REPUTATION (NOT TAYLOR'S VERSION): REGULATING ARTIFICIAL INTELLIGENCE HALLUCINATED DEEPFAKES OF PUBLIC FIGURES

Khang-Christopher Duc Truong[*]

Abstract

The rapid advancement in complexity and capability of generative artificial intelligence (AI) has generated many issues in current U.S. law, including tort liability of reputational harm. As generative AI becomes more accessible and prevalent in the digital world, public figures, such as Taylor Swift, will become increasing targets of deepfakes. Unlike deliberate creation, AI hallucinations exposes uncertainty in identifying a tortfeasor of a deepfake causing reputational harm. Although AI generated deepfakes can currently be easily fact checked, as generative AI evolves, deepfakes will become more realistic and indistinguishable from real photos. This Note examines whether current law provides an opportunity for victims of hallucinated deepfakes to recover from reputational harm. Furthermore, regulations, such as mandatory information storage and tagging, are provided without upending current tort law and third-party liability law.

TABLE OF CONTENTS

* J.D., University of Illinois College of Law, 2024; B.S. Biomedical Engineering, Texas A&M University, 2017. I would like to thank Professor Jennifer K. Robbennolt for her indispensable feedback and support in the writing of this Note. I would also like to thank my friends, family, and mentors for their continual encouragement throughout law school.

I. INTRODUCTION

Teenage Country Star Wins Horizon Award, Captivating Listeners With her Single "Tim McGraw" and Self-Titled Debut Album.[1]

"Girl next door" and America's sweetheart faces backlash after public feuds with Kanye West, Kim Kardashian, and Katy Perry.[2]

Not only a Pop Star, but also a Feminist Icon and Ally for the LGBTQ+ Community.[3]

A History of Exes and a Host of Hints Fuel Speculation of Her Musical Inspiration.[4]

Billionaire Multi-Grammy Award Winner Demands College Student Stop Tracking Her Highly Scrutinized Jet.[5]

As her career progressed from aspiring country artist to international pop sensation, Taylor Alison Swift became a media darling and no stranger to many

1. *Taylor Swift Timeline*, OFF. TIMELINE (Nov. 2, 2023), https://www.officetimeline.com/blog/taylor-swift-timeline [https://perma.cc/V2KJ-YD88].

2. Charlotte Chilton, *The Evolution of Taylor Swift from Country Star to Pop Icon*, COUNTRY LIVING (Oct. 28, 2021), https://www.countryliving.com/life/entertainment/g36298587/evolution-taylor-swift/ [https://perma.cc/8BJ5-5P9H]. *See* OpenAI, *Response to: "What Is Taylor Swift's Reputation and Public Image?*", CHATGPT 3.5 (Mar. 3, 2024), https://chat.openai.com/ (entering query into "Message ChatGPT" box).

3. Susan Harmeling, *The Politics of Pop: DEI Lessons from Taylor Swift's Advocacy*, FORBES (Jan. 24, 2024), https://www.forbes.com/sites/susanharmeling/2024/01/23/the-politics-of-pop-dei-lessons-from-taylor-swifts-advocacy/ [https://perma.cc/UBF7-3DQX]. *See* CHATGPT 3.5, *supra* note 2 (generating information about Taylor Swift based on the prompt "What is Taylor Swift's Reputation and Public Image?").

4. Eric E. Surbano, *The Most Iconic Taylor Swift Songs About Her Exes*, LIFESTYLE ASIA (Oct. 26, 2023, 1:04 PM), https://www.lifestyleasia.com/bk/entertainment/taylor-swift-songs-about-exes-all-too-well-lover/ [https://perma.cc/M289-P9W9].

5. Drew Harwell, *Taylor Swift Threatens Legal Action Against Student Who Tracks Her Jet*, WASH. POST (Feb. 6, 2024), https://www.washingtonpost.com/technology/2024/02/06/taylor-swift-jet-tracking-legal-threat/ [https://perma.cc/BF6G-P87H]; Isabella O'Malley, *Why Taylor Swift's Globe-Trotting in Private Jets is Getting Scrutinized*, ASSOCIATED PRESS NEWS (Feb. 2, 2024), https://apnews.com/article/taylor-swift-climate-jet-carbon-emissions-kelce-chiefs-02ac425d24281bd26d73bfdf4590bc82 [https://perma.cc/QY89-5JCH].

headlines similar to the sample ones above.[6] The media has tracked every minute detail, good or bad, about her career and life.[7] They scrutinized and speculated on her accolades, commercial successes, romantic partners, celebrity feuds, clothes, political advocacy, and even presidential endorsements.[8] Despite the media attention and controversies, Taylor Swift has always been seen as in control of her brand, her public image, and her reputation.[9] For example, after receiving major backlash from her public feud, she reframed the bad publicity with the release of her album *Reputation*.[10]

In 2023, Taylor Swift had one of her best years.[11] Her ongoing sixth tour grossed over $1 billion in revenue, boosting her to billionaire status and drawing envy from countries that did not make the tour route.[12] She released a movie of her tour and two re-recordings of her old albums to commercial success across the world.[13] Time magazine recognized her commercial and critical accomplishments, announcing her as their Person of the Year.[14] Her superfans and sports newscasters obsessed over her budding relationship with a professional football player.[15] After a year of receiving overwhelmingly positive press, Taylor Swift's public image could have been hijacked by headlines such as *Popstar Poses for Provocative Pics*, and *T-Swift is Team Trump*.[16] Unlike the sample headlines introduced at the beginning of this Note, these more controversial sample headlines narrate fake events depicted by AI-generated images that flooded social media.[17]

---

6. Scottie Andrew, *Why Taylor Swift Is the Media's Favorite Subject — Even When the Story Isn't About Her*, CNN (Jan. 8, 2024, 5:52 PM EST), https://www.cnn.com/2024/01/08/entertainment/taylor-swift-golden-globes-cec/index.html [https://perma.cc/TJ9Z-GSJL].

7. *See id.* (noting how Swift's mention an article will increase viewership).

8. *Id.*; *See, e.g.*, Edith Olmsted, *Could Taylor Swift Swing the 2024 Election at the Super Bowl?*, THE NEW REPUBLIC (Feb. 7, 2024), https://newrepublic.com/article/178750/taylor-swift-gets-political-at-super-bowl [perma.cc/4VC7-SUBP].

9. Janya Sindhu, *Taking Control of Your Brand the Taylor Swift Way*, MEDIUM (Jan. 4, 2020), https://bettermarketing.pub/taking-control-of-your-brand-the-taylor-swift-method-6111fa501af [https://perma.cc/Z2YR-8L8K].

10. Chilton, *supra* note 2.

11. Megan Camponovo, *Taylor Swift's Biggest 2023 Moments*, WBALTV (Dec. 13, 2023, 12:16 PM EST), https://www.wbaltv.com/article/taylor-swift-2023/46106514 [https://perma.cc/Z8EY-D6D7].

12. *Id.*; Kathleen Magramo, *Singapore Defends Taylor Swift's Exclusive Southeast Asia Stop After Neighbors Cry Foul*, CNN (Mar. 5, 2024, 8:33 PM EST), https://www.cnn.com/2024/03/05/asia/singapore-taylor-swift-southeast-asia-intl-hnk/index.html [https://perma.cc/W2AS-THKZ].

13. Camponovo, *supra* note 11.

14. *Id.*

15. *Why Are People So Obsessed with Taylor Swift and Travis Kelce? Rolling Stone Staff Has Theories*, ROLLING STONE (Nov. 16, 2023), https://www.rollingstone.com/culture/culture-lists/taylor-swift-travis-kelce-relationship-mania-rolling-stone-staff-theories-1234879881/ [https://perma.cc/Q9RF-JLGB].

16. *See* Jess Weatherbed, *Trolls Have Flooded X with Graphic Taylor Swift AI Fakes*, THE VERGE (Jan. 25, 2024, 10:04 AM CST), https://www.theverge.com/2024/1/25/24050334/x-twitter-taylor-swift-ai-fake-images-trending [https://perma.cc/5C9F-Q5YS] (reporting how explicit AI-generated nonconsensual pornography of Swift attracts wide-spread attention); *see also* Kat Tenbarge, *Taylor Swift Deepfakes on X Falsely Depict Her Supporting Trump*, NBC NEWS (Feb. 7, 2024, 6:30 PM CST), https://www.nbcnews.com/tech/internet/taylor-swift-deepfake-x-falsely-depict-supporting-trump-grammys-flag-rcna137620 [https://perma.cc/5YBT-CEHW] (recalling how Trump supporters created AI deepfakes of Swift to falsely show her support of their candidate).

17. *Id.*

In January 2024, sexually explicit artificial intelligence (AI) generated images of Taylor Swift spread throughout social media websites.[18] Within 17 hours, one of these images "attracted more than 45 million views, 24,000 reposts, and hundreds of thousands of likes and bookmarks" on one website.[19] One month later, AI-generated videos and images of Taylor Swift supporting Donald Trump surfaced.[20] One post of such videos has over 10.3 million views.[21] These AI-generated photos called "deepfakes" are synthetic imagery created to facilitate fraud, misinformation, and reputational harm.[22]

Many public figures, like Taylor Swift, maintain their reputation by controlling the narrative of their actions.[23] In some instances, public figures must handle not only the public response to their actions but also the media speculation that turns out to be complete fabrications or conspiracy theories.[24] Now, with the mainstream access to generative AI, those fabrications can come in the form of deepfakes.[25] Normally, public figures can seek legal remedy for falsehoods that impact their reputation through defamation laws or misappropriation of their likeness through right of publicity laws.[26] In a case like Taylor Swift's, where the deepfakes are not created or distributed for a commercial purpose, a right of publicity claim is most likely unavailable.[27] This recent case involving Taylor Swift highlights gaps in defamation and other falsehood laws when AI-generated deepfakes are spread purely for sensationalism and misinformation.[28]

Part II will discuss the rapid development and technological structure of generative AI and the gradually evolving legal landscape surrounding AI. Part III will discuss the potential legal liability to software developers when generative AI generates deepfakes of a public figure. Part IV will argue that federal regulations that will ease identification of tortfeasors and deepfakes will

18. Weatherbed, *supra* note 16.

19. *Id.*

20. Tenbarge, *supra* note 16.

21. *Id.*

22. *Understanding the Different Types of Generative AI Deepfake Attacks*, IPROOV (Nov. 30, 2023), https://www.iproov.com/blog/generative-ai-attack-types-explained [https://perma.cc/DJE7-CK5H].

23. *See* Chilton, *supra* note 2 (detailing the changes that Taylor Swift has made to her brand throughout the years).

24. *See* Jacob Shamsian, *The 18 Wildest Celebrity Conspiracy Theories on the Internet*, BUS. INSIDER (Nov. 21, 2017, 9:05 AM CST), https://www.businessinsider.com/celebrity-conspiracy-theories-2017-5 [https://perma.cc/2LJL-WQE2] (detailing examples of Avril Lavigne and Katy Perry dealing with conspiracy theories based on pure speculation).

