Recommended Amendments to Legislative Proposals on Deepfakes

13th May 2024

Policy team
policy@futureoflife.org

Contents

The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence (AI). Since its founding, FLI has taken a leading role in advancing key disciplines such as AI governance, AI safety, and trustworthy and responsible AI, and is widely considered to be among the first civil society actors focused on these issues. FLI was responsible for convening the first major conference on AI safety in Puerto Rico in 2015, and for publishing the Asilomar AI principles, one of the earliest and most influential frameworks for the governance of artificial intelligence, in 2017. FLI is the UN Secretary General's designated civil society organization for recommendations on the governance of AI and has played a central role in deliberations regarding the EU AI Act's treatment of risks from AI. FLI has also worked actively within the United States on legislation and executive directives concerning AI. Members of our team have contributed extensive feedback to the development of the NIST AI Risk Management Framework, testified at Senate AI Insight Forums, participated in the UK AI Summit, and connected leading experts in the policy and technical domains to policymakers across the US government.

FLI's wide-ranging work on artificial intelligence can be found at futureoflife.org.

The NO FAKES Act of 2023

The NO FAKES Act of 2023 was introduced by Sens. Coons, Blackburn, Klobuchar, and Tillis.

- Creates a property right over one's own likeness in relation to digital replicas which are nearly indistinguishable from the actual image, voice, or visual likeness of that individual.
- Imposes civil liability for the non-consensual production, publication, distribution, or transmission of a digital replica.
- States that it would not be a defense for the defendant to have included a disclaimer claiming that the digital replica was unauthorized or that they did not participate in its creation, development, distribution, or dissemination.

ANALYSIS

This proposal would enable victims to bring civil liability claims against persons who produce, publish, transmit, or otherwise distribute deepfakes of them. While this might capture model developers or providers whose tools enable the production of such deepfakes, this interpretation should be made clearer by the bill. Moreover, civil liability for breaching a property right will not be as strong of a deterrent as criminal liability, which better reflects the level of harm deepfakes pose to the public.

Recommended Amendment

Amend § 2(c) as follows:

(1) IN GENERAL.—Any person that, in a manner affecting interstate or foreign commerce (or using any means or facility of interstate or foreign commerce), engages in an activity described in paragraph (2) shall be jointly and severally liable in a civil action brought under subsection (d) for any damages sustained by the individual or rights holder injured as a result of that activity.

(2) ACTIVITIES DESCRIBED.—An activity described in this paragraph is either of the following:

(A) The production of a digital replica without consent of the applicable individual or rights holder.

(B) The publication, distribution, or transmission of, or otherwise making available to the public, an unauthorized digital replica, if the person engaging in that activity has knowledge that the digital replica was not authorized by the applicable individual or rights holder.

(C) The publication, distribution, or transmission of, or otherwise making available to the public, a system or application which employs artificial intelligence to produce a digital replica without consent of the applicable individual or rights holder, if the system or application can generate such unauthorized digital replicas without extraordinary effort.

The NO AI FRAUD Act of 2024

The NO AI FRAUD Act of 2024 was introduced by Rep. Salazar along with other House lawmakers.

- Creates a property right in each person's rights and likeness.
- Imposes civil liability for distributing, transmitting, or otherwise making available to the public a "personalized cloning service", with a remedy of at least $50,000.
- Defines personalized cloning services as algorithms, software, tools, or other technologies, services, or devices the primary purpose or function of which is to produce digital voice replicas or digital depictions of particular, identified individuals.
- Also imposes civil liability for publishing, performing, distributing, transmitting, or otherwise making available to the public a deepfake.

ANALYSIS

This proposal is strong in that it specifically incorporates model developers and providers through a civil liability provision for persons distributing, transmitting, or otherwise publishing a "personalized cloning service". However, the definition of "personalized cloning service" is limited by requiring the "primary purpose" of the AI system to be creating digital replicas. Because generative AI systems are often multi-purpose, this definition would therefore fail to capture systems which were not specifically designed to generate harmful deepfakes but which are actively being used for this purpose. Furthermore, the proposal requires actual knowledge that the individual or rights holder did not consent to the conduct in order to hold the creator, distributor, or other contributor liable. This standard is remarkably hard to meet and may discourage moderation of content produced by these services to ensure that there is plausible deniability of knowledge of what would otherwise be a violation of the bill. The remedy for providing such a service would also start at only $50,000 per violation. While the generation of many deepfakes would constitute multiple violations for those creating the deepfakes themselves, the publication of a service used to create them, i.e. the source of the deepfake supply chain, would seemingly only constitute a single violation regardless of the number of deepfakes produced by the service. This penalty may therefore be insufficient for incentivizing the incorporation of robust safeguards by model developers and providers.

