SENATE JUDICIARY COMMITTEE
Senator Thomas Umberg, Chair
2023-2024 Regular Session

AB 3211 (Wicks)
Version: June 10, 2024
Hearing Date: June 18, 2024
Fiscal: Yes
Urgency: No
CK

SUBJECT

California Provenance, Authenticity and Watermarking Standards

DIGEST

This bill establishes the California Provenance, Authenticity and Watermarking Standards Act, which requires a generative AI (GenAI) system provider to, among other things, 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 its systems. The bill requires 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, and to use state-of-the-art techniques to detect and label inauthentic text content that is uploaded or distributed by users. The bill requires GenAI providers and platforms to produce annual impact assessment reports. The bill also requires recording device manufacturers to enable options for embedding provenance data into recordings.

EXECUTIVE SUMMARY

Certain forms of media – audio recordings, video recordings, and still images – can be powerful evidence of the truth. While such media have always been susceptible to some degree of manipulation, fakes were relatively easy to detect. The rapid advancement of AI technology, specifically the wide-scale introduction of GenAI models, has made it drastically cheaper and easier to produce synthetic content, audio, images, text, and video recordings that are not real, but that are so realistic that they are virtually impossible to distinguish from authentic content, including so-called "deepfakes."

Among other things, this bill works to ensure that providers of these GenAI systems are equipping consumers with a tool to identify when specific content has been generated by their systems, places requirements on large online platforms to disclose the provenance data of content and to detect and label inauthentic content, and requires recording device manufacturers to allow for the embedding of watermarks on recordings. This bill is sponsored by the California Initiative for Technology & Democracy and supported by a wide variety of groups, including SEIU California and NextGen CA. The bill is opposed by Oakland Privacy and a coalition of industry groups, including Netchoice. Should this pass out of this Committee, it will next be heard in the Senate Governmental Organization Committee.

PROPOSED CHANGES TO THE LAW

Existing law:

1) Defines "deepfake" as audio or visual content that has been generated or manipulated by artificial intelligence which would falsely appear to be authentic or truthful and which features depictions of people appearing to say or do things they did not say or do without their consent. (Gov't Code § 11547.5.)

2) Defines "digital content forgery" as the use of technologies, including artificial intelligence and machine learning techniques, to fabricate or manipulate audio, visual, or text content with the intent to mislead. (Gov't Code § 11547.5.)

3) Defines "digital content provenance" as the verifiable chronology of the original piece of digital content, such as an image, video, audio recording, or electronic document. (Gov't Code § 11547.5.)

4) Requires, upon appropriation by the Legislature, the Secretary of Government Operations Agency (GovOps) to evaluate the following:
   a) The impact of the proliferation of deepfakes on state government, California-based businesses, and residents of the state.
   b) The risks, including privacy risks, associated with the deployment of digital content forgery technologies and deepfakes on state and local government, California-based businesses, and residents of the state.
   c) Potential privacy impacts of technologies allowing public verification of digital content provenance.
   d) The impact of digital content forgery technologies and deepfakes on civic engagement, including voters.
   e) The legal implications associated with the use of digital content forgery technologies, deepfakes, and technologies allowing public verification of digital content provenance.
   f) The best practices for preventing digital content forgery and deepfake technology to benefit the state, California-based businesses, and California residents, including exploring whether and how the adoption of a digital content provenance standard could assist with reducing the proliferation of digital content forgeries and deepfakes. (Gov't Code § 11547.5(b).)

5) Requires the Secretary of GovOps to develop a coordinated plan to accomplish all of the following:
    a) Investigate the feasibility of, and obstacles to, developing standards and technologies for state departments for determining digital content provenance.
    b) Increase the ability of internet companies, journalists, watchdog organizations, other relevant entities, and members of the public to meaningfully scrutinize and identify digital content forgeries and relay trust and information about digital content provenance to content consumers.
    c) Develop or identify mechanisms for content creators to cryptographically certify authenticity of original media and nondeceptive manipulations.
    d) Develop or identify mechanisms for content creators to enable the public to validate the authenticity of original media and nondeceptive manipulations to establish digital content provenance without materially compromising personal privacy or civil liberties. (Gov't Code § 11547.5(c).)

6) Establishes the California Department of Technology (CDT) within GovOps, under the supervision of the Director of Technology (Director), also known as the State Chief Information Officer. (Gov. Code Sec. 11545(a).)