25. Geoff Mulvihill, *What to Know About How Lawmakers are Addressing Deepfakes like the Ones that Victimized Taylor Swift*, ASSOCIATED PRESS NEWS (Jan. 31, 2024), https://apnews.com/article/deepfake-images-taylor-swift-state-legislation-bffbc274dd178ab054426ee7d691df7e [https://perma.cc/Z9CK-9T44].

26. *See, e.g.*, Matthew Ferraro & Louis Tompros, *Celebrity Disinformation Victims Have Panoply of Remedies*, LAW360 (Sept. 29, 2020, 4:05 PM EDT), https://www.law360.com/articles/1312609/ [https://perma.cc/PUW2-6GV8] (presenting California defamation law as an option for celebrity disinformation).

27. Kevin Frazier, *Swift Justice? Assessing Taylor's Legal Options in Wake of AI-Generated Images*, TECH POLICY PRESS (Feb. 27, 2024), https://www.techpolicy.press/swift-justice-assessing-taylors-legal-options-in-wake-of-aigenerated-images/ [https://perma.cc/QM2J-G7VD].

28. Sara H. Jodka, *Manipulating Reality: The Intersection of Deepfakes and the Law*, REUTERS (Feb. 1, 2024, 11:01 AM CST), https://www.reuters.com/legal/legalindustry/manipulating-reality-intersection-deepfakes-law-2024-02-01/ [https://perma.cc/HN78-7983].

facilitate existing state tort laws without offending free speech and hampering innovation.

II. BACKGROUND

A. The Generation of Artificial Intelligence

In 1950, the race for creating AI most famously entered mainstream academic thought with Alan Turing's paper *Computing Machinery and Intelligence*.[29] Turing imagined a process that involved asking simple yes or no questions to test for successful imitation of human thought.[30] In the early days of development, AI, valued for its potential to learn, reason, and problem solve like human beings,[31] flourished in various government projects.[32] Although government interest waned in the 90s, private companies began to invest in AI, producing headlines with their advancements.[33] AI models beat world champions in complicated games, such as chess and Go, reflecting the complexity that AI had achieved.[34] Within 70 years of development, attempts to simulate human intelligence evolved from answering yes or no questions to playing chess, and, now, to limitedly mimicking human speech and thought.[35]

Although AI has been around for a while, its explosion in recent years is due to technological developments enabling the effective development of generative AI.[36] Generative AI describes algorithms and software models that create new content, including audio, code, images, text, simulations, and videos.[37] Unlike AI models that came before it, generative AI not only identifies

29. Rockwell Anyoha, *The History of Artificial Intelligence*, HARV.: THE GRADUATE SCH. ARTS & SCI. (Aug. 28, 2017), https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/ [https://perma.cc/5ZU4-Y4XJ].

30. *Id.*; A. M. Turning, *Computing Machinery and Intelligence*, 236 MIND 433, 445 (1950); B.J. Copeland, *Artificial Intelligence*, BRITANNICA (Oct. 11, 2023), https://www.britannica.com/technology/artificial-intelligence [perma.cc/XY5W-Y2VR].

31. *See* Copeland, *supra* note 30 (asserting that most AI research has focused on learning, reasoning, problem solving, perception, and using language).

32. Anyoha, *supra* note 29.

33. *See id.* ("[i]ronically, in the absence of government funding and public hype, AI thrived.").

34. *See id.* (recalling how grand master Gary Kasparov lost to an AI chess playing computer program); *see also* Danielle Muoio, *Why Go is So Much Harder for AI to Beat Than Chess*, BUSINESS INSIDER (Mar. 10, 2016, 12:32 PM CST), https://www.businessinsider.com/why-google-ai-game-go-is-harder-than-chess-2016-3 [https://perma.cc/2VRA-2TCM] (explaining that the "sophisticated pattern recognition" shown in playing a game of Go has potential driverless cars).

35. *See generally* Copeland, *supra* note 30 (describing the development of AI and the current use of AI in the twenty first century).

36. *See* Amanda Napitu, *150+ Artificial Intelligence Statistics You Need to Know in 2024—Who is Using It & How?*, TECHNOPEDIA (Jan. 17, 2024), https://www.techopedia.com/artificial-intelligence-statistics [https://perma.cc/PB4S-LW37] (discussing the growth and projections for artificial intelligence in various sectors); Mark Webster, *149 AI Statistics: The Present & Future of AI at Your Fingertips*, AUTHORITY HACKER, https://www.authorityhacker.com/ai-statistics/ [https://perma.cc/JF38-C9L2] (last updated Oct. 25, 2024) (listing various statistics of increased usage and attention on AI in 2024).

37. *What is Generative AI?*, MCKINSEY & CO. (Apr. 2, 2023), https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai [https://perma.cc/N2VD-43WV]; *see also* Ben Dickson, *How 2022 Became the Year of Generative AI*, VENTUREBEAT (Nov. 11, 2022, 9:18 AM), https://venturebeat.com/ai/how-2022-became-the-year-of-generative-ai/ [https://perma.cc/9NYB-3GEG]

patterns from existing content but also can create new content based on those patterns.[38] This key distinguishing quality and the development process necessary to leverage this quality have generated much of the legal issues surrounding generative AI.[39]

B. Prompt: How do I Build and Use a Generative AI Model?

A generative AI model operates by creating new content based on patterns that it has identified in existing content and based on user input.[40] Before generating new content, a generative AI model must first be developed to identify those patterns.[41] Aside from writing the underlying code of the generative AI model, software developers must create a dataset, or a collection of existing content, and then "train the model," or feed the dataset to the model for analysis.[42]

In creating a dataset, massive amounts of existing content are collected.[43] In optimizing a generative AI model that can create a variety of new content and respond to a variety of user inputs, the model must be trained on a dataset that is large, diverse, and relevant.[44] Relevancy depends on the purpose of the generative AI model.[45] For example, DALL·E 2, a generative AI model designed to generate images, is trained on a dataset containing "hundreds of millions" of captioned images.[46] In another example, ChatGPT, a model designed to generate text on a wide range of topics, is trained on multiple datasets of over hundreds of billions of website data and books.[47]

Software developers can also take extra steps to filter the existing content in the dataset that could result in biased, over- or under-generalized, or unsavory content generation.[48] For example, the DALL·E 2 software developers decided that images of graphic violence and sexually explicit content did not contribute to their intended purpose for the model, so they removed those images from their

(attributing the recent productization of generative AI to the technological developments, starting in 2014, that resulted in "better algorithms, larger models and bigger datasets").

38. MCKINSEY & CO., *supra* note 37; *AI, ML, and Generative AI: Key Differences and Applications*, RACKSPACE TECH. (June 14, 2023), https://www.rackspace.com/blog/distinctions-ai-ml-generative-ai [https://perma.cc/N6MC-VCBB].

39. *See infra* Section II.C (discussing background of the legal landscape of generative AI).

40. Akash Takyar, *How to Build a Generative AI Solution: A Step-By-Step Guide*, LEEWAYHERTZ, https://www.leewayhertz.com/how-to-build-a-generative-ai-solution/ [https://perma.cc/RS22-XAS9] (last visited Oct. 12, 2023).

41. *Id.*

42. *See id.* (explaining that building a generative AI software model requires designing and training the generative AI software model).

43. *Id.*

44. *Id.*

45. *Id.*

46. *DALL·E 2 Pre-Training Mitigations*, OPENAI (June 28, 2022), https://openai.com/research/dall-e-2-pre-training-mitigations [https://perma.cc/3FTG-XS55].

47. Dennis Layton, *ChatGPT — Show me the Data Sources*, MEDIUM (Jan. 30, 2023), https://medium.com/@dlaytonj2/chatgpt-show-me-the-data-sources-11e9433d57e8 [https://perma.cc/4DRX-RYXS].

48. Andrii Bas, *How to Build Generative AI Solutions from Scratch*, UPTECH, https://www.uptech.team/blog/how-to-build-generative-ai [https://perma.cc/U4FM-YJUD] (last updated May 31, 2024).

dataset.[49] After that initial filter, the DALL·E 2 software developers discovered that the model had a significant bias toward creating images of men overall, necessitating further filtering of the dataset to reduce the gender bias.[50] Because curating a dataset is time and resource intensive, some software developers will start with existing datasets (e.g. ChatGPT) instead of creating a completely new dataset (e.g. DALL·E 2).[51] After finalizing a collection of existing content, additional preprocessing of the datasets can be undertaken to optimize the quality and relevance before training.[52]

In training the model, the software developer feeds the finalized dataset to the generative AI model, allowing the model to analyze the data.[53] Training involves the model extracting patterns, similarities, and differences from the existing content in the dataset.[54] After training, the software developer tests the model, simulating prompts that users would input, and evaluates the quality of content generation.[55] Any step throughout the reiterative development process can be redone until the software developer deems the generative AI model ready for use.[56]

In employing a generative AI model, a user inputs a prompt into a generative AI model, and the model utilizes varying amounts of pattern and predictive analysis to create new content.[57] Generative AI models have been adapted for a variety of uses such as image editing, customer service chatbots, and text generation.[58] The design of the model determines the form of the prompts that a user can input.[59] While prompts can take any form, many generative AI models require text prompts from users.[60] One of the more prominent types of generative AI models, AI image generators receive textual prompts from users and generates images.[61]

As generative AI gets more sophisticated, deepfakes, or artificial images or videos, will become harder to distinguish from real images or videos.[62] Deepfakes have been used to create fake videos of political figures so much that

49. *OPENAI*, *supra* note 46.

50. *Id.*

51. *See* Layton, *supra* note 47 (listing the multiple preexisting datasets that OpenAI, the software developer, used to curate the dataset for training ChatGPT); Kevin Pocock, *What Was Dall-E 2 Trained On?*, PC GUIDE, https://www.pcguide.com/apps/what-was-dall-e-2-trained-on/ [https://perma.cc/W2R6-ML9M] (May 10, 2023).

52. Takyar, *supra* note 40.

53. *Id.*

54. *Id.*

55. *Id.*

56. *Id.*

57. *Id.*; *Gartner Experts Answer the Top Generative AI Questions for Your Enterprise*, GARTNER, https://www.gartner.com/en/topics/generative-ai [https://perma.cc/3273-9FPE] (last visited Oct. 15, 2024).

58. Bas, *supra* note 48.

59. GARTNER, *supra* note 57.

60. *See id.* (listing many applications of generative AI models that receive text prompts); *see also Explore Beyond the Canvas with Generative Expand*, ADOBE, https://helpx.adobe.com/photoshop/using/generative-expand.html [https://perma.cc/E86G-VGRE] (last updated Oct. 14, 2024) ("Expand the dimensions of your image and generate content using text prompts with Generative Expand in Photoshop on the desktop.").