Recommended Amendment

Amend §3(a)(3) as follows:

(3) The term "personalized cloning service" means an algorithm, software, tool, or other technology, service, or device ~~the primary purpose or function of which~~ which can, without extraordinary effort, produce one or more digital voice replicas or digital depictions of particular, identified individuals.

Amend §3(c)(1) as follows:

(1) IN GENERAL. -- Any person or entity who, in a manner affecting interstate or foreign commerce (or using any means or facility of interstate or foreign commerce), and without consent of the individual holding the voice or likeness rights affected thereby--

(A) distributes, transmits, or otherwise makes available to the public a personalized cloning service;

Continued →

(B) publishes, performs, distributes, transmits, or otherwise makes available to the public a digital voice replica or digital depiction ~~with knowledge~~ that the person knows or should have known ~~that the digital voice replica or digital depiction~~ was not authorized by the individual holding the voice or likeness rights affected thereby; or

(C) materially contributes to, directs, or otherwise facilitates any of the conduct proscribed in subparagraph (A) or (B) ~~with knowledge~~ and knows or should have known that the individual holding the affected voice or likeness rights has not consented to the conduct, shall be liable for damages as set forth in paragraph (2).

Amend §3(c)(2) as follows:

(A) The person or entity who violated the section shall be liable to the injured party or parties in an amount equal to the greater of—

(i) in the case of an unauthorized distribution, transmission, or other making available of a personalized cloning service, fifty thousand dollars ($50,000) per ~~violation~~ injured party or the actual damages suffered by the injured party or parties as a result of the unauthorized use, plus any profits from the unauthorized use that are attributable to such use and are not taken into account in computing the actual damages; and

The Preventing Deepfakes of Intimate Images Act

The Preventing Deepfakes of Intimate Images Act was introduced by Rep. Morelle and Rep. Kean.

- Prohibits the disclosure of intimate digital depictions, defining such depictions broadly to include sexually explicit content or altered images.
- Outlines both civil and criminal penalties for unauthorized disclosure, provides relief options for victims, and includes exceptions for certain disclosures, such as those made in good faith to law enforcement or in matters of public concern.

ANALYSIS

The bill successfully creates a two-pronged approach to combating non-consensually generated, sexually explicit material by introducing both civil and criminal penalties. The bill draws on the definition of 'disclose' outlined in the Violence Against Women Act Reauthorization Act of 2022 which is to "transfer, publish, distribute, or make accessible." While this definition might encompass the actions performed by platforms which facilitate the distribution of deepfakes and developers who enable the creation of deepfakes, the language of the bill must be clarified in order to clearly capture these key entities in the supply chain.

Recommended Amendment

Amend §1309A(b) "Right of Action" by adding the following clause:

"(1) IN GENERAL.—Except as provided in subsection (e), an individual who is the subject of an intimate digital depiction that is disclosed, in or affecting interstate or foreign commerce or using any means or facility of interstate or foreign commerce, without the consent of the individual, where such disclosure was made by a person who knows that, or recklessly disregards whether, the individual has not consented to such disclosure, may bring a civil action against that person in an appropriate district court of the United States for relief as set forth in subsection (d).

"(2) RIGHT OF ACTION AGAINST DEVELOPERS.–Except as provided in subsection (e), an individual who is the subject of an intimate digital depiction that is disclosed, in or affecting interstate or foreign commerce or using any means of facility of interstate or foreign commerce, without the consent of the individual may bring a civil action against a developer or provider of an artificial intelligence system if both of the following apply:

"(A) the artificial intelligence system was used to create, alter, or distribute the intimate digital depiction; and

"(B) the developer or provider of the artificial intelligence system intentionally or negligently failed to implement reasonable measures to prevent such use.

[...]

The AI Labeling Act of 2023

The AI Labeling Act of 2023 was proposed by Sens. Schatz and Kennedy.

- Requires that each AI system that produces text, image, video, audio, or multimedia AI-generated content include on such AI-generated content a clear and conspicuous notice.
- Requires that the notice include an identification that the content is AI-generated, the identity of the tool used to create the content, and the date and time the content was created.
- Requires that, to the greatest extent possible, the disclosure must also be permanent or difficult to remove.
- Imposes a responsibility on developers and third-party licensees to "implement reasonable procedures to prevent downstream use of such system without the disclosures required". There is also a duty to include in any licenses a clause prohibiting the removal of the disclosure notice.

ANALYSIS

Disclosing the tool used to generate the content in the notice can make it easier to identify and hold developers accountable for the harmful deepfakes their models create. This can also help identify models that are particularly vulnerable to being employed for the creation of deepfakes due to insufficient design safeguards. Though we suggest the authors clarify that identification of the tool used to create the content should include, at a minimum, the model name, model version, and the name of the model's developer, we applaud the authors' foresight in including this requirement. We further note that many deepfakes, such as those which are sexually abusive in nature, are harmful in and of themselves, even if it is known that the content has been manipulated. As such, while useful, labeling of AI-generated content alone is likely insufficient to discourage deepfake proliferation and incentivize responsible design of generative AI systems without imposing liability for harms from the deepfakes themselves. That said, this may best be accomplished through complementary legislation.

