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 Provenance, Authenticity and Watermarking 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.

Beginning February 1, 2025, this bill, the California Provenance, Authenticity and Watermarking Standards Act, would require a generative artificial intelligence (AI) ~~system provider, as defined, to take certain actions to assist in the disclosure of provenance data to mitigate harms caused by inauthentic content, including placing imperceptible and maximally indelible watermarks containing provenance data into content created by an AI system that the generative AI system provider makes available.~~ *provider to, among other things,*

place imperceptible and maximally indelible watermarks containing provenance data into synthetic content produced or significantly modified by a generative AI system that the provider makes available, as those terms are defined. The bill would require, within 24 hours of discovering a vulnerability or failure in ~~an~~ *a generative* AI system, a generative AI ~~system~~ provider to report the vulnerability or failure to the Department of Technology and to notify other generative AI ~~system~~ providers, as specified. The bill would also require a conversational AI system, as defined, to clearly and prominently disclose to users that the conversational AI system receives synthetic content.

Beginning March 1, 2025, this bill would require a large online platform, as defined, to, among other things, use labels to prominently disclose the provenance data found in watermarks or digital signatures in content distributed to users on its platforms, as specified. The bill would require a large online platform to ~~use state-of-the-art techniques, including, but not limited to, analysis of user behavioral signals indicating usage of synthetic content, to detect and label inauthentic text content that is uploaded or distributed by individual users or networks of users.~~ *require a user that uploads or distributes content on its platform to disclose whether the content is synthetic content, as specified.*

Beginning January 1, 2026, this bill would require newly manufactured digital cameras and recording devices sold, offered for sale, or distributed in California to offer users the option to place an authenticity watermark and provenance watermark in the content produced by that device. The bill would require the authenticity watermark and provenance watermark to be compatible with widely used industry ~~standards, as specified. If a digital camera or recording device purchased in California prior to January 1, 2026, is capable of receiving a software or firmware update that would enable a user to place an authenticity watermark and provenance watermark on the content created by the device, the bill would require the device's manufacturer to offer that update.~~ *standards. If technically feasible, the bill would require a camera and recording device manufacturer, as defined, to offer to a user of a digital camera or recording device purchased in California prior to January 1, 2026, a software or firmware update enabling the user to place an authenticity watermark and provenance watermark on the content created by the device.*

Beginning January 1, 2026, and annually thereafter, this bill would also require ~~specified entities~~ *generative AI providers and large online*

platforms to produce a Risk Assessment and Mitigation Report that assesses the risks posed ~~by synthetic content and the harms that have been or could be caused by synthetic content,~~ *and harms caused by synthetic content generated in their systems or hosted on their platforms,* as prescribed. *The bill would require the report to be audited by qualified, independent auditors who are required to assess and either validate or invalidate the claims made in the report, as specified.*

This bill would provide that a violation of its provisions may result in an administrative penalty, assessed by the department, of up to $1,000,000 or 5% of the violator's annual global revenue, whichever is greater. The bill would require the department to adopt regulations as necessary to implement and carry out the purposes of this act and to review and update those regulations as needed.

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 produce inauthentic images, audio, video, and text 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 GenAI content to be clearly disclosed and labeled.

(c) Failing to appropriately label GenAI content can skew election results, enable academic dishonesty, and erode trust in the online information ecosystem.

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

(e) The Legislature should push for the creation of tools that allow Californians to assess the authenticity of online content.

(f) The Legislature should require online platforms to label inauthentic 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 Provenance, Authenticity, and Watermarking Standards

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

(a) "AI red-teaming" means a structured testing effort to find flaws and vulnerabilities in an AI system, including, but not limited to, harmful or discriminatory outputs, unforeseen or undesirable system behaviors, limitations, or potential risks associated with misuse of the 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) "Authentic content" means images, videos, audio, or text created by human beings 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, removal of background noise in audio, and spelling or grammar corrections in text.

(d) "Conversational AI system" means chatbots and other audio- or video-based systems that can hold humanlike conversations through digital media, including, but not limited to, online calling, phone calling, video conferencing, messaging, application or web-based chat interfaces, or other conversational interfaces. Conversational AI systems include, but are not limited to, chatbots for customer service or entertainment purposes embedded in internet websites and applications.

(e) "Digital signature" means a digital method that allows a user to sign a piece of authentic or synthetic content with their name or device information, verifying that they created the content.

(f) "Generative AI hosting platform" means an online repository or other internet website that makes generative AI systems available for download.