61. ADOBE, *supra* note 60.

62. *What the Heck Is a Deepfake?*, UNIV. OF VA., https://security.virginia.edu/deepfakes [https://perma.cc/A9GV-GMBR] (last visited Aug. 7, 2024).

the Federal Election Commission and multiple states have begun to create statutory and regulatory schemes to combat deepfakes used for political discourse.[63] In the case with Taylor Swift, a public figure, no analogous statute or regulation exist.[64] As will be discussed, as deepfakes of celebrities or other public figures become more prevalent and realistic, the public will unknowingly spread these deepfakes, causing further reputational harm.[65]

Although software developers invest a substantial amount of time and resources in developing a generative AI model, many operational and development issues arise regardless, resulting in legal and ethical issues.[66] The operation and development processes that have caused legal issues are the curating of massive amounts of content for the dataset, storing and processing of sensitive information in datasets and prompts, and generating false or explicit content.[67] In the next section, this Note examines the government and legal profession responses in addressing these issues.

C. The Legal Landscape of Generative AI

Although AI currently has a stronghold on the attentions of the government and the public, the growing concerns did not mature into formal actions until OpenAI released ChatGPT, exposing various regulatory issues.[68] The next two subsections summarize the responses from the government and the legal profession respectively, highlighting the lagging and patchwork-like efforts to address the variety areas of law and regulation affected by generative AI models.[69]

---

63. Matthew Shapanka & Samuel Klein, *As States Lead Efforts to Address Deepfakes in Political Ads, Federal Lawmakers Seek Nationwide Policies*, COVINGTON (Apr. 18, 2024), https://www.insidepoliticallaw.com/2024/04/18/as-states-lead-efforts-to-address-deepfakes-in-political-ads-federal-lawmakers-seek-nationwide-policies/ [https://perma.cc/6NPU-DALR].

64. *See* Mulvihill, *supra* note 25 (addressing potential legals issues involving existing law but not new law).

65. *See id.* (explaining that deepfakes are "appearing online more often, in several forms").

66. *See* Brain John Aboze, *Risks of Large Language Models: A Comprehensive Guide*, DEEPCHECKS (Aug. 7, 2023), https://deepchecks.com/risks-of-large-language-models/ [https://perma.cc/3BTB-6BVR] (listing various risks with generative AI model use such as prompt misinterpretation, data privacy and security concerns, misinformation, intellectual property infringement, and hallucinations); *see also infra* Section II.C (discussing background of the legal landscape of generative AI).

67. Aboze, *supra* note 66.

68. *See* Nicole Cunningham, *Top 5 Stories of the Week: Generative AI Dominates the News Again*, VENTUREBEAT (Mar. 18, 2023, 5:17 AM), https://venturebeat.com/ai/top-5-stories-of-the-week-generative-ai-dominates-the-news-again/ [https://perma.cc/T659-29HJ] (describing that five generative AI news stories dominated the news cycle); *see also* Müge Fazlioglu, *US Federal AI Governance: Laws, Policies and Strategies*, IAPP (Nov. 2023), https://iapp.org/resources/article/us-federal-ai-governance/ [https://perma.cc/G924-PKQV] (reporting the early responses of various U.S. governmental bodies in 2020); *see generally* Annabelle Nyst, *History of ChatGPT: A Timeline of the Meteoric Rise of Generative AI Chatbots*, SEARCH ENGINE J., https://www.searchenginejournal.com/history-of-chatgpt-timeline/488370/ [https://perma.cc/4D5N-XLXV] (last updated Oct. 9, 2024) (reporting OpenAI's 2019 publication of their research on ChatGPT's generative AI model and their 2020 release of ChatGPT).

69. *See infra* Subsections II.C.1 and II.C.2 (discussing background of government's response, legal profession's response, and courts' response to generative AI).

1. Government Response

Despite the emerging technological potentiality of generative AI models since 2014, the 2019 release of the first consumer friendly generative AI model, ChatGPT, revealed the ill preparedness of not only the U.S. government but also the world governments.[70] Before 2019, most AI regulations and laws centered around self-driving vehicles and national security.[71] Because of the limited scope of AI at the time, the government policies focused on funding and encouraging AI research to "address broad social problems."[72] After 2019, the world governments, recognizing the rapid expansion of AI into unaddressed applications, began to bolster their AI policies to include risk management.[73] Not unlike the rest of the world, the U.S. federal government began to consider the AI issues of data privacy, intellectual property, discrimination, and misinformation.[74]

Due to the enormous amounts (often in the millions or billions) of content that get curated in training datasets or inputted as prompts, generative AI models expose some personally sensitive information or IP protected works to unauthorized use.[75] In response, U.S. President Biden's Administration highlighted commitment to ensuring individual data privacy and content bias amounting to discrimination.[76] The United States Patent and Trademark Office (USPTO) and the United States Copyright Office (USCO) have questioned whether current U.S. policy supports the intellectual property rights of purely AI-generated works, refusing to recognize AI authorship or inventorship until further notice.[77] While the Congressional Research Service has affirmed the

70. *See supra* note 68 and accompanying text (explaining that concerns about AI did not mature into action until the release of ChatGPT); *see also* Dickson, *supra* note 37 (attributing the recent productization of generative AI to the technological developments, starting in 2014, that resulted in "better algorithms, larger models and bigger datasets").

71. Fazlioglu, *supra* note 68.

72. *Id.*

73. *See id.* (stating examples such as: Singapore's 2019 Model AI Governance Framework that adds "practical examples of organizational-level AI governance" into their existing National AI Strategy; China's 2023 Administrative Measures for Generative Artificial Intelligence Services that aims to prevent discriminating, spreading of misinformation, and infringing of intellectual property rights by generative AI content; Canada's 2023 proposed Artificial Intelligence and Data Act that updates the existing information privacy laws to include AI concerns; U.K.'s 2023 white paper release that provides a framework to balance AI innovation with its risk; and the European Parliament's 2023 potential adoption of the Artificial Intelligence Act that would ban or limit "high-risk applications of AI.").

74. *Id.*

75. *See* Aboze, *supra* note 66 (discussing AI training); Dylan Walsh, *The Legal Issues Presented by Generative AI*, MIT SLOAN (Aug. 28, 2023), https://mitsloan.mit.edu/ideas-made-to-matter/legal-issues-presented-generative-ai [https://perma.cc/6R8L-KCQJ] (describing lawsuits where software developers allegedly violated their own privacy policies or infringed on copyrighted material).

76. *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*, WHITE HOUSE OFF. OF SCI. & TECH. POLICY, https://www.whitehouse.gov/ostp/ai-bill-of-rights/ [https://perma.cc/S6CG-EFS4] (last visited Oct. 14, 2024).

77. *See Copyright in the Age of Artificial Intelligence*, U.S. COPYRIGHT OFF., https://www.copyright.gov/events/artificial-intelligence/ [https://perma.cc/6FX6-SMDB] (last visited Oct. 14, 2024) (reporting a symposium held by the United States Copyright Office and the World Intellectual Property Organization to discuss using AI in creating potentially copyrightable works); Kathi Vidal, *With Artificial Intelligence Speeding the Innovation Process, What Does That Mean for Invention and a Properly Balanced*

USCO's stance, some legislators have showed some openness to changing U.S. patent law to consider AI-assisted inventorship.[78] While the executive and legislative branches have prioritized data privacy, intellectual property rights, and preventing discriminatory practices, little has been said about generative AI content that amounts to tortious conduct.[79]

Due to recent cases of deepfakes, two federal actions specifically targeting deepfakes have been announced.[80] The Federal Communications Commission has adopted a rule banning AI-generated robocalls.[81] A federal legislator introduced The Disrupt Explicit Forged Images and Non-Consensual Edits, or DEFIANCE Act to target those who create or distribute sexually explicit deepfakes.[82] Additionally, ten states have enacted laws to target sexually explicit deepfakes or politically motivated deepfakes.[83]

2. Legal Profession and Court Response

As the government response slowly ramps up, lawsuits over generative AI models have rapidly materialized, forcing courts to consider these issues with a patchwork of laws.[84] Many of the issues spotlighted and advanced in court deal with intellectual property rights and data privacy.[85] In regard to the intellectual property rights cases, judges easily rejected granting protection for AI-assisted or created works, affirming the USPTO's and USCO's regulatory decisions.[86]

Patent System?, U.S. PAT. & TRADE OFF. (Apr. 18, 2023), https://www.uspto.gov/blog/director/entry/with-artificial-intelligence-speeding-the [perma.cc/MPP4-XUE3] (announcing the USPTO's request for public comments about inventorship of AI-assisted invention); James Love, *We Need Smart Intellectual Property Laws for Artificial Intelligence*, SCI. AM. (Aug. 7, 2023), https://www.scientificamerican.com/article/we-need-smart-intellectual-property-laws-for-artificial-intelligence/ [https://perma.cc/Z8UE-YBSD] (describing the intent of legislatures to address AI-assisted inventorship of intellectual property).

78. Christopher T. Zirpoli, GENERATIVE A.I. AND COPYRIGHT L. 3 (2023); Gopal Ratnam, *Congress Ponders Whether AI Should Have the Power of the Patent*, ROLL CALL (June 13, 2023), https://rollcall.com/2023/06/13/congress-ponders-whether-ai-should-have-the-power-of-the-patent/ [https://perma.cc/P762-FSZZ].

79. *Supra* notes 73–74 and accompanying text; *cf.* Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Oct. 30, 2023) (mentioning "disinformation" as a potential social harm caused by AI but not traditional torts like defamation); *US FTC Opens Investigation into OpenAI over Misleading Statements*, REUTERS (July 13, 2023, 5:05 PM CDT), https://www.reuters.com/technology/us-ftc-opens-investigation-into-openai-washington-post-2023-07-13/ [https://perma.cc/V2G5-M764] (reporting the first request for information about ChatGPT generating statements that cause reputational harm to consumers).

80. *Infra* notes 81–82.

81. Max Burns, *Why Deepfakes Like the Biden Robocall are a Threat — Even to Those Who Don't Fall for It*, MSNBC (Feb. 26, 2024, 5:39 PM), https://www.msnbc.com/opinion/msnbc-opinion/fake-biden-robocall-ai-elections-rcna140570 [https://perma.cc/WL88-RE55].

82. Solcyré Burga, *How a New Bill Could Protect Against Deepfakes*, TIME (Jan. 31, 2024 4:34 PM), https://time.com/6590711/deepfake-protection-federal-bill/ [https://perma.cc/4ZYU-PU4S].

83. Mulvihill, *supra* note 25.

84. *See* Elizabeth Vandesteeg et al., *The Crucial Role of U.S. Privacy Laws on Unregulated Generative AI*, LEVENFELD PEARLSTEIN (Sept. 27, 2023), https://www.lplegal.com/content/old-laws-new-tech-part-1-u-s-privacy-laws-unregulated-generative-ai/ [https://perma.cc/564T-4WHZ] ("That is because, although there are, as of yet, no AI regulatory laws, there are privacy laws and copyright laws that govern how data can be used."); Walsh, supra note 75 (detailing lawsuits over AI).