7) Provides that the duties of the Director include:

    a) Advising the Governor on the strategic management and direction of the state's information technology (IT) resources.
    b) Establishing and enforcing state IT strategic plans, policies, standards, and enterprise architecture, as specified.
    c) Minimizing overlap, redundancy, and cost in state IT operations by promoting the efficient and effective use of information technology.
    d) Providing technology direction to agency and department chief information officers to ensure the integration of statewide technology initiatives, compliance with IT policies and standards, and the promotion of the alignment and effective management of IT services.
    e) Working to improve organizational maturity and capacity in the effective management of IT; and establishing performance management and improvement processes to ensure state IT systems and services are efficient and effective. (Gov. Code § 11545(b).)

This bill:

1) Defines the relevant terms, including:
    a) "Digital fingerprint" means a unique value that can be used to identify identical or similar digital content.

b) "Digital signature" means a method based on cryptography that allows a user or entity to digitally sign content with provenance data in order to verify that the user or entity participated in the creation of the content.
c) "Generative AI system" means an artificial intelligence system that generates derived synthetic content, including images, videos, audio, text, and other digital content.
d) "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 and can facilitate the sharing of synthetic content.
e) "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.
f) ″Nonsynthetic 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 content 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.
g) "Potentially deceptive content" means synthetic content that is so similar to nonsynthetic content that it could reasonably be mistaken as nonsynthetic content.
h) "Provenance data" means information about the history of the content, including, but not limited to, the following:
    i. The name of the generative AI provider or the camera or recording device manufacturer.
    ii. The name and version number of the AI system that generated the content or the operating system, version of the operating system, or the application used to capture, create, or record the content.
    iii. The time and date of the content's creation and any additional modifications of the content.
    iv. The portions of content that have been changed by a generative AI system, if applicable.
i) "Synthetic content" means information, including images, videos, audio, and text, that has been produced or significantly modified by a generative AI system.
j) "Watermark" means information that is embedded into content for the purpose of communicating the provenance, history of modification, or history of conveyance.

2) Requires a GenAI provider to place an imperceptible and maximally indelible watermark into synthetic content produced or significantly modified by the provider's GenAI system, as provided. To the greatest extent possible,

watermarks shall be designed to communicate information that identifies content as synthetic and identifies the provider in the event that a sample of synthetic content is corrupted, downscaled, cropped, or otherwise damaged.

3) Requires a GenAI provider to make available to the public a watermark decoder that is easy to use and adheres to relevant national or international standards.

4) Requires a GenAI provider to conduct AI red-teaming exercises involving third-party experts to test whether watermarks can be easily removed or whether the provider's GenAI systems can be used to falsely add watermarks to otherwise nonsynthetic content. Red-teaming exercises shall be conducted before the release of any new Gen AI system and annually thereafter, as provided. A provider shall make summaries of these exercises publicly available, as specified, and full reports are to be sent to CDT. Details that pose an immediate risk to public safety or provide information that could be used to disable or circumvent the functionality of watermarks must be removed.

5) Requires a GenAI system capable of producing potentially deceptive content to generate and store, in a searchable online database in a manner that can be retrieved by a viewer of the content, a digital fingerprint of and provenance data for any piece of potentially deceptive content that they produce. This provenance shall not include personally identifiable information.

6) Prohibits providers and distributors of software and online services from making available a system, application, tool, or service that is designed to remove watermarks from synthetic content.

7) Prohibits a GenAI hosting platform from making available a GenAI system that does not place maximally indelible watermarks, as required herein.

8) Requires a GenAI provider to, within 96 hours of discovering a material vulnerability or failure in a GenAI system related to the erroneous or malicious inclusion or removal of provenance information or watermarks, report the vulnerability or failure to CDT, as provided. Providers must also notify other providers that may be affected by similar vulnerabilities or failures.

9) Requires a provider to notify other GenAI providers that may be affected by similar vulnerabilities or failures in a manner that allows the other provider to harden their own AI systems against similar risks, but that does not compromise the reporting provider's systems or disclose the reporting provider's confidential or proprietary information. A provider must also use commercially reasonable efforts to 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

provenance data. A provider shall not be required to notify an affected party whose contact information the provider has not previously collected or retained.