(g) "Generative AI provider" means an organization or individual that creates, codes, substantially modifies, or otherwise produces a generative AI system.

(h) "Generative AI system" means an artificial intelligence system that generates derived synthetic content, including images, videos, audio, text, and other digital content.

(i) "Inauthentic content" means synthetic content that is so similar to authentic content that it could be mistaken as authentic.

(j) "Large online platform" means a public-facing internet website, web application, or digital application, including a social network, video sharing platform, messaging platform, advertising network, or search engine that had at least 1,000,000 California users during the preceding 12 months.

(k) "Maximally indelible watermark" means a watermark that is designed to be as difficult to remove as possible using state-of-the-art techniques and relevant industry standards.

(l) "Provenance data" means data that identifies the origins of synthetic content, including, but not limited to, the following:

(1) The name of the generative AI provider.

(2) The name and version number of the AI system that generated the content.

(3) The time and date of the creation.

(4) The portions of content that are synthetic.

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

(n) "Watermark" means information that is embedded into a generative AI system's output for the purpose of conveying its synthetic nature, identity, provenance, history of modifications, or history of conveyance.

(o) "Watermark decoders" means freely available software tools or online services that can read or interpret watermarks and output the provenance data embedded in them.

22949.90.1. (a) A generative AI provider shall do all of the following:

(1) Place imperceptible and maximally indelible watermarks containing provenance data into synthetic content produced or significantly modified by a generative AI system that the provider makes available.

(A) If a sample of synthetic content is too small to contain the required provenance data, the provider shall, at minimum, attempt to embed watermarking information that identifies the content as synthetic and provide the following provenance information in

order of priority, with clause (i) being the most important, and clause (iv) being the least important:

(i) The name of the generative AI provider.

(ii) The name and version number of the AI system that generated the content.

(iii) The time and date of the creation of the content.

(iv) If applicable, the specific portions of the content that are synthetic.

(B) To the greatest extent possible, watermarks shall be designed to retain information that identifies content as synthetic and gives the name of the provider in the event that a sample of synthetic content is corrupted, downscaled, cropped, or otherwise damaged.

(2) Develop downloadable watermark decoders that allow a user to determine whether a piece of content was created with the provider's system, and make those tools available to the public.

(A) The watermark decoders shall be easy to use by individuals seeking to quickly assess the provenance of a single piece of content.

(B) The watermark decoders shall adhere, to the greatest extent possible, to relevant national or international standards.

(3) Conduct AI red-teaming exercises involving third-party experts to test whether watermarks can be easily removed from synthetic content produced by the provider's generative AI systems, as well as whether the provider's generative AI systems can be used to falsely add watermarks to otherwise authentic content. Red-teaming exercises shall be conducted before the release of any new generative AI system and annually thereafter.

(A) If a provider allows their generative AI systems to be downloaded and modified, the provider shall additionally conduct AI red-teaming to assess whether their systems' watermarking functionalities can be disabled.

(B) A provider shall make summaries of its AI red-teaming exercises publicly available in a location linked from the home page of the provider's internet website, using a clearly labeled link that has a similar look, feel, and size relative to other links on the same web page. The provider shall remove from the summaries any details that pose an immediate risk to public safety.

(C) A provider shall submit full reports of its AI red-teaming exercises to the Department of Technology within six months of conducting a red-teaming exercise pursuant to this section.

(b) A generative AI provider may continue to make available a generative AI system that was made available before the date upon which this act takes effect and that does not have watermarking capabilities as described by paragraph (1) of subdivision (a), if either of the following conditions are met:

(1) The provider is able to retroactively create and make publicly available a decoder that accurately determines whether a given piece of content was produced by the provider's system with at least 99 percent accuracy as measured by an independent auditor.

(2) The provider conducts and publishes research to definitively demonstrate that the system is not capable of producing inauthentic content.

(c) Providers and distributors of software and online services shall not make available a system, application, tool, or service that is designed to remove watermarks from synthetic content.

(d) Generative AI hosting platforms shall not make available a generative AI system that does not place maximally indelible watermarks containing provenance data into content created by the system.

(e) (1) Within 24 hours of discovering a vulnerability or failure in a generative AI system, a generative AI provider shall report the vulnerability or failure to the Department of Technology.

(A) A provider shall notify other generative AI providers that may be affected by similar vulnerabilities or failures in a manner that allows the other generative AI provider to harden their own AI systems against similar risks.