85. *See* Walsh, *supra* note 75 (detailing types of recent lawsuits).

86. *See, e.g.*, Thaler v. Perlmutter, 687 F. Supp. 3d 140, 150 (D.D.C. 2023) (affirming the denial of copyright registration for AI-generated work); Douglas R. Nemec & Laura M. Rann, *AI and Patent Law:*

These court decisions seem to be legally sound and final until new legislation or agency rules are introduced.[87] As for privacy lawsuits, many have been filed and, with overlapping state and federal privacy laws, are not as quickly resolved.[88]

A recent filing by Mark Walters, in July 2023, has raised a new question that does not involve intellectual property or data privacy.[89] Walters filed a defamation lawsuit against OpenAI in Georgia State court, claiming ChatGPT generated a complete fabrication about him that was libelous and harmful to his reputation.[90] A recently identified problem with generative AI models are AI hallucinations—AI-generated "false, misleading or illogical information, but presents it as if it were a fact."[91] If a user were to post or repeat an AI hallucination, the factual circumstances more closely resemble traditional tort cases, opening the user to liability.[92] Walters's case introduces a uniquely generative AI models situation in which an algorithm made false statements to a user about Walters.[93] In attempting to satisfy the elements of traditional torts, Walters must overcome some obstacles: (1) The generative AI models most likely lack the knowledge, intent, or recklessness to conduct tortious acts; (2) the software developer might not be responsible for preventing hallucinations; and (3) the hallucination was only presented to one person.[94]

The next part of the note analyzes whether a software developer would be liable in a tort suit for their generative AI model hallucinations that result in a deepfake of a public figure.

Balancing Innovation and Inventorship, SKADDEN (Apr. 2023), https://www.skadden.com/insights/publications/2023/04/quarterly-insights/ai-and-patent-law [https://perma.cc/AF56-PKSJ] (discussing similar results in another case).

87. *See supra* note 77 and accompanying text (reporting that while USPTO and USCO have requested public opinion, both offices have decided to grant IP protection under current policy); Zachary Small, *As Fight Over A.I. Artwork Unfolds, Judge Rejects Copyright Claim*, N.Y TIMES (Aug. 21, 2023), https://www.nytimes.com/2023/08/21/arts/design/copyright-ai-artwork.html [https://perma.cc/V9TS-G8SW] (commenting on the Thaler decision, the USCO "believes the court reached the correct result").

88. *See* Walsh, *supra* note 75 ("There are also 'way too many cases to count' centered on privacy concerns . . . ."); *see, e.g.*, Haim Ravia & Dotan Hammer, *Google and OpenAI were Hit with Lawsuits on the Use of AI*, PEARL COHEN (Jul. 30, 2023), https://www.pearlcohen.com/google-and-openai-were-hit-with-lawsuits-on-the-use-of-ai/ [https://perma.cc/426W-SSMQ] (reporting multiple class action lawsuits over personal data scraping by Alphabet, Microsoft, and OpenAI).

89. Rebecca Cahill, *OpenAI Defamation Lawsuit: The First of Its Kind*, SYRACUSE L. REV. (June 22, 20023), https://lawreview.syr.edu/openai-defamation-lawsuit-the-first-of-its-kind/ [https://perma.cc/YM8H-997D].

90. *Id.*

91. Ellen Glover, *What are AI Hallucinations?*, BUILT IN (Oct. 2, 2023), https://builtin.com/artificial-intelligence/ai-hallucination [https://perma.cc/W7UF-QT3V].

92. *See* Clay Calvert, *Defamation Law and Generative AI: Who Bears Responsibility for Falsities?*, AM. ENTER. INST. (Aug. 22, 2023), https://www.aei.org/technology-and-innovation/defamation-law-and-generative-ai-who-bears-responsibility-for-falsities/ [https://perma.cc/8UHW-XCA6] ("[A]nyone who uses generative AI to produce information about a person and then conveys it to someone else may be legally responsible if it is false and defamatory.").

93. *See* Cahill, *supra* note 89 (discussing the facts behind Walters' suit).

94. *See id.* (assessing Walters' case).

III. ANALYSIS

Although AI -generated falsities have become prevalent enough to garner the moniker "hallucinations," there has been a dearth of attention and response from the White House and congress.[95] This part will analyze the available private causes of action when a generative AI model generates a false image amounting to a deepfake of a person. First, though, because many generative AI models are hosted online, generative AI model software developers or those who deploy them could be granted immunity under 47 U.S.C. § 230 (Protection for private blocking and screening of offensive material) ("Section 230").[96]

A. Section 230

With the goal of balancing first amendment protections on the Internet and online platform's freedom to operate, Congress passed Section 230.[97] This statue states, "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[98] This specific section has protected many social media companies from liability for defamatory comments users post on their platforms.[99] While legislators have begun to scrutinize the continued need for Section 230,[100] legal scholars disagree on whether the current iteration of Section 230 grants immunity to generative AI model software developers or web hosts.[101] A group of scholars have interestingly pointed out that Section 230 creates a paradox for generative AI model software developers: that efforts to create hallucination-proof datasets would likely result in the software developer losing Section 230 immunity.[102] This will be discussed later in this section.

The portion of Section 230 at issue here states:

> No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.[103]

---

95. *See* Glover, *supra* note 91 (explaining AI hallucinations); *see supra* note 79 and accompanying text (detailing the policy priorities on intellectual property and data privacy, more so than tortious conduct).

96. *See, e.g.*, *ChatGPT 3.5*, OPENAI, https://chat.openai.com/ (last visited Oct. 14, 2024) (online AI chatbox); Protection for Private Blocking and Screening of Offensive Material, 47 U.S.C. § 230(c)(1).

97. 47 U.S.C. § 230(b).

98. *Id.* at (c)(1).

99. *Cf. Section 230*, ELEC. FRONTIER FOUND., https://www.eff.org/issues/cda230 [https://perma.cc/SVK4-TZNV] (last visited Oct. 14, 2024) (providing examples of websites defended by 47 U.S.C. § 230(c)(1)).

100. Bryan Mena & Duncan Agnew, *Republicans and Democrats Both Want to Repeal Part of a Digital Content Law, but Experts Say That Will be Extremely Tough*, TEX. TRIBUNE (Jan. 21, 2021), https://www.texastribune.org/2021/01/21/section-230-internet-social-media/ [https://perma.cc/QQ5V-XC2G].

101. *Compare* Matt Perault, *Section 230 Won't Protect ChatGPT*, 3 J. FREE SPEECH L. 363, 364 (2023) (arguing that Section 230 does provide immunity); Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. FREE SPEECH L. 489, 495 (2023) (also arguing that Section 230 does not provide immunity) *with.* Peter Henderson et al., *Where's the Liability in Harmful AI Speech*, 3 J. FREE SPEECH L. 589, 619–626 (2023) (showing less certainty about Section 230 immunity).

102. Henderson et al., *supra* note 101, at 625.

103. 47 U.S.C.§ 230(c)(1).

The determinative question here is whether software developers of generative AI model are only "providers . . . of an interactive computer service" (granting Section 230 immunity) or also an "information content provider" (falling out of Section 230 immunity).[104]

Section 230 defines "interactive computer service" as "any information service, system, or access software provider that provides or enables computer access by multiple users to a computer server, including specifically a service or system that provides access to the Internet and such systems operated or services offered by libraries or educational institutions."[105] Generative AI models unquestionably meet this definition as an interactive computer service and, the software developers, who are often the web host, fall under this definition.[106]

Section 230 defines "information content provider" as "any person or entity that is responsible, in whole or in part, for the creation or development of information provided through the Internet or any other interactive computer service."[107] Whether generative AI qualifies as an information content provider determines liability.[108] In particular, the most dispositive and debated question is whether a particular generative AI model merely republishes existing content or is creating new content as to constitute the generative AI model's speech.[109]

Although generative AI models are "creating new content," the AI models are essentially using varying amounts of predictive pattern recognition and regurgitation for content generation.[110] This type of content generation is thus derivative of content and information provided by others.[111] Although the derivative nature of AI-generated content bolsters arguments for Section 230 immunity, the existing case law does not provide such protections for humans.[112] For example, if an individual were to repost or republish another person's online text that was already posted, then that individual would be immunized from liability by Section 230.[113] On the other hand, if an individual were to copy each word of another person's online text and created their own post, then that individual would be open to liability.[114] The threshold question now being how much modification, addition, or editing arises to material contribution, resulting in status as an information content provider.[115]

---

104. Volokh, *supra* note 101, at 494–95.

105. 47 U.S.C. § 230(f)(2).

106. *See* Avi Weitzman & Jackson Herndon, *Generative AI: The Next Frontier for Section 230 of the Communications Decency Act*, LAW.COM (June 26, 2023, 10:00AM), https://assets.ctfassets.net/t0ydv1wnf2mi/4LqDaZLG1NvNeWhAoGLxRq/f0f003214772a4c67369de054a9870b8/NYLJ706202344999Paul.pdf [https://perma.cc/2XPV-BZM5] (stating that AI will likely establish themselves as interactive computer services).

107. *Id*.

108. *Id*.

109. *See, e.g.*, *id.* (concluding that the current state of U.S. law and policy does not provide an answer to whether generative AI creates content); s*ee* Perault, *supra* note 101, at 364 (concluding that generative AI does develop content as to preclude Section 230 immunity).

110. *See* Weitzman & Herndon, *supra* note 106 (explaining that the output is based on training data containing millions of examples of media and the computer learns to create the desired output).

111. Volokh, *supra* note 101, at 496.

112. *Id*.

113. *Id*.

114. *Id*.

115. *Id.* at 498.

Although the Supreme Court has not ruled on a list of requirements to classify an information content provider, the appellate courts provide some consistency in the factors considered.[116]

One factor that the courts have considered is whether the original source of the information intended the information to be private.[117] In *Batzel v. Smith*, a distributer of an electronic newsletter published the plaintiff's private email in the newsletter.[118] The appellate court remanded, holding that a distributor is immunized under Section 230 if an original provider furnished the email "under circumstances in which a reasonable person in the position of the [distributor] would conclude that the information was provided for publication on the Internet or other interactive computer service."[119] Applying this ruling, a generative AI model would most likely not be protected by Section 230 if it generated and cited an exact copy of defamatory content that was intended to be private.[120] As mentioned before, the massive data scraping for training datasets could result in private information being used without consent.[121]

In *Parker v. Google*, an author sued Google for archiving and displaying defamatory messages about him that were posted online on a third-party website.[122] Additionally, the author claimed the display by Google included the defamatory messages creating "an unauthorized biography of Plaintiff".[123] The appellate court held that Google was a provider clearly intended for protection under Section 230.[124] It would follow that generative AI models escape liability if making minor additions to the presentation of the information so long as the content of the information is not affected.[125]

While multiple courts have reiterated the providers' freedom to publish, that freedom does not completely immunize all publishing decisions by providers.[126] In *Roca Labs v. Consumer Opinion*, the court cited another case that trimming content for length does not preclude a provider from Section 230 immunity, so long as the edits are "unrelated to the illegality."[127] In *Lewis v. Google*, the court referenced that section 230 shields providers from liability of "all publication decisions . . . with respect to content generated entirely by third parties."[128] In *Pennie v. Twitter*, the court pointed out that Section 230 provides flexibility to providers to edit and filter content without becoming liable for

116. Batzel v. Smith, 333 F.3d 1018, 1034 (9th Cir. 2003); Roca Labs, Inc. v. Consumer Op. Corp., 140 F. Supp. 3d 1311, 1319–1321 (M.D. Fla. 2015).