The AI Disclosure Act of 2023

The AI Disclosure Act of 2023 was introduced by Rep. Torres in 2023.

- Requires that any output generated by AI includes: "Disclaimer: this output has been generated by artificial intelligence."
- States that failing to include this disclaimer would violate the Federal Trade Commission Act (15 U.S.C. 57a(a)(1)(B)) regarding unfair or deceptive acts or practices.

ANALYSIS

These types of disclaimers can be easily removed from the outputs of generative AI, such as videos, images, and audio, meaning malicious actors could continue to utilize deepfakes for deceptive purposes, which would be made more deceptive by absence of a disclaimer. It is also not clear which party would be responsible for the lack of disclaimer in the output. It would be most reasonable for the model developer to bear this responsibility given that they can designs models to ensure all outputs incorporate this disclaimer, but without any requirement to include information identifying the generative AI model used to create the content, enforcement would be extremely challenging.

Recommended Amendment

Amend §2(a) as follows:

(a) Disclaimer Required.—Generative artificial intelligence shall include on any output generated by such artificial intelligence the following: "Disclaimer: this output has been generated by artificial intelligence."

(b) Information Required.—Generative artificial intelligence shall embed in any output generated by such artificial intelligence the following information concerning the artificial intelligence model used to generate that output:

(1) The model name;

(2) The model version;

(3) The name of the person responsible for developing the model; and

(4) The date and time the output was generated by that model.

[...]

The DEEPFAKES Accountability Act

The DEEPFAKES Accountability Act was introduced by Rep. Clarke.

- Requires that deepfakes include a contents provenance disclosure, such as a notice that the content was created using artificial intelligence.
- Requires that audiovisual content includes no less than one verbal statement to this effect and a written statement at the bottom of any visual component.
- Imposes a criminal penalty for a failure to comply with disclosure requirements where the noncompliance is:
  - intended to cause violence or physical harm;
  - intended to incite armed or diplomatic conflict;
  - intended to interfere in an official proceeding;
  - in the course of criminal conduct related to fraud;
  - intended to humiliate or harass a person featured in the deepfake in a sexual manner;
  - by a foreign power intending to influence a domestic public policy debate, interfere in an election, or engage in other unlawful acts.
- Also applies the criminal penalty where a person removes or meaninfully obscures the disclosures, with the intent of distribution.
- Grants a private right of action against a person who fails to compy with the disclosure requirements, even if there is only a "tangible risk" of suffering the enumerated harms.

ANALYSIS

While the inclusion of criminal penalties may increase compliance and reflects the range of harms posed by deepfakes, this penalty would only apply for failure to include a contents provenance disclosure. Therefore, producing and sharing harmful deepfakes would still be permitted. Moreover, the bill includes an exception to the disclosure requirements for deepfakes created by an officer or employee of the United States in furtherance of public safety or national security, which could create a wide loophole. Finally, the contents provenance provision does not appear to require disclosure of the specific model that was used to generate the deepfake content; as such it would be difficult to identify the upstream model provider responsible for a harmful deepfake.

Recommended Amendment

Amend §1041 as follows:

"(c) Audiovisual Disclosure.—Any advanced technological false personation records containing both an audio and a visual element shall include—

"(1) not less than 1 clearly articulated verbal statement that identifies the record as containing altered audio and visual elements, ~~and~~ a concise description of the extent of such alteration, a timestamp of when the content was generated or altered, and, where applicable, a description including the name, version, and developer of the artificial intelligence system used to generate or alter the record;

Continued →

"(2) an unobscured written statement in clearly readable text appearing at the bottom of the image throughout the duration of the visual element that identifies the record as containing altered audio and visual elements, and a concise description of the extent of such alteration; and

"(3) a link, icon, or similar tool to signal that the content has been altered by, or is product of, generative artificial intelligence or similar technology.

"(d) Visual Disclosure.—Any advanced technological false personation records exclusively containing a visual element shall include an unobscured written statement in clearly readable text appearing at the bottom of the image throughout the duration of the visual element that identifies the record as containing altered visual elements, and either—

"(1) a concise description of the extent of such alteration, a timestamp of when the content was generated or altered, and, where applicable, a description including the name, version, and developer of the artificial intelligence system used to generate or alter the record; or

"(2) a clearly visible link, icon, or similar tool to signal that the content has been altered by, or is the product of, generative artificial intelligence or similar technology, along with a description, either linked or embedded, which provides the timestamp of when the content was generated or altered, and, where applicable, a description including the name, version, and developer of the artificial intelligence system used to generate or alter the record.