(B) A provider shall notify affected parties, including, but not limited to, online platforms, researchers or users who received incorrect results from a watermark decoder, or users who produced AI content that contained incorrect or insufficient watermarking data. A provider shall not be required to notify an affected party whose contact information the provider has not previously collected or retained.

(2) A provider shall make any report to the Department of Technology under this subdivision publicly available in a location linked from the home page of the provider's internet website with a clearly labeled link that has a similar look, feel, and size relative to other links on the same web page. If public disclosure of the

report under this subdivision could pose public safety risks, a provider may instead do either of the following:

(A) Post a summary disclosure of the reported vulnerability or failure.

(B) Delay, for no longer than 30 days, the public disclosure of the report until the public safety risks have been mitigated. If a provider delays public disclosure, they shall document their efforts to resolve the vulnerability or failure as quickly as possible in order to meet the reporting requirements under this subdivision.

(f) (1) A conversational AI system shall clearly and prominently disclose to users that the conversational AI system generates synthetic content.

(A) In visual interfaces, including, but not limited to, text chats or video calling, a conversational AI system shall place the disclosure required under this subdivision in the interface itself and maintain the disclosure's visibility in a prominent location throughout any interaction with the interface.

(B) In audio-only interfaces, including, but not limited to, phone or other voice calling systems, a conversational AI system shall verbally make the disclosure required under this subdivision at the beginning and end of a call.

(2) In all conversational interfaces of a conversational AI system, the conversational AI system shall, at the beginning of a user's interaction with the system, obtain a user's affirmative consent acknowledging that the user has been informed that they are interacting with a conversational AI system. A conversational AI system shall obtain a user's affirmative consent prior to beginning the conversation.

(3) Disclosures and affirmative consent opportunities shall be made available to a user in the language in which the conversational AI system is communicating with the user.

(4) The requirements under this subdivision shall not apply to conversational AI systems that do not produce inauthentic content.

(g) This section shall become operative on February 1, 2025.

22949.90.2. (a) For purposes of this section, the following definitions apply:

(1) "Authenticity watermark" means a watermark of authentic content that includes the name of the device manufacturer.

(2) "Camera and recording device manufacturer" means the makers of a device that can record photographic, audio, or video

content, including, but not limited to, video and still photography cameras, mobile phones with built-in cameras or microphones, and voice recorders.

(3) "Provenance watermark" means a watermark of authentic content that includes details about the content, including, but not limited to, the time and date of production, the name of the user, details about the device, and a digital signature.

(b) (1) Beginning January 1, 2026, newly manufactured digital cameras and recording devices sold, offered for sale, or distributed in California shall offer users the option to place an authenticity watermark and provenance watermark in the content produced by that device.

(2) A user shall have the option to remove the authenticity and provenance watermarks from the content produced by their device.

(3) Authenticity watermarks shall be turned on by default, while provenance watermarks shall be turned off by default.

(4) Newly manufactured digital cameras and recording devices subject to the requirements of this subdivision shall clearly inform a user of the existence of the authenticity and provenance watermarks settings upon the user's first use of the camera or the recording function on the recording device.

(A) When a camera or audio recording application is open, a newly manufactured digital camera or recording device shall have a clear indicator that a watermark is being applied.

(B) A newly manufactured digital camera or recording device shall allow the user to adjust the watermarks settings.

(5) Authenticity and provenance watermarks shall, if enabled, be applied to authentic content produced using third-party applications that bypass default camera or recording applications in order to offer camera or audio recording functionalities.

(c) The authenticity watermark and provenance watermark, as required by subdivision (b), shall be compatible with widely used industry standards.

(d) Beginning January 1, 2026, a camera and recording device manufacturer shall offer a software or firmware update enabling a user to place an authenticity watermark and provenance watermark on the content created by the device to a user of a digital camera or recording device purchased in California prior to January 1, 2026, if technically feasible.

22949.90.3. (a) Beginning March 1, 2025, a large online platform shall use labels to prominently disclose the provenance data found in watermarks or digital signatures in content distributed to users on its platforms.

(1) The labels shall indicate whether content is fully synthetic, partially synthetic, authentic, authentic with minor modifications, or does not contain a watermark.

(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. A disclosure in audio content shall occur at the beginning and end of a piece of content and shall be presented in a prominent manner and at a comparable volume and speaking cadence as other spoken words in the content. A disclosure in video content should be legible for the full duration of the video.