117. *Smith*, 333 F.3d at 1034.

118. *Id.* at 1022.

119. *Id.* at 1034.

120. Volokh, *supra* note 101, at 496.

121. Ravia & Hammer, *supra* note 88.

122. Parker v. Google, Inc., 422 F. Supp. 2d 492 (E.D. Pa. 2006), *aff'd*, 242 Fed. Appx. 833 (3d Cir. 2007).

123. *Id.* at 500.

124. Parker v. Google, 242 Fed. Appx. 833, 838 (3d Cir. 2007).

125. *See id.* (holding that Google's archiving and caching of the data is not publication).

126. *See* Volokh, *supra* note 101, at 495 (explaining that Congress made the choice not to impose liability on companies that serve as intermediaries but defendants who "materially contribut[e]" are not immunized).

127. Roca Labs, Inc. v. Consumer Op. Corp., 140 F. Supp. 3d 1311, 1320 (M.D. Fla. 2015) (citing Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157, 1169 (9th Cir. 2008)).

128. Lewis v. Google LLC, 461 F. Supp. 3d 938, 954 (N.D. Cal. 2020) (citing Barnes v. Yahoo!, Inc., 570 F.3d 1096, 1105 (9th Cir. 2009)).

defamatory content that is not edited or filtered.[129] These cases seem to suggest that generative AI model software developers would not be protected by Section 230, but the volitive editing in those cases differ from generative AI models' algorithmic editing.[130] In *Fair Housing Council v. Roommates.com*, the court discusses what "development" is under Section 230.[131] The court decides that a search engine, being a "neutral tool," cannot contribute to any unlawfulness.[132] The court also states that a provider "who edits in a manner that contributes to the alleged illegality—such as by removing the word 'not' from a user's message . . . in order to transform an innocent message into a libelous one—is directly involved in [unlawfulness]."[133] In the case of a generative AI model, the model acts as both a "neutral tool" and the entity that could "edit in a manner that contributes to the unlawfulness."[134]

As shown, the evolving law around Section 230 involves defining whether the editing conducted during content generation or publishing amounts to content development as an information content provider.[135] An option to avoid these hallucinations, and thus liability of defamatory statements, would be curating the dataset more scrupulously or creating the content for the dataset from scratch.[136] While the mechanisms of Section 230 normally encourage website providers to refrain from developing content to maintain immunity, generative AI models create a conundrum where the simple option of avoiding content creation does not exist.[137] In the situation where software developers train generative AI models on datasets of third-party content, they do not actively engage in content development but increase the likelihood of hallucinations, potentially resulting in content development.[138] In the inverse situation where, in attempting to decrease hallucinations, software developers train generative AI models on datasets of content created by developers or implement stricter editing and filtering algorithms, they actively engage in content development.[139] This catch-22 leaves software developers with two options: (1) risking the possibility of a court finding that they are unprotected as information providers due to their generative AI models' hallucinations or (2) making the decision to be information providers by actively engaging in

129. Pennie v. Twitter, Inc., 281 F. Supp. 3d 874, 890 (N.D. Cal. 2017) (citing *Fair Hous. Council of San Fernando Valley*, 521 F.3d at 1174).

130. *See id.* (holding that there is flexibility for some editing); *Roca Labs*, 140 F. Supp. 3d at 1320 (holding that trimming for length does not jeopardize immunity); Weitzman & Herndon, *supra* note 106 (describing the use of algorithms to generate content).

131. *Fair Hous. Council of San Fernando Valley*, 521 F.3d at 1169.

132. *Id.* at 1171.

133. *Id.*

134. *See id.* at 1169 ("[A] website operator who edits in a manner that contributes to the alleged illegality . . . is . . . not immune."); GARTNER, *supra* note 57 (explaining that generative AI creates content in response to a user's request and creates new content that reflects the training data but does not repeat it).

135. *Infra* Part III.A.

136. Henderson et al., *supra* note 101, at 617.

137. *See id.* at 622 (explaining that the immunity applies to third-party content hosted, shared, or linked to, but "AI is itself generating new content").

138. *Infra* Part III.A.

139. Henderson et al., *supra* note 101, at 625–26.

creating, editing, and filtering content.[140] Making the less burdensome decision, some software developers have chosen option one, attaching a disclaimer that the content generated by generative AI models could be inaccurate.[141] Whether a disclaimer legally protects software developers is beyond the scope of this Note, but the effect a disclaimer has on the users' states of mind will be explored.[142]

Even if a generative AI model results in the loss of Section 230 immunity, a plaintiff must still prove the tort claim to hold a defendant liable.[143] The next sections examine the liability of software developers and website hosts of generative AI models against various tort claims of falsehoods.[144]

B. Defamation

When an individual publishes AI-generated defamatory text, the factual circumstances more closely resemble cases that traditional defamation jurisprudence can answer.[145] The factual circumstance outlined in Walter's suit puts forward questions of defamation liability that is less clear.[146] This section explores current defamation jurisprudence and its applicability to determining the liability of generative AI model software developers.

While many legal definitions of defamation exist and vary depending on state, the general prima facie of defamation involves: "1) a false statement purporting to be fact; 2) publication or communication of that statement to a third person; 3) fault amounting to at least negligence; and 4) damages, or some harm caused to the reputation of the person or entity who is the subject of the statement."[147] When public figures and officials are the subject of a defamatory statement, the fault standard rises to "actual malice."[148] Actual malice means a defendant communicated the defamatory statement "with knowledge that it was false or with reckless disregard of whether it was false or not."[149] Additionally, some cases recognize certain statements to be defamation *per se*; a statement

140. *See id.* at 622, 625–26 (explaining that AI is generating new content but attempting to decrease hallucinations can increase their potential liability).

141. *See, e.g.*, *ChatGPT 3.5*, OPENAI, https://chat.openai.com/ (last visited Oct. 13, 2024) (displaying "ChatGPT can make mistakes. Check important info." at the bottom of the screen").

142. *Infra* Section III.B. *see also* Henderson et al., *supra* note 101, at 637–39 (questioning whether ChatGPT's disclaimer avoids liability).

143. *See, e.g.*, Anthony v. Yahoo! Inc., 421 F. Supp. 2d 1257, 1263 (N.D. Cal. 2006) (holding that the defendant is not immunized by Section 230 and that the plaintiff may still proceed with their tort claims in the case).

144. *Infra* Sections III.B, III.C.

145. *See* Calvert, *supra* note 92 (analogizing the situation to when "a print newspaper cannot escape liability for publishing a defamatory comment just because it accurately attributes the comment to a source.").

146. *See Walters v. OpenAI*, KNOWING MACHS. (Jan. 16, 2024), https://knowingmachines.org/knowing-legal-machines/legal-explainer/cases/walters-v-openai [https://perma.cc/DTD4-6G6U] ("[T]he Georgia Superior Court preliminarily rejected OpenAI's arguments, denying the company's motion to dismiss and allowing the case to proceed.").

147. Wex Definitions Team, *Defamation*, CORNELL L. SCH.: LEGAL INFO. INST., https://www.law.cornell.edu/wex/defamation [https://perma.cc/45BC-WPS5] (last updated June 2023); *see also Is There a Federal Defamation Statute of Limitations?*, MULLEN L. FIRM, https://mullenlawfirm.com/federal-defamation-statute-of-limitations/ [https://perma.cc/F79S-NEND] (last visited Oct. 12, 2024).

148. Wex Definitions Team, *supra* note 147.

149. *Id.* (citing N.Y. Times Co. v. Sullivan, 376 U.S. 254, 279–280 (1964)).

that "is presumed to be damaging to a person's reputation without any additional proof of harm."[150] "These statements include those that impute a criminal offense, suggest someone is infected with a contagious disease, imply a lack of ability or integrity in one's trade or profession, or make false allegations about a person's sexual conduct."[151]

1. Purported Fact

Defamation requires the false statement to be asserted as a fact.[152] The First Amendment demands protection to limited categories of speech despite their harm to a person's reputation.[153] An informative holding is found in *Hustler Mag. v. Falwell*. There, the Supreme Court found that the First Amendment protects parody from defamation liability because parodic statements "could not reasonably have been interpreted as stating actual facts."[154] In its reasoning, the court focused on whether a person would reasonably perceive a statement to be asserted as fact.[155] This standard is further bolstered in *Masson v. New Yorker Mag.*, where the Supreme Court considered whether accompanying a statement with quotation marks definitively asserts the statement as fact.[156] There, the court acknowledged that quotations used in instances such as docudramas or historical fiction differ from quotations used in journalistic writing.[157] Reinforcing this standard, some state courts have affirmed that the central question is whether a statement would reasonably appear to assert a fact.[158]

Some could argue that generative AI models' generated content should not be seen as assertions of fact.[159] Unsurprisingly, in Walters's suit, OpenAI, motioning to dismiss, argued that ChatGPT-generated statements cannot reasonably be perceived as facts because of the disclaimers and "probabilistic" nature of generative AI models.[160] Skeptical by the initial arguments, the Georgia Superior Court preliminarily denied OpenAI's motion.[161] Agreeing with the Georgia Superior Court's skepticism, scholars have argued that disclaimers similar to ChatGPT's will not significantly change users' perception

150. Antolonplos & Assocs., *Defamation Per Se: Understanding the Legal Concept and Its Implications*, (Feb. 6, 2023), https://www.antonlegal.com/blog/defamation-per-se-understanding-the-legal-concept-and-its-implications/ [https://perma.cc/N5GF-QYD7].

151. *Id.*

152. MULLEN L. FIRM, *supra* note 147; Wex Definitions Team, *supra* note 147.

153. *See, e.g.*, Milkovich v. Lorain J. Co., 497 U.S. 1, 2 (1990) (explaining that opinions "having no provably false factual connotation" are protected even if harmful); *see, e.g.*, Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 50 (1988) (holding that parody is protected because it "could not reasonably have been interpreted as stating actual facts").

154. *Falwell*, 485 U.S. at 46, 50 (1988).

155. *Id.* at 52.

