AMENDED IN SENATE AUGUST 23, 2024

AMENDED IN SENATE AUGUST 22, 2024

AMENDED IN SENATE JUNE 24, 2024

AMENDED IN SENATE JUNE 10, 2024

AMENDED IN ASSEMBLY APRIL 18, 2024

AMENDED IN ASSEMBLY MARCH 21, 2024

CALIFORNIA LEGISLATURE—2023–24 REGULAR SESSION

ASSEMBLY BILL No. 3211

Introduced by Assembly Member Wicks

February 16, 2024

An act to add Chapter 41 (commencing with Section 22949.90) to Division 8 of the Business and Professions Code, relating to artificial intelligence.

LEGISLATIVE COUNSEL'S DIGEST

AB 3211, as amended, Wicks. California Digital Content Provenance Standards.

Existing law requires the Secretary of Government Operations to develop a coordinated plan to, among other things, investigate the feasibility of, and obstacles to, developing standards and technologies for state departments to determine digital content provenance. For the purpose of informing that coordinated plan, existing law requires the secretary to evaluate, among other things, the impact of the proliferation of deepfakes, as defined.

This bill, the California Digital Content Provenance Standards, would require a generative artificial intelligence (AI) provider, as provided,

to, among other things, apply provenance data to synthetic content produced or significantly modified by a generative AI system that the provider makes available, as those terms are defined, and to conduct adversarial testing exercises, as prescribed. The bill would prohibit, among other things, providers and distributors of software and online services from making available a system, application, tool, or service that is designed for the primary purpose of removing provenance data from synthetic content, as provided.

This bill would require a newly manufactured recording device sold, offered for sale, or distributed in California to offer users the option to apply difficult to remove provenance data to nonsynthetic content produced by that device and would require the application of that provenance data to be compatible with state-of-the-art adopted and relevant industry standards. If technically feasible and secure, the bill would require a recording device manufacturer to offer a software or firmware update enabling a user of a recording device manufactured before July 1, 2026, and purchased in California to apply difficult to remove provenance data to the nonsynthetic content created by the device and decode any provenance data attached to nonsynthetic content created by the device.

This bill would require a large online platform, as defined, capable of disseminating specified content to use labels to disclose, as specified, any machine-readable provenance data detected in synthetic content that is distributed on its platform. If content uploaded to or distributed on a large online platform by a user does not contain specified provenance data or if the content's provenance data cannot be interpreted or detected, the bill would require a large online platform to label the content as having unknown provenance. The bill would require a large online platform to use a visual disclosure that contains specified information, including the copyrightholder or licensor information, when labeling and disclosing provenance data of sound recordings and music videos.

Beginning July 1, 2026, and annually thereafter, this bill would require a large online platform to produce a transparency report that identifies moderation of deceptive synthetic content on their platform and would authorize that report to include, among other things, instances where synthetic or potentially deceptive content was identified and removed by the platform, as applicable.

This bill would authorize the Department of Technology (department) to assess specified administrative penalties for prescribed violations of

the bill's provisions, including an administrative penalty of up to ~~$500,000~~ *$100,000* for each violation that is intentional or is the result of grossly negligent conduct, to be deposited in the Digital Content Provenance Administrative Fund which the bill would establish in the State Treasury. The bill would, upon appropriation by the Legislature for this express purpose, authorize the expenditure of moneys in the fund by the department to administer these provisions.

This bill would make its provisions operative on July 1, 2026.

Vote: majority. Appropriation: no. Fiscal committee: yes. State-mandated local program: no.

The people of the State of California do enact as follows:

SECTION 1. The Legislature finds and declares all of the following:

(a) Generative artificial intelligence (GenAI) technologies are increasingly able to synthesize images, audio, and video content in ways that are harmful to society.

(b) In order to reduce the severity of the harms caused by GenAI, it is important for photorealistic synthetic content to be clearly disclosed and labeled.

(c) Failing to appropriately label synthetic content created by GenAI technologies can skew election results, enable defamation, and erode trust in the online information ecosystem.