(c) A large online platform shall use state-of-the-art techniques to detect and label synthetic content that has had watermarks removed or that was produced by generative AI systems without watermarking functionality.

(d) (1) A large online platform shall require a user that uploads or distributes content on its platform to disclose whether the content is synthetic content.

(2) A large online platform shall include prominent warnings to users that uploading or distributing synthetic content without disclosing that it is synthetic content may result in disciplinary action.

(3) A large online platform may provide users with an option to indicate that the user is uncertain whether the content they are uploading or distributing is synthetic content. If a user uploads or distributes content and indicates that they are uncertain of whether the content is synthetic content, a large online platform shall indicate that the uploaded or distributed content is possibly synthetic and shall display that indication in a manner that is visible or audible to viewers or listeners of the content.

(e) A large online platform shall use state-of-the-art techniques to detect and label text-based inauthentic content that is uploaded by users.

(f) A large online platform shall make accessible a verification process for users to apply a digital signature to authentic content. The verification process shall include options that do not require disclosure of personal identifiable information.

22949.90.4. (a) (1) Beginning January 1, 2026, and annually thereafter, generative AI providers and large online platforms shall produce a Risk Assessment and Mitigation Report that assesses the risks posed and harms caused by synthetic content generated by their systems or hosted on their platforms.

(2) The report shall include, but not be limited to, assessments of the distribution of AI-generated child sexual abuse materials, nonconsensual intimate imagery, disinformation related to elections or public health, plagiarism, or other instances where synthetic or inauthentic content caused or may have the potential to cause harm.

(b) The report required under subdivision (a) shall be audited by qualified, independent auditors who shall assess and either validate or invalidate the claims made in the report. Auditors shall use state-of-the-art techniques to assess reports, and shall adhere to relevant national and international standards.

22949.90.5. A violation of this chapter may result in an administrative penalty, assessed by the Department of Technology, of up to one million dollars ($1,000,000) or 5 percent of the violator's annual global revenue, whichever is greater.

22949.90.6. Within 90 days of the date upon which this act takes effect, the Department of Technology shall adopt regulations to implement and carry out the purposes of this chapter. The department shall review and update its regulations relating to the implementation of this chapter as needed, including, but not limited to, adopting specific national or international standards for provenance, authenticity, watermarking, and digital signatures, as long as the standards do not weaken the provisions of this chapter.

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.

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

~~(a) In light of the widespread adoption of generative artificial intelligence (AI) technologies that are increasingly able to generate inauthentic images, audio, video, and text content, sometimes called "deepfakes," it is increasingly important that the provenance of this content be clearly disclosed and labeled in ways that could both prevent harms, reduce the severity of harms, and also make it more difficult and costly for bad actors to cause these harms.~~

~~(b) (1) The harms caused by inauthentic content that is presented as authentic span a wide gamut of fields. These harms can be costly and deeply damaging to individuals and society. Some prominent categories of harm include scams and fraud, child sexual abuse material (CSAM) and nonconsensual intimate imagery (NCII), disinformation, and plagiarism and academic integrity.~~

~~(A) The Federal Trade Commission reports that people lost $8.8 billion to scams in 2022. This is only expected to increase with broad access to generative AI tools, with one study already finding a 1,265-percent increase in phishing scam emails since the fourth quarter of 2022.~~

~~(B) A recent study found that a version of one of the most popular AI image generator tools was trained on CSAM, which facilitates production in derivative models of both CSAM and NCII of children. Another recent study found 34 different AI "undressing" tools that produce realistic naked images based on clothed photographs of individuals. Many of these tools only work on women. One recent case involved a high school where 30 different teenage girls were victims of these tools.~~

~~(C) From politics to public safety, disinformation can have a multiplicity of negative impacts on society. It can skew election results, as may have been the case with a convincing audio deepfake in Slovakia. In the lead up to the 2024 United States elections, we have already seen candidate- and party-produced advertisements using deepfake audio and video content. A deepfake of an explosion at the Pentagon caused a brief but significant stock market dip.~~

~~(D) Educational institutions have struggled to adopt clear best practices for assigning students at-home writing assignments since the widespread use of high-quality AI text generation tools.~~

~~(2) More broadly, erosion of trust in the information ecosystem, sometimes referred to as "truth decay," can be shown to increase polarization and stands to be greatly exacerbated by our ongoing~~

failure to establish norms for clearly disclosing and labeling the provenance of digital media.