156. *See* Masson v. New Yorker Mag., Inc., 501 U.S. 496, 512 (1991) ("Punctuation marks, like words, have many uses. Writers often use quotation marks, yet no reasonable reader would assume that such punctuation automatically implies the truth of the quoted material.") (citing Baker v. Los Angeles Examiner, 42 Cal. 3d 254, 263 (1986)).

157. *Id.* at 513.

158. *See, e.g.*, Takieh v. O'Meara, 497 P.3d 1000, 1006 (Ariz. Ct. App. 2021) (explaining that defamatory statement must be assertions of a false statement of fact).

159. Volokh, *supra* note 101, at 498.

160. KNOWING MACHS., *supra* note 146.

161. *Id.*

of generative AI model's generating factual assertions.[162] Particularly, disclaimers of uncertainty or risk of mistake in an asserted statement does not prevent defamation liability because the asserted statement can still reasonably be perceived as factual.[163]

Additionally, software developers' touting of generative AI models' general reliability and significant feats undermine their ability to argue that LLMs are just "probabilistic" tools.[164] OpenAI's website promotes using ChatGPT to "Get answers."[165] OpenAI's, like many other software developers', business requires building trust and credibility in a generative AI models' ability to produce reasonably accurate facts.[166] Contrast this with an Incorrect Quote Generator.[167] This AI tool can generate funny and fake quotes, attributing them to famous people.[168] While attributing fake quotes to famous people could lead to liability for the software developer, the presentation of the tool and its generated content could not reasonably be seen as factual assertions.[169] The advertised benefits of the tool do include the tool's reliability to produce true or factual content; the title of the tool being upfront of its falsity.[170] Despite the disclaimer and pattern recognition mechanism of generative AI models, the AI-generated content can reasonably be perceived as factual assertions.[171]

Although defamation is often a written or spoken statement, photos can also be defamatory.[172] In *Jewell v. NYP Holdings, Inc.*, a district court found that a picture with a caption that implied false criminal liability of a plaintiff was defamatory.[173] In *Burton v. Crowell Publishing C*o., the Second Circuit found that an ad using optical trickery to depict a plaintiff with enlarged genitals was defamatory.[174] In that case, Judge Learned Hand wrote:

> "[E]verybody would at once see that it was the camera, and the camera alone that had made the unfortunate mistake. If the advertisement is a libel, it is such in spite of the fact that it asserts nothing whatever about the plaintiff, even by the remotest implication. It does not profess to depict him as he is; it does not

162. *See* Henderson et al., *supra* note 101, at 638–39 (assuming users will disregard the disclaimer); *see* Volokh, *supra* note 101, at 500–01 (explaining disclaimers of uncertainty do not prevent defamation liability).

163. Volokh, *supra* note 101, at 500–01.

164. *Id.* at 498 ("And OpenAI has touted ChatGPT as a generally pretty reliable . . . source of assertions of fact, not just as a source of entertaining nonsense.") (footnote omitted); *see* KNOWING MACHS., *supra* note 146 (reporting the preliminary rejection of OpenAI's "probabilistic" argument against the perception that AI-generated content are factual assertions).

165. *ChatGPT*, OPENAI, https://openai.com/chatgpt/overview/ (last visited Oct. 12, 2024).

166. Volokh, *supra* note 101, at 499, 501.

167. *Incorrect Quotes Generator*, YTTAGS.COM, https://www.yttags.com/incorrect-quotes-generator.php [https://perma.cc/JV7X-3FBG] (last visited Oct. 12, 2024).

168. *Id.*

169. *See supra* notes 152–158 and accompanying text (arguing that certain instances of direct quotes could not reasonably be perceived as factual assertions).

170. YTTAGS.COM, *supra* note 167.

171. Volokh, *supra* note 101, at 498 ("And OpenAI has touted ChatGPT as a generally pretty reliable . . . source of assertions of fact, not just as a source of entertaining nonsense.") (footnote omitted).

172. *See* Jewell v. NYP Holdings, Inc., 23 F. Supp. 2d 348 at 364–365 (S.D.N.Y. 1998) ("A photograph which is otherwise an accurate picture of a plaintiff may nonetheless be actionable where the caption suggests something defamatory and false about the plaintiff.").

173. *Id.*

174. Burton v. Crowell Pub. Co., 82 F.2d 154, 156 (2d Cir. 1936).

> exaggerate any part of his person so as to suggest that he is deformed; it is patently an optical illusion and carries its correction on its face as much as though it were a verbal utterance which expressly declared that it was false. It would be hard for words so guarded to carry any sting, but the same is not true of caricatures, and this is an example; for notwithstanding all we have just said, it exposed the plaintiff to overwhelming ridicule."[175]

Both cases suggest that an image with a caption implying a falsehood, and an image altered to present a falsehood can be defamatory.[176]

2. Publication

In *Walters v. OpenAI*, OpenAI argues that ChatGPT's statement does not meet the "publication to a third-party" requirement.[177]

a. Inference of Third-Party Disclosure

As previously stated, Walters faces three obstacles; one of which deal with this publication requirement.[178] In this case, Walters presumably was the only person who saw the alleged defamatory comment generated by ChatGPT.[179] In *Bloodworth v. KS City Board of Police Commissioners*, the Eighth Circuit affirmed that "'[c]ommunication of defamatory matter only to the plaintiff who then discloses to third parties ordinarily does not subject defendant to liability,' unless the defendant 'intends, or has reason to suppose, that in the ordinary course of events the matter will come to the knowledge of some third person.'"[180] Eugene Volokh, citing various lower court cases, argues that this extends to computerized systems.[181]

Eugene Volokh argues automated statements generated by computerized systems can be reasonably inferred to satisfy publication to a third-party.[182] In *Finlay v. MyLife.com*, a website host posted an online profile of the plaintiff with false information that could be seen by prospective employers.[183] The plaintiff contended that the online profile made the defamatory statements available to prospective employers.[184] The district court found that the public availability of statements on an online profile gives rise to a reasonable inference of third-party disclosure.[185] In a more factually similar publication process, the district court,

175. *Id.* at 155.
176. *Jewell*, 23 F. Supp. 2d at 364–65; *Burton*, 82 F.2d at 156.
177. KNOWING MACHS., *supra* note 146.
178. *Supra* notes 93–94 and the accompanying text.
179. *See* Cahill, *supra* note 93 ("After receiving the information, Riehl contacted SAF to disconfirm what ChatGPT had told him. He never repeated any of it in article of his own. He presumably shared the information with Walters as well, although he has not confirmed that detail.").
180. Bloodworth v. Kan. City Bd. of Police Comm'rs, 89 F.4th 614, 622 (2023) (quoting Overcast v. Billings Mut. Ins. Co., 11 S.W.3d 62, 70 (Mo. 2000).
181. *See* Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. FREE SPEECH L. 489, 505–06 (2023) (citing cases where the court inferred publication to third parties in defamation cases).
182. *Id.* at 505.
183. Finlay v. MyLife.com Inc., 525 F. Supp. 3d 969, 974–75 (D. Minn. 2021).
184. *Id.* at 983.
185. *Id.*

in *Shaunfield v. Experian Info. Sols., Inc.*, held that the plaintiff reasonably plead that a credit reporting agency supplied defamatory statements to third parties.[186] The plaintiff alleged that he requested and received his own credit report with numerous false statements from the credit reporting agency.[187] The court found that, based on the facts contended, the plaintiff sufficiently raised a reasonable inference that false credit information was disclosed to third parties.[188]

Based on the Georgia Superior Court's rejection of the motion to dismiss, it could be assumed that the Court found a similar reasonable inference of third-party disclosure.[189] Most likely third-party disclosure will be reasonably inferred and satisfied by AI-generated statements.[190]

b. Volition to Communicate

As foreshadowed in Section III.A, a generative AI model does not have volitive intent.[191] In defamation, "publication of defamatory matter is its communication intentionally or by a negligent act to one other than the person defamed."[192] An important note, emphasized by Volokh: "the 'intentionally or by a negligent act' in this section refers to the act of *communicatio*n . . . . It doesn't refer to knowledge or negligence as to the *falsehood of the statement* . . . ."[193] While a generative AI model does not have the volition to act, courts will most likely turn to the volitional acts of the software developers.[194]

3. Fault in Falsity

In *Walters v. OpenAI*, the court preliminarily rejected OpenAI's argument that users of ChatGPT owned the rights to the ChatGPT outputs and, thus, were solely responsible for them.[195] Scholars do not agree with OpenAI's argument, either definitively arguing that software developers are liable for the AI-generated statements or, at the very least, open to the inquiry.[196] The default

186. Shaunfield v. Experian Info. Sols., Inc., 991 F. Supp. 2d 786, 803 (N.D. Tex. 2014).

187. *Id.* at 795.

188. *Id.* at 803.

189. *See supra* notes 177–88 and accompanying text (reasoning that the discovery of false statements and the public availability, either online or as requested, of the false statements give rise to a reasonable inference of third-party disclosure).

190. *Id.*

191. *See supra* notes 130–42 and accompanying text (explaining that generative AI models do not engage in volitive editing).

192. RESTATEMENT (SECOND) OF TORTS § 577(1) (1977).

193. Volokh, *supra* note 101, at 505.

194. *See* Henderson et al., *supra* note 101, at 636–37 ("The real target in these cases will likely be the company that runs the AI (assuming it is a real company with assets, like OpenAI, and not simply open-source software being passed around by individuals).").

195. KNOWING MACHS., *supra* note 146.

196. *See* Volokh, *supra* note 101, at 500 ("To begin with, such disclaimers can't operate as contractual waivers of liability: Even if the AIs' *users* are seen as waiving their rights to sue based on erroneous information when they submit a query despite seeing the disclaimers, that can't waive the rights of the *third parties* who might be libeled."); *see* Henderson et al., *supra* note 101, at 625–26 ("Absent Section 230 immunity, deployers of generative AI could or should be liable for generated speech or actions in some circumstances.").

standard of fault for private citizens is negligence.[197] When public figures and officials are the subject of a defamatory statement, the fault standard rises to "actual malice."[198] Actual malice means a defendant communicated the defamatory statement "with knowledge that it was false or with reckless disregard of whether it was false or not."[199] As previously noted, a generative AI model does not have the requisite volition to be at fault for negligence or actual malice.[200] As such, this subsection will examine the different standards of fault for the falsity of a defamatory statement.