(d) The Legislature should act to adopt standards pertaining to the clear disclosure and labeling of synthetic content, in order to alleviate harms caused by the misuse of GenAI technologies.

(e) The Legislature should push for the creation of tools that allow Californians to assess the provenance of content distributed online and the ways in which content has been significantly altered or completely synthesized by GenAI.

(f) The Legislature should require online platforms to label synthetic content produced by GenAI.

(g) Through these actions, the Legislature can help to ensure that Californians remain safe and informed.

SEC. 2. Chapter 41 (commencing with Section 22949.90) is added to Division 8 of the Business and Professions Code, to read:

Chapter 41. California Digital Content Provenance Standards

22949.90. For purposes of this chapter, the following definitions apply:

(a) "Adversarial testing" means a structured testing effort to find flaws and vulnerabilities in a generative AI system's ability to attach robust provenance data to synthetic content created by the system and access potential risks associated with misuse of the generative AI system to attach false provenance data to digital content generated outside of the generative AI system.

(b) "Artificial intelligence" or "AI" means an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments.

(c) "Digital fingerprint" means a unique value that can be used to identify identical or similar digital content.

(d) "Digital signature" means a cryptography-based method that identifies the user or entity that attests to the information provided in the signed section.

(e) "Generative AI hosting platform" means an online repository or other internet website that makes a generative AI system available for use by a California resident, regardless of whether the terms of that use include compensation.

(f) "Generative AI provider" or "GenAI provider" means an organization or individual that creates, codes, substantially modifies, or otherwise produces a generative AI system that is made publicly available for use by a California resident, regardless of whether the terms of that use include compensation.

(g) "Generative AI system" or "GenAI system" means an artificial intelligence system that can generate derived synthetic content, including images, videos, and audio, and that emulates the structure and characteristics of the system's training data.

(h) "Large online platform" means a public-facing social media platform, as defined in Section 22675, video-sharing platform, messaging platform, advertising network, or standalone search engine that displays content to viewers who are not the creator or collaborator and had at least 2,000,000 unique monthly California users during the preceding 12 months.

(i) "Metadata" means structural or descriptive information about data.

(j) "Nonsynthetic content" means images, videos, or audio captured in the physical world by natural persons using a recording device and is without any modifications or with only minor modifications that do not lead to significant changes to the perceived contents or meaning of the content. Minor modifications include, but are not limited to, changes to brightness or contrast of images and removal of background noise in audio.

(k) "Provenance data" means data that records the origin or history of digital content and is communicated using state-of-the-art techniques based on widely adopted and relevant industry standards. "Provenance data" may be communicated using digital fingerprinting to associate metadata with digital content, attaching metadata to digital content, including through the use of a digital signature, or embedding of watermarks in digital content.

(*l*) "Provenance detection tool" means a software tool or online service that can read or interpret a watermark, metadata, or digital signature, and output the associated provenance data.

(m) "Synthetic content" means ~~information, including~~ images, videos, and audio, that has been produced or significantly modified by a generative AI system.

(n) "Watermark" means information covertly embedded into digital content, including image, audio, and video, for the purpose of communicating the provenance, history of modification, or history of conveyance.

22949.90.1. (a) A generative AI provider whose GenAI system is capable of producing digital content that would falsely appear to a reasonable person to depict real-life persons, objects, places, entities, or events shall do all of the following:

(1) (A) Apply provenance data, either directly or through the use of third-party technology, to synthetic content produced or significantly modified by a generative AI system that the GenAI provider makes available. The GenAI provider shall make the provenance data difficult to ~~remove,~~ *remove or disassociate,* taking into account the accuracy of the provenance data, the quality of the content produced or significantly modified by the generative AI system, and widely accepted industry standards on ~~the robustness of~~ provenance data.

(B) The application of provenance data to synthetic content, as required by subparagraph (A), shall, at minimum, be difficult to ~~remove~~*, remove or disassociate,* identify the digital content as synthetic, and communicate the following provenance data in order of priority, with clause (i) being the most important, and clause (iv) being the least important:

(i) The synthetic nature of the content.