(c) (1) For the above-described reasons, this act provides for the phased introduction of the California Provenance, Authenticity and Watermarking Standards (PAWS), which aim to provide the public with tools to understand how the content that they see across digital media was produced and if it is, in fact, authentic.

(2) The PAWS require all producers of media to embed maximally indelible and privacy-preserving content provenance data into the content that they generate, whether AI generated or authentic. The PAWS also require all large online platforms to display clearly understandable labels on content that alert users to its provenance, or to the absence of available provenance data.

(3) The provenance data required by PAWS is narrowly tailored to provide consumers with factual information about the authenticity or inauthenticity of images, audio, video, or text content in order to prevent consumer deception.

(4) While the PAWS alone will not fully eliminate the harms described above, it has the capacity to dramatically reduce these harms by signaling to the would-be audiences for and victims of these harms that inauthentic content will not be easily mistaken for authentic content. The PAWS will also significantly reduce the burden on large online platforms that would like to display content provenance data to their users, but are not able to do so because of the lack of an industrywide standard for embedding of provenance data in their content.

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 PROVENANCE, AUTHENTICITY AND WATERMARKING STANDARDS

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

(a) "AI red-teaming" means a structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of artificial intelligence (AI) and is most often performed by dedicated "red teams" that adopt adversarial methods to identify flaws and vulnerabilities, including, but not limited to, harmful or

~~discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with misuse of the system.~~

~~(b) "AI system" means a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs, including, but not limited to, predictions, content, recommendations, or decisions that may influence physical or virtual environments.~~

~~(c) "Authentic content" means images, videos, audio clips, or text created by human beings 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, removal of background noise in audio, and spelling or grammar corrections in text.~~

~~(d) "Camera and recording device manufacturers" means the makers of a device that can record photographs, audio, or video content, including, but not limited to, video and still photography cameras, mobile phones with built-in cameras or microphones, and voice recorders.~~

~~(e) "Conversational AI systems" means chatbots and other audio- or video-based systems that can hold humanlike conversations through digital media, including, but not limited to, online calling, phone calling, video conferencing, messaging, application or web-based chat interfaces, or other conversational interfaces. Conversational AI systems includes chatbots for customer service or entertainment purposes embedded in internet websites and applications.~~

~~(f) "Digital signature" means a digital method that allows a user to sign a piece of authentic or synthetic content with their name or device information, verifying that they created the content, and that is included in the content's provenance data.~~

~~(g) "Generative AI system" means the class of AI models that emulates the structure and characteristics of input data to generate derived synthetic content, including images, videos, audio, text, and other digital content. The synthetic content generated may, but does not necessarily have to, be of the same content type as the input data.~~

~~(h) "Generative AI system distributors" means organizations or individuals that distribute generative AI systems, or substantial~~

~~components thereof, such as model weights, in ways that can be downloaded and used by individuals locally on their own hardware, or modified or incorporated into other products or services.~~

~~(i) "Generative AI system providers" means organizations or individuals that make AI systems, or substantial components thereof, available on the market, put them into service, provide them as standalone models or embed them in other systems or products, or provide them under free and open source licenses as a service, or through other large online platforms. Generative AI system distributors include repositories or hosting internet websites that make AI systems available for download, even if those repositories are not the original makers of the AI systems they make available, and providers of conversational AI systems.~~

~~(j) "Inauthentic content" means synthetic content that is so similar to authentic content that it could be mistaken as authentic.~~

~~(k) "Large online platform" means a public-facing internet website, web application, or digital application, including a social network, video sharing platform, messaging platform, advertising network, or search engine that had at least 1,000,000 California users during the preceding 12 months.~~

~~(*l*) "Legacy generative AI systems" means generative AI systems created and released before the date upon which this act takes effect.~~

~~(m) "Maximally indelible watermarks" means watermarks that are as difficult to remove as is possible with currently available techniques and that are tested through AI red-teaming.~~

~~(n) "Provenance data" means data that includes the name of the AI system that generated the content, the underlying AI models that were part of the AI system, the time and date of the creation of the content, and, if applicable, which specific portions of the content are synthetic.~~

~~(o) "Synthetic content" means information, including images, videos, audio clips, and text, that has been significantly modified or generated by algorithms, including by AI.~~

~~(p) "Tamper-evident" means a type of watermark that contains provenance data that cannot be modified without leaving evidence of tampering.~~