10) Requires a conversational AI system to clearly and prominently disclose to users that the conversational AI system generates synthetic content, as specified. In all conversational interfaces of such a system, the system shall, at the beginning of a user's interaction, obtain a user's affirmative consent acknowledging that the user has been informed that they are interacting with a conversational AI system before beginning the conversation, as specified.

11) Makes the above section operative on February 1, 2025.

12) Requires, beginning January 1, 2026, newly manufactured recording devices sold, offered for sale, or distributed in California to offer users the option to place a watermark into content produced by that device. A user shall have the option to customize the types of provenance data communicated by these watermarks, including by removing any personally identifiable information (PII). PII cannot be included by default.

13) Requires the recording devices to clearly inform users of the existence of the watermark settings upon a user's first use of the recording function and to contain a clear indicator that a watermark is being applied. A watermark shall, if enabled, be applied to nonsynthetic content produced using third-party applications that bypass default recording applications in order to offer recording functionalities and must be compatible with widely used industry standards.

14) Requires, beginning January 1, 2026, if technically feasible, a recording device manufacturer to offer a software or firmware update enabling a user of a recording device manufactured before January 1, 2026, and purchased in California to place a watermark on the content created by the device and to decode the provenance data.

15) Requires a large online platform, beginning March 1, 2025, to use labels to disclose the provenance data of content distributed on its platform, as specified. The labels shall prominently display whether content is fully synthetic, partially synthetic, nonsynthetic, nonsynthetic with minor modifications, or does not contain a watermark. A user shall be able to click or tap on a label to inspect provenance data in an easy-to-understand format.

16) Requires a large online platform to use state-of-the-art techniques to detect and label synthetic content that has had watermarks removed or that was produced by GenAI systems without watermarking functionality. If the platform is not able

to detect the provenance data of content, then the platform shall label the content as unknown provenance.

17) Provides that if content uploaded to or distributed on a platform does not contain provenance data, or if the content's provenance data cannot be interpreted or detected by the platform, a platform shall require the user to disclose whether the content is synthetic content or if they are uncertain. A large online platform shall include prominent warnings to users that uploading or distributing synthetic content without disclosing that it is synthetic content may be a violation of platform policy.

18) Requires a large online platform to use state-of-the-art techniques to detect and label text-based potentially deceptive content that is uploaded by users.

19) Requires a large online platform to make accessible some functionality for users to apply a digital signature to nonsynthetic content, and include options that do not require disclosure of PII.

20) Requires a large online platform that can detect potentially deceptive content that does not contain watermarks that comply with applicable industry standards to generate and store, in an online database to be shared and made privately accessible by all other online platforms and CDT, digital fingerprints and any associated provenance data for these images. This provenance data shall not include PII.

21) Authorizes CDT to share access to these databases with coordinating bodies acting to facilitate more rapid and computationally efficient detection and labeling of synthetic content.

22) Requires GenAI providers and large online platforms, beginning January 1, 2026, and annually thereafter, to produce a Risk Assessment and Mitigation Report that assesses the risks posed and harms caused by synthetic content generated by their GenAI systems or hosted on their GenAI hosting platforms. This must include assessments of the distribution of illegal GenAI-generated child sexual abuse materials, nonconsensual intimate imagery, disinformation related to elections or public health, plagiarism, or other instances where synthetic or potentially deceptive content caused or may have the potential to cause harm.

23) Requires the report to incorporate information known to the GenAI provider or large online platform about known harms caused by synthetic content generated by their systems or hosted on their platforms, as informed by reports submitted to, and confirmed by, the provider or platform, and independent investigation as appropriate, including, for example, illegal material. The report must 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.

24) Provides that a violation of these provisions may result in an administrative penalty, assessed by CDT, of up to $1 million or five percent of the violator's annual global revenue, whichever is greater.

25) Requires CDT, within 90 days of the effective date of this bill, to adopt regulations to implement and carry out the purposes of the bill. CDT is required to review and update its regulations as needed, including adopting specific national or international standards for provenance, authenticity, watermarking, and digital signatures, as long as the standards do not weaken these provisions.

26) Includes a severability clause.

COMMENTS

1. Blurring reality: AI-generated content

Generative AI is a type of artificial intelligence that can create new content, including text, images, code, or music, by learning from existing data. Generative AI models can produce realistic and novel artifacts that resemble the data they were trained on, but do not copy it. For example, generative AI can write a poem, draw a picture, or compose a song based on a given prompt or theme. Generative AI enables users to quickly generate new content based on a variety of inputs. Generative AI models use neural networks to identify the patterns and structures within existing data to generate new and original content.