Actual malice means a defendant communicated the defamatory statement "with knowledge that it was false or with reckless disregard of whether it was false or not."[201] To prevent actual malice, "[p]ublishers do indeed need to take time and effort to investigate potential errors once they are aware of them."[202]

One notable example of actual malice is in the alteration of quotes.[203] In *Masson*, a magazine article and book contained altered statements presented as direct quotes from an interviewee.[204] The interviewee claimed that the altered statements had several errors, portraying him in an unflattering light.[205] The court held that when "the alteration results in a material change in the meaning conveyed by the statement," the defendant satisfies the requisite knowledge of falsity for actual malice.[206] As previously noted, generative AI models lack volitive intent.[207] Scholars have hypothesized that courts most likely will require actual malice to the falsity of a particular fact and not the general risk of generating falsehoods.[208]

C. False Light

Scholars have argued that false light tort claims should be similarly treated like defamation claims.[209] The purpose of false light claims is to provide a remedy for falsehoods that would be "considered objectionable by the average person."[210] Although only some states recognize false light as a distinct legal remedy from defamation, the general principles remain the same in the states that do.[211] One particular issue with false light claims and most likely the reason

197. Wex Definitions Team, *supra* note 147.
198. *Id.*
199. *Id.* (citing N.Y. Times Co. v. Sullivan, 376 U.S. 254, 279–80 (1964)).
200. *See* Henderson et al., *supra* note 101, at 635–36 ("Of course, the AI itself isn't a person, doesn't have money, and can't be sued.").
201. Wex Definitions Team, *supra* note 147.
202. Volokh, *supra* note 101, at 517.
203. Masson v. New Yorker Mag., Inc., 501 U.S. 496, 517 (1991).
204. *Id.* at 499–500.
205. *Id.* at 501.
206. *Id.* at 517.
207. *See supra* notes 130–42 and accompanying text (explaining that generative AI models do not engage in volitive editing).
208. Henderson et al., *supra* note 101, at 641–42.
209. Volokh, *supra* note 101, at 549.
210. *False Light*, CORNELL L. SCH.: LEGAL INFO. INST., https://www.law.cornell.edu/wex/false_light [https://perma.cc/822L-KUU9] (last updated Dec. 2022).
211. *Invasion of Privacy: False Light*, FINDLAW (Aug. 3, 2023), https://www.findlaw.com/injury/torts-and-personal-injuries/invasion-of-privacy--false-light.html [https://perma.cc/8GJ8-M6WP].

for their exclusion from some state books is the potential chilling effect on free speech.[212] Generally, false light claims require that: (1) the false light in which the other was placed would be highly offensive to a reasonable person, and (2) the actor had knowledge of or acted in reckless disregard as to the falsity of the publicized matter and the false light in which the other would be placed.[213]

1. Highly Offensive

False light does not necessarily mean that specifical false statements were made about the individual.[214] Misleading implications that the average person would find highly offensive are often satisfactory.[215] The lower standard of "highly offensive" implicates the previously mentioned free speech issues that defamation does not.[216] In *Braun v. Flynt*, the Fifth Circuit found that providing a plaintiff's picture in an overtly sexual magazine was highly offensive but not defamatory.[217] This case suggests that a hallucinated deepfake would most likely satisfy the highly offensive standards of false light claims more often than the higher defamatory standard.[218]

2. Publication

Scholars note that false light penalizes speech that gives publicity to incorrect factual assertions by communicating it to the public at large or to so many persons that the matter must be regarded as substantially certain to become one of public knowledge.[219] Publication to a small group of people does not satisfy the publication requirement, but a publication to a small newspaper, posting on a shop window, or even sending mail individually to a large group of people does satisfy the requirement.[220] This suggests that a generative AI model that generates a hallucination to multiple people will satisfy the publication requirement.[221]

---

212. *Id*.
213. RESTATEMENT (SECOND) OF TORTS §§ 652E (AM. L. INST. 1977).
214. FINDLAW, *supra note 211*.
215. *Id.*
216. CORNELL L. SCH.: LEGAL INFO. INST., *supra* note 210.
217. Braun v. Flynt, 726 F.2d 245, 258 (5th Cir. 1984).
218. *See id.* ("We hold that there can be no duplicative awards of damages arising from a single publication merely because that publication yields causes of action under state law for both defamation and invasion of privacy.").
219. *See Publicly Placing Person in False Light*, USLEGAL, https://privacy.uslegal.com/what-constitutes-a-violation/publicly-placing-person-in-false-light/# [https://perma.cc/UVX7-2UYE] (last visited Oct. 13, 2024) ("In order to be actionable, the false light in which the plaintiff is placed must be highly offensive to a reasonable person. Although it is not necessary that the plaintiff be defamed, publicity placing one in a highly offensive false light will in most cases be defamatory as well.").
220. *Id.*
221. *See id.* ("In the context of a false-light claim, giving publicity means making a matter public, by communicating it to the public at large, or to so many persons that the matter must be regarded as substantially certain to become one of public knowledge. Communication of a fact to a single person or even to a small group of persons is not publicity.").

3. Fault in Falsity

To safeguard free speech, the Supreme Court placed the same actual malice standard as defamation on false light claims for public figures.[222] For false light claims, the plaintiff must prove actual knowledge or reckless disregard as to both the falsity of the publicized matter and the false light in which the other would be placed.[223] Here, generative AI models retain the same issue of lack of volitive intent illustrated in Section III.B.3, presenting the biggest hurdle to false light claims against generative AI models.[224]

IV. RECOMMENDATION

The ambiguous applicability of current legal regimes[225] has confounded legislators on the sufficiency of the current law, resulting in reactionary, poorly thought-out legislative proposals.[226]

A. Tort Liability is a Hallucination – A Mistake

After a committee hearing about deepfakes, no legislation other than the No Section 230 Immunity for AI Act, which would remove Section 230 protection for AI, was proposed.[227] This proposed legislation to strip Section 230 protection has been criticized for dampening First Amendment rights and technological innovation.[228]

Congress originally enacted Section 230 to prevent chilling of Internet users' speech from restrictions on Internet website hosts, software developers, and other online services that provide a platform for users.[229] Two significant

222. *See* Time, Inc. v. Hill, 385 U.S. 374, 387 (1967) ("Factual error, content defamatory of official reputation, or both, are insufficient for an award of damages for false statements unless actual malice-knowledge that the statements are false or in reckless disregard of the truth- is alleged and proved.").

223. FINDLAW, *supra* note 211.

224. *See id.* ("To have actual malice, the defendant must have knowledge of the falsehood or must otherwise act with reckless disregard in publicizing it.").

225. *See* Jason Cohen, *The Senate's 'No Section 230 Immunity for AI Act' Would Exclude Artificial Intelligence Developers' Liability Under Section 230*, TENN. STAR (Dec. 11, 2023), https://tennesseestar.com/news/the-senates-no-section-230-immunity-for-ai-act-would-exclude-artificial-intelligence-developers-liability-under-section-230/admin/2023/12/11/ [https://perma.cc/8B75-GN3T] ("NetChoice, a group whose members include companies like Google and TikTok, says the definition of artificial intelligence in Hawley and Blumenthal's bill is so vague and broad that it could apply to much more technology than what is typically considered AI.").

226. *Id*.

227. *Id.*

228. *See* Hayley Tsukayama et. al, *Congress Should Not Rush to Regulate Deepfakes*, ELECTRONIC FRONTIER FOUND. (June 24, 2019), https://www.eff.org/deeplinks/2019/06/congress-should-not-rush-regulate-deepfakes [https://perma.cc/YYD4-UB4M] ("Before Congress drafts legislation to regulate deepfakes, lawmakers should carefully consider what types of content new laws should address, what our current laws already do, and how further legislation will affect free speech and free expression."); *See* Adam Thierer & Shoshana Weissmann, *Without Section 230 Protections, Generative AI Innovation Will Be Decimated*, R STREET (Dec. 6, 2023), https://www.rstreet.org/commentary/without-section-230-protections-generative-ai-innovation-will-be-decimated [https://perma.cc/H892-K4TQ] ("The combined result of these proposals would decimate algorithmic innovation and see America shooting itself in the foot as the global race for AI supremacy heats up.").

229. *Section 230: Legislative History*, ELEC. FRONTIER FOUND., https://www.eff.org/issues/cda230/legislative-history [https://perma.cc/W645-GYDU] (last visited Mar. 8, 2024).

legal events prompted action from Congress.[230] First, in *Cubby, Inc. v. CompuServe, Inc*, a district court found that a website host cannot be liable for users' libel posted on the host's website.[231] Specifically, the court found that because the website host did not review user posts, the website host could not have the requisite knowledge to be held liable.[232] Second, in *Stratton Oakmont, Inc. v. Prodigy Servs. Co.*, a district court found that a website host can be liable for users' posts when the website host takes on some moderation of posts.[233] In response, Congress passed Section 230 to (1) protect development of free speech online and (2) allow website hosts and services to moderate with their own standards.[234] Congress explicitly distinguishes the internet from traditional mediums of communication like newspapers.[235]

Scholars argue that stripping Section 230 protection for generative AI models would run against the express purpose of Section 230.[236] They attribute the United States' early success in digital markets and ecommerce because the Section 230 protection.[237] In fact, Australia, which does currently impose some civil penalties to website hosts for hosting defamatory content, is proposing to limit the liability of deepfakes in response to Taylor Swift's case.[238] Although removing Section 230 immunity would not automatically impose civil liability, software development companies would severely limit or remove their generative AI models from public use.[239] Section 230 encourages internet service providers to open up their services for use by internet users, giving internet users more tools and mediums to express their thoughts.[240] Abolishing or stripping liability would hurt internet users and their ability to speak online.[241] This Note argues that legislative proposals to impose civil liability onto generative AI models, either through removing Section 230 protection or broadening tort liability,[242] would chill free speech and hamper economic and

230. *Id.*
231. *Id.*
232. *Id.*
233. *Id.*
234. *Id.*
235. *Id.*
236. *See* Thierer & Weissmann, *supra* note 228 ("Moreover, although scholars debate whether Section 230 ought to apply to AI, there is a very real case for it applying. Courts have regularly applied Section 230 protections to outputs generated by algorithms, including Google's snippets that summarize each search result because they are derived from third-party information."); Adam Thierer, *Artificial Intelligence, Section 230 & Looming Liability Threats*, MEDIUM (April 2, 2024), https://medium.com/@AdamThierer/artificial-intelligence-section-230-looming-liability-threats-a40ad32c67e3 [https://perma.cc/9T6D-XYBQ].
237. *Id.*
238. Jeannie Marie Paterson, *'Picture to Burn': The Law Probably Won't Protect Taylor (or Other Women) from Deepfakes*, PURSUIT (Feb. 8, 2024), https://pursuit.unimelb.edu.au/articles/picture-to-burn-the-law-probably-won-t-protect-taylor-or-other-women-from-deepfakes [https://perma.cc/583P-VNK7].
239. Jennifer Huddleston, *Does Section 230 Cover Generative AI?*, CATO INSTITUTE (Dec. 6, 2023, 11:56 AM), https://www.cato.org/blog/does-section-230-cover-generative-ai [https://perma.cc/F5E2-75PM].
240. 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/3YTF-6PEJ].
241. *Id.*
242. *See* Thierer & Weissmann, *supra* note 228 ("S. 1993 would also have a deleterious effect on economic output. The United States was the unambiguous winner of the first round of the web wars. Digital markets and

technological development. The next part will recommend regulation that allows easier identification of actual tortfeasors, tortfeasors' intent, and deepfakes.