(ii) The name of the generative AI provider.

(iii) If feasible for the provenance technique used, the time and date the provenance data was applied.

(iv) If applicable and feasible for the provenance technique used, the specific portions of the content that are synthetic.

(2) (A) A generative AI provider shall create and make available to the public a provenance detection tool or permit users to use a provenance detection tool provided by a third party. The provenance detection tool shall be based on broadly adopted industry standards and, if technically feasible, meet the following criteria:

(i) The tool allows a user to assess whether digital content was created or altered by a generative AI system.

(ii) The tool allows a user to determine how digital content was created or altered by a generative AI system.

(iii) The tool outputs any provenance data that is detected in the content.

(iv) The tool is publicly accessible through the generative AI provider's or the third-party's internet website, its mobile application, or an application programming interface, as applicable.

(v) The tool allows a user to upload content or provide a uniform resource locator (URL) linking to online content.

(B) A generative AI provider or third party shall put in place a process to collect user feedback related to the efficacy of the provenance detection tool described in subparagraph (A) and incorporate any feedback into any attempt to improve the efficacy of the tool.

(C) A generative AI provider that creates or makes available a provenance detection tool pursuant to subparagraph (A) may limit access to the decoder to ensure the robustness and security of their provenance data techniques.

(3) (A) Conduct adversarial testing exercises following relevant guidelines from the National Institute of Standards and Technology. The adversarial testing exercises shall assess both of the following:

(i) The robustness of provenance data methods.

(ii) Whether the generative AI provider's GenAI systems can be used to add false provenance data to content generated outside of the system.

(B) Adversarial testing exercises required by this paragraph shall be conducted before the general audience release of any new tool or method used to apply provenance data to synthetic content produced or significantly modified by a generative AI system that the GenAI provider makes available.

(C) In the event that a generative AI provider utilizes a third-party tool or method to apply provenance data, the generative AI provider may rely on the testing conducted by the provider of the third-party tool or method pursuant to paragraph (2).

(D) A generative AI provider shall submit full reports of its adversarial testing exercises to the Department of Technology within 90 days of conducting an adversarial testing exercise pursuant to this paragraph. The report shall address any material, systemic failures in a generative AI system related to the erroneous or malicious inclusion or removal of provenance data.

(E) *(i)* Upon the request of an accredited academic institution, a generative AI provider shall make available a summary or report of its adversarial testing exercises.

(ii) The provider may deny a request if providing a summary or report to the relevant institution would undermine the robustness or security of its provenance data techniques.

(F) This paragraph does not require the disclosure of trade secrets, as defined in Section 3426.1 of the Civil Code.

(b) Providers and distributors of software and online services shall not make available a system, application, tool, or service that is designed for the primary purpose of removing provenance data from synthetic content in a manner that would be reasonably likely to deceive a consumer of the origin or history of the content.

(c) Generative AI hosting platforms shall not make available a generative AI system that does not allow a GenAI provider, to the greatest extent possible and either directly providing functionality or making available the technology of a third-party vendor, to apply provenance data to content created or substantially modified

by the system in a manner consistent with specifications set forth in paragraph (1) of subdivision (a).

22949.90.2. (a) (1) A newly manufactured recording device sold, offered for sale, or distributed in California shall offer users the option to apply difficult to remove provenance data to nonsynthetic content produced by that device.

(2) A user shall have the option to not apply provenance data and any other information attached to nonsynthetic content produced by their device and to customize the types of provenance data attached to nonsynthetic content produced by their device, including by removing any personally identifiable information. Personally identifiable information, including geolocation, shall not be included in provenance data by default.

(3) Recording devices subject to the requirements of this subdivision shall clearly inform users of the existence of the settings relating to provenance data upon a user's first use of the recording function on the recording device.

(4) When a recording device's recording function is in use, the recording device shall contain a clear indicator when provenance data is being applied.