~~(q) "Watermark" means embedded information, typically difficult to remove, into outputs created, including into photographs, videos, audio clips, or text, for the purposes of~~

~~verifying the authenticity of the output or the identity or characteristics of its provenance, modifications, or conveyance.~~

~~(r) "Watermark decoders" means freely available software tools or online services that can read or interpret watermarks and output the provenance data embedded in them.~~

~~22949.90.1. (a) A generative AI system provider shall do all of the following:~~

~~(1) Place imperceptible and maximally indelible watermarks containing provenance data into content created by an AI system that the generative AI system provider makes available. If the content is manipulated in ways that are intended to remove a watermark, or if a content sample, including a cropped portion of an image, a truncated segment of a generated piece of audio or video content, or a small amount of text content, is damaged or becomes too small to contain the required provenance data, the generative AI system provider shall, at minimum, embed watermarking information that identifies the content as synthetic and gives the name of the generative AI system provider.~~

~~(2) Provide downloadable software tools or online services to determine whether a piece of content was created with the provider's system and make those tools available to all large online platforms and the public.~~

~~(A) A generative AI system provider shall produce the software tools or online services in ways that are easy to use manually by individuals seeking to quickly assess the provenance of a single piece of content and to use in an efficient and automated fashion by online platforms or researchers seeking to assess the provenance of hundreds or thousands of pieces of content per minute.~~

~~(B) A generative AI system provider shall produce the software tools or online services to be interoperable to the greatest extent possible with decoders made available by other providers or organizations. The decoders shall adhere, to the greatest extent possible, to relevant national or international standards, if available.~~

~~(3) Conduct AI red-teaming exercises that involve third-party experts to test whether watermarks can be easily removed from their synthetic content and whether AI systems could be used to falsely add watermarks to authentic content that indicates that it is inauthentic.~~

~~(A) If a generative AI system provider intends to allow their systems to be downloaded and modified, the provider shall conduct~~

~~AI red-teaming to assess whether the systems' watermarking functionalities can be disabled to generate deceptive, inauthentic content.~~

~~(B) A generative AI system provider shall make summaries of its AI red-teaming exercises publicly available in a location linked from the home page of the generative AI system provider's internet website with a clearly labeled link that has a similar look, feel, and size relative to other links on the same web page and shall remove from the summaries details that may pose immediate public security. A generative AI system provider shall submit full reports of its AI red-teaming exercises to the Department of Technology within six months from the date upon which this act takes effect, and annually thereafter.~~

~~(b) A generative AI system provider may continue providing legacy generative AI systems that do not have watermarking capabilities as described by paragraph (1) of subdivision (a) if either of the following applies:~~

~~(1) The generative AI system provider is able to retrospectively create and make publicly available a decoder that accurately determines whether content generated by the legacy generative AI system is synthetic with at least 99 percent accuracy as measured by an independent auditor.~~

~~(2) The generative AI system provider conducts and publishes research to definitively demonstrate that the legacy generative AI system is not capable of producing inauthentic content of a quality that is sufficiently realistic to be potentially mistaken for authentic content.~~

~~(c) A generative AI system provider, generative AI system distributor, or other provider, or a distributor of software or online services shall not make available or distribute an AI system, application, tool, or service that has the capacity to remove watermarks from synthetic content or an AI system that has the capacity to remove watermarks from synthetic content but was not explicitly designed for that purpose.~~

~~(d) (1) If a generative AI system provider distributes a generative AI system in a way that allows the generative AI system to be modified by others, the generative AI system provider shall ensure that the generative AI system does not allow for removal of the system's watermarking functionality.~~

~~(2) A generative AI system provider shall not distribute a generative AI system or make a generative AI system available for use if the system's watermarking functionality can be removed by others.~~

~~(e) (1) Within 24 hours of discovering a vulnerability or failure in an AI system, a generative AI system provider shall report the vulnerability or failure to the Department of Technology. A generative AI system provider shall also notify other generative AI system providers that may be affected by similar vulnerabilities or failures in a manner that allows the other generative AI system provider to harden their own AI systems against similar risks.~~

~~(2) A generative AI system provider shall also notify any affected party. Affected parties include, but are not limited to, an online platform, researchers or users who received incorrect results from a watermark decoder, or users who produced AI content that contained incorrect or insufficient watermarking data.~~

~~(3) A generative AI system provider shall make any report to the Department of Technology under this subdivision publicly available in a location linked from the home page of the generative AI system provider's internet website with a clearly labeled link that has a similar look, feel, and size relative to other links on the same web page.~~