"(e) Audio Disclosure.—Any advanced technological false personation records exclusively containing an audio element shall include,

"(1) at the beginning of such record, a clearly articulated verbal statement that identifies the record as containing altered audio elements and a concise description of the extent of such alteration, a timestamp of when the content was generated or altered, and, where applicable, the name, version, and developer of the artificial intelligence system used to generate or alter the record; and

"(2) in the event such record exceeds two minutes in length, not less than 1 additional clearly articulated verbal statement and additional concise description at some interval during each two-minute period thereafter.

The Protecting Consumers from Deceptive AI Act

The Protecting Consumers from Deceptive AI Act was introduced by Reps. Eshoo and Dunn.

- Directs the National Institute of Standards and Technology to establish task forces to develop watermarking guidelines for identifying content created by AI.
- Requires providers of generative AI applications to ensure that content created or modified by their application includes machine-readable disclosures acknowledging its AI origin.
- Requires providers of generative AI applications to make available to users the ability to incorporate, within the metadata of content created or modified by the application, information including the AI application used to create or modify the content.
- Requires providers of covered online platforms to prominently display disclosures included in the content accessed through their platform.
- States that violations of these watermarking provisions would constitute unfair or deceptive practices under the FTC Act.

ANALYSIS

The development of content provenance standards through dedicated NIST task forces would be valuable in establishing best practices for watermarking. While the proposal requires that application providers enable users to include the model name and version in the metadata of AI-generated content, application providers and model developers should be required to automatically embed this information in the metadata in order to facilitate tracing back to the relevant developer and provider. This enables greater accountability and incentivizes design choices that incorporate protections against deepfakes by default. However, the bill does not address how users, developers, and platforms might be held accountable for the harmful deepfakes they create, publish, and distribute outside of simply requiring watermarking of these deepfakes.

Recommended Amendment

Amend §2(b)(1) as follows:

(1) Developers and providers of generative artificial intelligence ~~applications~~.—A person who develops, distributes, transmits, or makes available to users a generative artificial intelligence model or a software application based on generative artificial intelligence technology shall—

(A) ensure that audio or visual content created or substantially modified by such model or application incorporates (as part of such content and in a manner that may or may not be perceptible by unaided human senses) a disclosure that-- [...]

(D) ensure that such model or application ~~makes available to users the ability to incorporate~~ incorporates, within the metadata of content created or modified by such model or application, information regarding the generative artificial intelligence origin of such content, including tamper-evident information regarding—

(i) the name of such application, where applicable;

(ii) the name, ~~and~~ version, and developer of the generative artificial intelligence model utilized ~~by such application~~ to create or modify such content;

(iii) the date and time associated with the creation or modification of such content by such model or application; and

(iv) the portion of such content that was created or modified by such model or application.

The DEFIANCE Act of 2024

The DEFIANCE Act of 2024 is sponsored by Sens. Durbin, Hawley, Klobuchar, and Graham.

- Imposes civil liability for disclosing, soliciting, and intending to disclose a "digital forgery".
- Defines "digital forgery" as "any intimate visual depiction of an identifiable individual" created through means including artificial intelligence, regardless of whether it includes a label.

ANALYSIS

The bill explicitly states that sexually explicit deepfakes which include disclaimers should still be subject to civil liability, and it also prohibits the production and possession of sexual deepfakes in certain cases. However, it is not clear that it limits a provider's ability to supply individuals with deepfake-generating services, nor does it incentivize developers of AI systems to ensure that their models cannot produce this content. Furthermore, we note that the definition for a "digital forgery" may be too broad in that it could also capture manipulated content created by means other than artificial intelligence, such as traditional image editing software. AI-powered production and alteration of visual content arguably necessitates a distinct approach from traditional tools, as AI technology trivializes the process of producing realistic deepfakes at scale.

Recommended Amendment

Amend §2 CIVIL ACTION RELATING TO DISCLOSURE OF INTIMATE IMAGES as follows:

(b) CIVIL ACTION.—Section 1309(b) of the Consolidated Appropriations Act, 2022 (15 U.S.C. 6851(b)) is amended—

(1) in paragraph (1)—

(A) by striking paragraph (A) and inserting the following:

"(A) IN GENERAL.—Except as provided in paragraph (5)—

"(i) an identifiable individual whose intimate visual depiction is disclosed, in or affecting interstate or foreign commerce or using any means or facility of interstate or foreign commerce, without the consent of the identifiable individual, where such disclosure was made by a person who knows or recklessly disregards that the identifiable individual has not consented to such disclosure, may bring a civil action against the developer or provider of an artificial intelligence system used to generate or modify the depiction, the ~~that~~ person that disclosed the intimate digital depiction, or any combination thereof, in an appropriate district court of the United States for relief as set forth in paragraph (3);