(5) The option to apply provenance data to nonsynthetic content produced by a recording device, as described by paragraph (1), shall also be applied to nonsynthetic content produced using third-party applications that bypass default recording applications in order to offer recording functionalities.

(6) The application of provenance data shall be compatible with state-of-the-art widely adopted and relevant industry standards.

(b) If technically feasible and secure, a recording device manufacturer shall offer a software or firmware update enabling a user of a recording device manufactured before July 1, 2026, and purchased in California to do both of the following:

(1) Apply difficult to remove provenance data to the nonsynthetic content created by the device.

(2) Decode any provenance data attached to the nonsynthetic content created by the device.

22949.90.3. (a) A large online platform capable of disseminating content that would falsely appear to a reasonable person to depict real-life persons, objects, places, entities, or events shall use labels to disclose any machine-readable provenance data detected in synthetic content distributed on its platform.

(1) To the extent technically feasible, the labels shall indicate whether provenance data is available.

(2) A user shall be able to click or tap on a label to inspect provenance data in an easy-to-understand format.

(b) The disclosure required under subdivision (a) shall be readily legible to an average viewer or, if the content is in audio format, shall be clearly audible.

(c) If content uploaded to or distributed on a large online platform by a user does not contain provenance data or if the content's provenance data cannot be interpreted or detected by the platform using technically feasible methods, a large online platform shall label the content as having unknown provenance.

(d) A large online platform shall add the following provenance data to digital content published on their platform:

(1) The name of the platform on which the content was published.

(2) The date and time of publishment on the platform.

(3) The term "unknown creation process" if the digital content did not contain any previously applied provenance data at the time it was published on the platform.

(e) (1) Notwithstanding anything to the contrary in this section, for purposes of labeling and disclosing provenance data of sound recordings and music videos, a large online platform shall use a visual, not an audio, disclosure for sound recordings and music videos that contains all of the following:

(A) The artist.

(B) The track.

(C) The copyrightholder or licensor information.

(2) A large online platform shall comply with the visual disclosure requirement described in paragraph (1) to the extent that those sound recordings and music videos have not been solely generated by a GenAI system, extended or modified by a GenAI system without the authorization of the copyrightholder whose work has been modified or extended, or modified by a GenAI system to imitate or be readily identifiable as another person and that other person has not authorized the modification.

(f) This section shall not apply to any product, service, website, or application that provides predominantly non-user-generated video game, television, streaming, or movie experiences.

22949.90.4. (a) Beginning July 1, 2026, and annually thereafter, a large online platform shall produce a transparency report that identifies moderation of deceptive synthetic content on their platform.

(b) The report required by subdivision (a) may include assessments of the distribution of illegal generative AI-generated child sexual abuse materials, nonconsensual intimate imagery, disinformation related to elections or public health, or other instances where synthetic or potentially deceptive content was identified and removed by the platform.

22949.90.5. The Department of Technology may assess an administrative penalty pursuant to the following:

(a) If a violation of this chapter is intentional or is the result of grossly negligent conduct, a penalty of up to ~~five~~ *one* hundred thousand dollars ~~($500,000)~~ *($100,000)* for each violation.

(b) If a violation of this chapter is unintentional or is not the result of grossly negligent conduct, a penalty of up to ~~fifty~~ *twenty-five* thousand dollars ~~($50,000)~~ *($25,000)* for each violation.

22949.90.6. (a) The Digital Content Provenance Administrative Fund is hereby created in the State Treasury.

(b) All penalties collected by the Department of Technology under Section 22949.90.5 shall be deposited in the Digital Content Provenance Administrative Fund.

(c) Upon appropriation by the Legislature for this express purpose, moneys in the Digital Content Provenance Administrative Fund may be expended by the Department of Technology to administer this chapter.

22949.90.7. This chapter shall become operative on July 1, 2026.

22949.91. The provisions of this chapter are severable. If any provision of this chapter or its application is held invalid, that invalidity shall not affect other provisions or applications that can be given effect without the invalid provision or application.

O