342. Two of the central challenges in this context are likely to be issues relating to First Amendment protection as well as to the immunity provided by Section 230. In the context of the First

proposed. First, companies may indeed be exempt from liability in such cases, but those who have changed the code might be held liable. Second, an adjusted standard of care can be applied, requiring companies to only release their code under an open-source license prohibiting the removal of basic safeguards. A company that uses the license to limit the removal of the safeguards will be viewed as having fulfilled its duty of care, even if subsequent users violate the license and remove them. Finally, accessing and altering code requires a certain level of expertise and can therefore be considered a type of friction. Even if safeguards are eventually removed in some derivative products, their inclusion in released open-code software adds a level of difficulty for those who want to remove them. Thus, while open-source code may present some challenges to the application of legal requirements mandating safeguards in generative-AI tools, it by no means makes these safeguards superfluous.

CONCLUSION

As generative AI tools continue to evolve, they bring both immense promise and alarming risks. The rise of sexually explicit deepfakes has become a critical concern, with the number of deepfakes and their victims steadily increasing. This manipulated content inflicts severe pain and suffering on its targets, often resulting in lasting psychological harm. It is imperative that the legal system recognize the gravity of this phenomenon and take decisive action. Existing scholarship does not offer a plausible way forward. Legislation and regulation are too slow to act, and platform and user liability do not provide a suitable framework.

Cognizant of these challenges, we highlight a critical oversight in previous research. We propose that companies developing and deploying generative AI systems should be held liable for the harms created by their products under the doctrine of manufacturer liability. As manufacturers of the software driving deepfake technology, these companies owe a duty of care towards foreseeable victims, particularly those targeted by sexually explicit deepfakes. We discussed the challenges our proposal faces and the positive spillover effects it could generate. We also outlined the practical measures that courts could require companies to implement when developing the doctrine of manufacturer liability in the context of generative AI. While there are issues regarding standing in the context of political deepfakes and damages in the context of both political and sexual deepfakes, this Article demonstrates that those difficulties can be overcome. By relying on the common law to develop this area as legislation

Amendment, see Kyle Langvardt, *Crypto's First Amendment Hustle*, 26 YALE J.L. & TECH. 130, 152 (2023) ("[W]hen the interference with expression is merely incidental . . . courts apply intermediate scrutiny. . . . [C]ourts applying the allegedly formidable 'code is speech' principle have repeatedly upheld schemes that impose heavy, even excessive regulatory penalties on software developers.").

works its way through governing bodies, the courts can begin to place the burden of helping to reduce the deepfake phenomenon back on the manufacturers.