B. Shallow, Real Regulations for Deepfakes

Learned Hand wrote "[t]he only reason why the law makes truth a defense is not because a libel must be false, but because the utterance of truth is in all circumstances an interest paramount to reputation . . . ."[243] In the instance of hallucinated deepfakes, the truth is that generative AI models and their software developers are not speakers but should have a duty to help facilitate the truth.[244] Here, this Note recommends easily implementable technological methods that would ease identification of actual tortfeasors, tortfeasors' intent, and deepfakes.

1. Storage of Prompt History

Agency regulations requiring private entities to store records is common in many areas of regulation.[245] For example, the U.S. Customs and Border Protection requires importers, exporters, and various entities involved in trade through the U.S. border to maintain records.[246] The regulation specifies the required content of the stored records, the minimal retention period, and allowed methods of storage.[247] In another example, the U.S. Securities and Exchange Commission requires parties involved with securities trade to maintain and preserve electronic records.[248] The rules mandate the format, content, and accessibility of electronic records.[249]

Currently, some generative AI models already save and store user data, including user prompts and responses to those user prompts.[250] While there are data privacy issues, in addition to requiring prompt history storage, regulations could also require redactions of user prompts for sensitive information.[251] Additionally, storage of that amount of data is not uncommon as Google has been storing user search history.[252]

---

ecommerce first took off here and American innovators have been global leaders ever since."); *See* Mackey & Mullin, *supra* note 240 ("Engine has estimated that without Section 230, many startups and small services would be inundated with costly litigation that could drive them offline.").

243. Burton v. Crowell Pub. Co., 82 F.2d 154, 156 (2d Cir. 1936).

244. Paterson, *supra* note 238.

245. *See* 19 CFR § 163.5 (2012) ("Any of the persons listed in §163.2 may maintain any records, other than records required to be maintained as original records under laws and regulations administered by other Federal government agencies.").

246. 19 CFR §§ 163.2, 163.5.

247. *Id*.

248. *Amendments to Electronic Recordkeeping Requirements for Broker-Dealers*, U.S. SEC. AND EXCH. COMM'N, https://www.sec.gov/investment/amendments-electronic-recordkeeping-requirements-broker-dealers [https://perma.cc/YB9R-EDQH] (last updated Feb. 28, 2023).

249. *Id*.

250. Bryan Arnott, *Yes, ChatGPT Saves Your Data. Here's How to Keep It Secure.*, FORCEPOINT (Sept. 13, 2023), https://www.forcepoint.com/blog/insights/does-chatgpt-save-data [https://perma.cc/EQ7K-BRDY].

251. *See Top 7 Automatic Redaction Tools that are Worth Trying (Updated 2024)*, REDACTABLE (2024), https://www.redactable.com/blog/automatic-redaction-tools [https://perma.cc/2PE2-B292] (listing automatic redaction tools that are currently available).

252. Tim Fisher, *How to Stop Google from Tracking Your Searches*, LIFEWIRE (July 16, 2021), https://www.lifewire.com/stop-google-from-tracking-your-searches-4123866 [https://perma.cc/69JF-V2M8].

The storage of prompt history would allow a review of the search history to determine if a deepfake was unintentionally created or purposefully created.

2. Attachment of Watermark and Metadata

Regulations that require marking of products exist in various areas.[253] For example, the U.S. Customs and Border Protection requires marking of Country of Origins for U.S. imports.[254] The regulations specify the acceptable forms of the marking, the method to determine which marking to apply, and when markings do apply.[255]

"A watermark is an image, overlay, or text that's placed over a digital asset."[256] Watermarks are commonly used for asset protection by helping identify the owner or source of a photo.[257] Some argue that watermarks on AI-generated pictures would be ineffective because any person with enough time, resources, and motivation could remove the watermark.[258] To combat this problem, "forensic watermarks" could be applied.[259] Forensic watermarks are embedded into the image itself instead of overlayed and is impossible to remove without damaging the image.[260] DALL-E, a popular AI image generator, currently applies a watermark to its generated content.[261]

Additional to the watermark, metadata could be required to input into AI-generated photos.[262] Metadata is data "describing and providing information about rights and administration of an image."[263] Metadata can be embedded into the photo or be a separate file.[264] Metadata can store information about the creator, caption, or other identifiers.[265]

For use with generative AI models, regulations could require a combination of watermarks and metadata. Watermarks could be used to identify when a picture was AI-generated, while metadata would be used to store information

253. *See, e.g.*, *Marking of Country of Origin on U.S. Imports*, U.S. CUSTOMS & BORDER PROT., https://www.cbp.gov/trade/rulings/informed-compliance-publications/marking-country-origin-us-imports [https://perma.cc/VUS3-UUUG] (last updated Aug. 12, 2020) (describing requirements for marking foreign products with their country of origin); Clare McVeigh, *What is a Watermark and Why is Digital Watermarking Software Important?*, MEDIAVALET (June 28, 2024), https://www.mediavalet.com/blog/watermarks-are-important [https://perma.cc/VGD4-RUV9] (describing how watermarks are used to mark photos).

254. U.S. CUSTOMS & BORDER PROT., *supra* note 253.

255. *Id.*

256. McVeigh, *supra* note 253.

257. *Id.*

258. Ashley Belanger, *4chan Daily Challenge Sparked Deluge of Explicit AI Taylor Swift Images*, ARS TECHNICA (Feb. 5, 2024, 3:02 PM), https://arstechnica.com/tech-policy/2024/02/4chan-daily-challenge-sparked-deluge-of-explicit-ai-taylor-swift-images/ [https://perma.cc/KWA8-24VN].

259. McVeigh, *supra* note 253.

260. *Id.*

261. Emilia David, *OpenAI is Adding New Watermarks to DALL-E 3*, THE VERGE (Feb. 6, 2024, 4:32 PM), https://www.theverge.com/2024/2/6/24063954/ai-watermarks-dalle3-openai-content-credentials [perma.cc/D6MY-74NB].

262. *Id.*; *see What is Photo Metadata?*, IPTC, https://www.iptc.org/standards/photo-metadata/photo-metadata/ [https://perma.cc/JS83-VF58] (last visited Mar. 8, 2024) (describing how metadata which identifies the creator can be embedded into images).

263. IPTC, *supra* note 262.

264. *Id.*

265. *Id.*

about which generative AI model was used and could even be used to embed the prompt that was used to generate the image.[266]

V. CONCLUSION

The mainstream availability and intricate development of generative AI models have exposed weaknesses with traditional falsehood torts, culminating in a most distasteful form: Taylor Swift deepfakes.[267] The general anonymity of the internet allows the original publisher of a deepfake to escape tort liability.[268] The astounding AI replication of a real photograph also makes it difficult for internet users to identify deepfakes.[269] Additionally, generative AI models tend to create false content called hallucinations that can result in depicting a public figure in an unsavory situation.[270] When a user publishes a hallucination, without knowing that it is a deepfake, the generative AI model developer's liability is called into question.

Because of the rapid development and introduction of generative AI to the public, legislators have worked at a lethargic pace in dealing with legal issues surrounding AI.[271] When they have pondered on these issues, most of the focus has been on data privacy, intellectual property, and national security.[272] Now that deepfakes of an international pop star have surfaced,[273] the legislators have reacted, but in the most extreme way. Legislators have proposed the removal of Section 230 which was specifically enacted to protect First Amendment rights of internet users while facilitating innovation.[274]

Section 230 provides protection to a provider of an interactive computer service from liability of users who use the interactive computer service.[275] Current court interpretations and rulings of Section 230 provide an ambiguous view of whether courts would decide if generative AI model developers are

266. *See id*. (describing the types of data that metadata can embed into an image); *see also* David, *supra* note 261 (explaining how OpenAI is already adding watermarks to indicate that a product was AI-generated).

267. *See supra* Part I (describing incident of sexually explicit artificial intelligence (AI) generated images of Taylor Swift spread throughout social media websites).

268. *See* McVeigh, *supra* note 253 (discussing that a watermark permits an original publisher of an image to be identified).

269. *See* UNIV. OF VA., *supra* note 62 (explaining the advancements of deepfake creation and how it can spread misinformation).

270. *See* Glover, *supra* note 91 (explaining the dangers of AI hallucinations of spreading misleading information).

271. *See* Walsh, *supra* note 75 (discussing how the development of AI has created a "legal minefield" with uncertainty in how to apply current copyright and privacy laws to new AI developments); Ryland Barton, *As Congress Lags, States have taken the Lead in Regulating the Emerging AI Industry*, NPR (Feb. 13, 2024, 5:43 PM), https://www.npr.org/2024/02/13/1231221329/as-congress-lags-states-have-taken-the-lead-in-regulating-the-emerging-ai-indust [https://perma.cc/WYT2-PV56].

272. *See* sources cited *supra* note 75 (explaining how privacy law, and intellectual property law have been applied to generative AI regulation).

273. *Supra* Part I (discussing viral deepfake created about Taylor Swift).

274. 47 U.S.C. § 230.

275. *Id.* § (c)(1); Weitzman & Herndon, *supra* note 106.

protected.[276] Even if a court does not grant Section 230 immunity, plaintiffs must still prove the underlying tort.[277]

The two most common falsehood claims available are defamation and false light.[278] Because both claims involve penalizing speech about public figures, the Supreme Court has required that actors engage in defamatory or highly offensive falsehoods with knowledge of or reckless disregard for the falsehoods.[279] The current jurisprudence of algorithms and computer-based systems indicate that a generative AI model does not act with the volitive intent arising to knowledge or reckless disregard.[280]

Because of the First Amendment safeguards erected by the legislature's enactment of Section 230 and the Supreme Court's standard of actual malice, it is not recommended to attempt to change those areas of law. The more prudent course, while still preserving the original purpose of Section 230,[281] would be to mandate regulations on AI software developers that would ease the burden of identifying deepfakes and their source.

276. *See supra* Part III.A (discussing how various courts have attempted to classify an information content provider).

277. *See, e.g.*, Anthony v. Yahoo! Inc., 421 F. Supp. 2d 1257, 1263 (N.D. Cal. 2006) (holding that the defendant is not immunized by Section 230 and that the plaintiff may still proceed with their tort claims in the case).

278. *See, e.g.*, Ferraro & Tompros, *supra* note 26 (discussing various legal claims against falsehoods like defamation, false light, and right of publicity).

279. *See* Wex Definitions Team, *supra* note 147 (citing N.Y. Times Co. v. Sullivan, 376 U.S. 254, 279–280 (1964)).

280. *See* Henderson et al., *supra* note 101 at 641 ("It seems unlikely that any software design could be said to act with reckless disregard for truth or actual knowledge that it would produce a false defamatory statement . . . .").

281. 47 U.S.C. § 230.