~~(4) If public disclosure of the report under this subdivision could pose public safety risks, a generative AI system provider may instead do either of the following:~~

~~(A) Post a summary disclosure of the reported vulnerability or failure.~~

~~(B) Delay, for no longer than 30 days, the public disclosure of the report until the public safety risks have been mitigated. If a generative AI system provider delays public disclosure, the provider shall document and demonstrate that they moved as quickly as possible to resolve the vulnerability or failure and meet the reporting requirements under this subdivision.~~

~~(f) (1) A conversational AI system shall clearly and prominently disclose to users that the conversational AI system receives synthetic content.~~

~~(A) In visual interfaces, including, but not limited to, text chats or video calling, a conversational AI system shall place the disclosure required under this subdivision in the interface itself~~

~~and maintain the disclosure's visibility in a prominent location throughout any interaction with the interface.~~

~~(B) In audio-only interfaces, including, but not limited to, phone or other voice calling systems, a conversational AI system shall verbally make the disclosure required under this subdivision at the beginning and end of a call.~~

~~(2) In all conversational interfaces of a conversational AI system, the conversational AI system shall, at the beginning of a user's interaction with the system, obtain a user's affirmative consent acknowledging that the user has been informed that they are interacting with a conversational AI system. A conversational AI system shall obtain a user's affirmative consent prior to beginning the conversation.~~

~~(3) Disclosures and affirmative consent opportunities shall be made available to a user in the language in which the conversational AI system is communicating with the user.~~

~~(4) The requirements under this subdivision shall not apply to conversational AI systems that produce content that could not be reasonably mistaken as authentic.~~

~~(g) This section shall become operative on February 1, 2025.~~

~~22949.90.2. (a) For purposes of this section, the following definitions apply:~~

~~(1) "Authenticity watermark" means a watermark of authentic content that includes the name of the device manufacturer.~~

~~(2) "Provenance watermark" means a watermark of authentic content that includes details about the content, including, but not limited to, the time and date of production, the name of the user, details about the device, and a digital signature.~~

~~(b) (1) Beginning January 1, 2026, newly manufactured digital cameras and recording devices sold, offered for sale, or distributed in California shall offer users the option to place an authenticity watermark and provenance watermark in the content produced by that device.~~

~~(2) A user shall have the option to remove the authenticity and provenance watermarks from the content produced by their device.~~

~~(3) Authenticity watermarks shall be turned on by default, while provenance watermarks shall be turned off by default.~~

~~(4) Newly manufactured digital cameras and recording devices subject to the requirements of this subdivision shall clearly inform a user of the existence of the authenticity and provenance~~

~~watermarks settings upon the user's first use of the camera or the recording function on the recording device.~~

~~(A) When a camera or audio recording application is open, a newly manufactured digital camera or recording device shall have a clear indicator that a watermark is being applied.~~

~~(B) A newly manufactured digital camera or recording device shall allow the user to adjust the watermarks settings.~~

~~(5) The authenticity watermark and provenance watermark shall persist in content produced using newly manufactured digital cameras or recording devices even if the content is created or recorded in third-party applications that offer camera or audio recording functionalities.~~

~~(c) The authenticity watermark and provenance watermark, as required in subdivision (b), shall be compatible with widely used industry standards, including the Coalition for Content Provenance and Authenticity's content credentials.~~

~~(d) Beginning January 1, 2026, a camera and recording device manufacturer shall offer a software or firmware update enabling a user to place an authenticity watermark and provenance watermark on the content created by the device to a user of a digital camera or recording device purchased in California prior to January 1, 2026. This requirement shall only apply if a digital camera or recording device purchased in California prior to January 1, 2026, is capable of receiving a software or firmware update that would enable the user to place an authenticity watermark and provenance watermark on the content created by the device.~~

~~22949.90.3. (a) Beginning March 1, 2025, a large online platform shall use labels to prominently disclose the provenance data found in watermarks or digital signatures in content distributed to users on its platforms by making use of the data contained in watermarks and digital signatures embedded in content using widely used industry standards, including the Coalition for Content Provenance and Authenticity's content credentials, and synthetic content decoders provided by generative AI system providers.~~

~~(1) The labels shall indicate whether content is fully synthetic, partially synthetic, authentic, authentic with minor modifications, or does not contain a watermark.~~

~~(2) A user shall be allowed to click or tap on a label to inspect all available 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. A disclosure in audio and video content shall occur at the beginning and end of a piece of content and shall be presented in a prominent manner and at a comparable volume and speaking cadence as other spoken words in the content.~~

~~(c) A large online platform shall use state-of-the-art techniques to detect and label synthetic content that has had watermarks removed or was produced by AI systems without watermarking functionality.~~

~~(d) (1) A large online platform shall require a user that uploads or distributes content on its platform to disclose whether the uploaded or distributed content is synthetic content.~~

~~(2) A large online platform shall include prominent warnings to users that uploading or distributing synthetic content without disclosing that it is synthetic content is not permissible and will result in disciplinary action by the large online platform.~~

~~(3) A large online platform may provide users with an option to indicate that the user is uncertain whether the content they are uploading or distributing is synthetic content. If a user uploads or distributes content and indicates that they are uncertain of whether the content is synthetic content, a large online platform shall indicate that the uploaded or distributed content is possibly synthetic and shall display that indication in a manner that is visible or audible to viewers or listeners of the content.~~

~~(e) (1) A large online platform shall use state-of-the-art techniques to detect and label inauthentic text content that is uploaded or distributed by individual users or networks of users.~~

~~(2) A large online platform may use a variety of methods to detect inauthentic text content, including, but not limited to, the following:~~

~~(A) Bulk analysis of collected text content from users or networks of users.~~

~~(B) Analysis of user behavioral signals indicating usage of synthetic content.~~

~~(C) Assessing large quantities of text generated by users or networks of users for watermarked content.~~

~~(D) Considering whether a user's typing cadence indicates authenticity or automation.~~

~~(E) Verification that users are matched to unique device identifications such as a subscriber identity module (SIM) card, international mobile equipment identity (IMEI), or multifactor authentication (MFA).~~

~~(3) A large online platform shall also consider account age, login frequency, connection to other identity verification services, frequency of content uploading or distributing, authenticity of original media content, and other on-platform behaviors by a user that could be used to detect undisclosed inauthentic content production.~~

~~(4) If a large online platform discovers that a user has uploaded or distributed inauthentic content and the user did not disclose that the uploaded or distributed content is synthetic content pursuant to subdivision (d), the large online platform shall disable the account of the user that uploaded or distributed the undisclosed inauthentic content.~~

~~(f) A large online platform shall make accessible a verification process for users to apply a digital signature to content created by a human being. The verification process shall include options to verify in a variety of methods that do not necessarily require disclosure of personal identifiable information, including, but not limited to, uploading a government-issued identification and matching picture identification or verifying that a user possesses a unique device with a SIM card and active phone number.~~

~~22949.90.4. (a) (1) Beginning January 1, 2026, and annually thereafter, generative AI system providers, generative AI system distributors, and large online platforms shall produce a Risk Assessment and Mitigation Report that assesses the risks posed by synthetic content and the harms that have been or could be caused by synthetic content.~~

~~(2) The report shall include, but is not limited to, assessments on the distribution of AI-generated child sexual abuse materials, nonconsensual intimate imagery, disinformation related to elections or public health, plagiarism, or other instances where synthetic or inauthentic content caused or may have the potential to cause harm.~~

~~(3) The report shall be audited by qualified, independent auditors who shall assess and either validate or invalidate the claims made in the report. An auditor shall assess the report by using state-of-the-art techniques and adhering to national and~~

~~international standards for the auditing of AI systems as they become available.~~

~~(b) The Department of Technology may authorize independent researchers associated with educational institutions or civil society organizations approved by the Department of Technology to access special researcher tools designed to facilitate large-scale and efficient analysis of content and application programming interfaces (APIs) from generative AI system providers and large online platforms for the purposes of generating test content, studying the efficacy of labeling and effects on users, and evaluating the overall effectiveness of this chapter in preventing harms caused by inauthentic content.~~

~~22949.90.5. A violation of this chapter may result in an administrative penalty, assessed by the Department of Technology, of up to one million dollars ($1,000,000) or 5 percent of the violator's annual global revenue, whichever is greater.~~

~~22949.90.6. Within 90 days of the date upon which this act takes effect, the Department of Technology shall adopt regulations as necessary to implement and carry out the purposes of this chapter. The department shall review and update its regulations relating to the implementation of this chapter as needed, including, but not limited to, adopting specific national or international standards for provenance, authenticity, watermarking, and digital signatures, so long as the standards do not weaken the provisions of this chapter.~~

~~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