The world has been in awe of the powers of this generative AI since the widespread introduction of AI systems such as the various iterations of ChatGPT. However, the capabilities of these advanced systems leads to a blurring between reality and fiction. The Brookings Institution lays out the issue:

> Over the last year, generative AI tools have made the jump from research prototype to commercial product. Generative AI models like OpenAI's ChatGPT and Google's Gemini can now generate realistic text and images that are often indistinguishable from human-authored content, with generative AI for audio and video not far behind. Given these advances, it's no longer surprising to see AI-generated images of public figures go viral or AI-generated reviews and comments on digital platforms. As such, generative AI models are raising concerns about the credibility of digital content and the ease of producing harmful content going forward.

> Against the backdrop of such technological advances, civil society and policymakers have taken increasing interest in ways to distinguish AI-generated content from human-authored content.[1]

One expert at the Copenhagen Institute for Future Studies estimates that should large generative-AI models run amok, up to 99 percent of the internet's content could be AI-generated by 2025 to 2030.[2] The problematic applications are seemingly infinite, whether it be deepfakes to blackmail or shame victims, misinformation in elections, false impersonations to commit fraud, or other nefarious purposes. Infamously, in January of this year, Taylor Swift was the victim of sexually explicit, nonconsensual deepfake images using AI that were widely spread across social media platforms.[3] Perhaps more disturbingly, a trend has emerged in schools of students creating such images: "At schools across the country, people have used deepfake technology combined with real images of female students to create fraudulent images of nude bodies. The deepfake images can be produced using a cellphone."[4]

In February of this year, voters in New Hampshire received robocalls that are purported to have used an AI voice resembling President Joe Biden advising them against voting in the presidential primary and saving their vote for the November general election.[5] Recently, a former federal judge urged the federal judiciary's Advisory Committee on Evidence Rules to update evidentiary rules regarding the admissibility of evidence believed to be AI generated.[6] But, in addition to concerns about the potential for AI-generated evidence to be admitted is the reverse, false claims that real evidence is synthetic. As more of the population becomes aware of the potential to realistically fake images, video, and text, some will use the skepticism that

[1] Siddarth Srinivasan, *Detecting AI fingerprints: A guide to watermarking and beyond* (January 4, 2024) Brookings Institution, https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/#:~:text=Google%20also%20recently%20announced%20SynthID,model%20to%20detect%20the%20watermark. All internet citations are current as of June 12, 2024.

[2] Lonnie Lee Hood, *Experts Say That Soon, Almost The Entire Internet Could Be Generated by AI* (March 4, 2022) The Byte, https://futurism.com/the-byte/ai-internet-generation.

[3] Brian Contreras, *Tougher AI Policies Could Protect Taylor Swift – And Everyone Else – From Deepfakes* (February 8, 2024) Scientific American, https://www.scientificamerican.com/article/tougher-ai-policies-could-protect-taylor-swift-and-everyone-else-from-deepfakes/.

[4] Hannah Fry, Laguna Beach High School investigates 'inappropriate' AI-generated images of students (April 2, 2024) Los Angeles Times, https://www.latimes.com/california/story/2024-04-02/laguna-beach-high-school-investigating-creation-of-ai-generated-images-of-students.

[5] Em Steck & Andrew Kaczynski, *Fake Joe Biden robocall urges New Hampshire voters not to vote in Tuesday's Democratic primary* (January 22, 2024) CNN, https://www.cnn.com/2024/01/22/politics/fake-joe-biden-robocall/index.html.

[6] Avalon Zoppo, *Threat of AI-Generated 'Deepfake' Evidence Needs Judiciary's Attention, Former Judge Says* (October 27, 2023) The National Law Journal, https://www.law.com/nationallawjournal/2023/10/27/threat-of-ai-generated-deepfake-evidence-needs-judiciarys-attention-former-judge-says/?slreturn=20240303000917.

creates to challenge the authenticity of real content, a phenomena coined the "liar's dividend."[7]

2. Taking action to identify synthetic content and address its usage

Earlier this year, the European Parliament signed the European Union AI Act. It highlights these very issues and obligates developers and deployers to assist in ensuring, to the extent feasible, that individuals are able to distinguish between original and AI-generated or manipulated content. The Act states:

> A variety of AI systems can generate large quantities of synthetic content that becomes increasingly hard for humans to distinguish from human-generated and authentic content. The wide availability and increasing capabilities of those systems have a significant impact on the integrity and trust in the information ecosystem, raising new risks of misinformation and manipulation at scale, fraud, impersonation and consumer deception. In light of those impacts, the fast technological pace and the need for new methods and techniques to trace origin of information, it is appropriate to require providers of those systems to embed technical solutions that enable marking in a machine readable format and detection that the output has been generated or manipulated by an AI system and not a human. Such techniques and methods should be sufficiently reliable, interoperable, effective and robust as far as this is technically feasible, taking into account available techniques or a combination of such techniques, such as watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods, fingerprints or other techniques, as may be appropriate. When implementing this obligation, providers should also take into account the specificities and the limitations of the different types of content and the relevant technological and market developments in the field, as reflected in the generally acknowledged state-of-the-art. Such techniques and methods can be implemented at the level of the system or at the level of the model, including general purpose AI models generating content, thereby facilitating fulfilment of this obligation by the downstream provider of the AI system. To remain proportionate, it is appropriate to envisage that this marking obligation should not cover AI systems performing primarily an assistive function for standard editing or AI systems not substantially altering the input data provided by the deployer or the semantics thereof.

[7] Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security* (July 14, 2018) 107 California Law Review 1753 (2019), https://ssrn.com/abstract=3213954.

It also specifically obligates deployers who use an AI system to generate or manipulate image, audio, or video content that "appreciably resembles existing persons, places or events and would falsely appear to a person to be authentic (deep fakes), should also clearly and distinguishably disclose that the content has been artificially created or manipulated by labelling the artificial intelligence output accordingly and disclosing its artificial origin."

There is currently an arms race in techniques for distinguishing between synthetic and authentic content and companies are declaring their commitment to identifying such content. There are various methods for deciphering AI-generated or altered content, although none are foolproof and all require updates as technology advances:

> There are several approaches that have been proposed for detecting AI-generated content. The four most prominent approaches are watermarking (in its various forms), which is the embedding of an identifiable pattern in a piece of content to track its origin; content provenance, which securely embeds and maintains information about the origin of the content in its metadata; retrieval-based detectors, where all AI-generated content is stored in a database that can be queried to check the origin of content; and post-hoc detectors, which rely on machine learning models to identify subtle but systematic patterns in AI-generated content that distinguish it from human-authored content.[8]

3. <u>Ensuring sufficient tools to establish the provenance of synthetic and nonsynthetic content</u>

This bill looks to implement some of these approaches to ensuring individuals can identify when content is modified by GenAI. It places a wide set of obligations on GenAI providers, recording device manufacturers, and large online platforms.

According to the author:

> The primary purpose in introducing the bill is to establish a comprehensive regulatory framework to mitigate the harmful impacts of synthetic or "deep fake" content. Specifically, the bill aims to address the interrelated problems stemming from the increasing proliferation and sophistication of generative AI technologies that can create synthetic or "deepfake" content that is difficult to distinguish from human-generated, non-synthetic content:
>
> 1. **Harms caused by potentially deceptive content presented as human generated content**: The bill acknowledges the wide range

[8] *See* footnote 1.

of potential harms caused by potentially deceptive AI generated content, including financial scams, non-consensual intimate imagery, disinformation (especially around elections and public health), and the erosion of trust in the digital information ecosystem. The PAWS Act seeks to reduce these harms by requiring clear disclosure of content provenance, making it harder for potentially deceptive content to be mistaken as human generated.

2. **Lack of transparency around provenance of digital media:** The bill addresses transparency concerns by mandating that generative AI providers embed imperceptible and indelible watermarks containing provenance data in all synthetic content they create, and prominently display this provenance data to users. The bill also establishes standards for recording devices to offer watermarking options for human generated content.
3. **Facilitation of harmful acts by bad actors:** The bill prohibits the distribution of tools designed to remove watermarks or manipulate provenance data, making it more difficult for bad actors to generate unlabeled, potentially deceptive content. It also requires conversational AI systems to disclose their artificial nature and obtain user consent, making it more difficult for bad actors to leverage AI generation for deception.

In sum, the bill seeks to establish clear standards and requirements around content provenance disclosure, watermarking, and labeling, with the goal of increasing transparency and reducing the ability of bad actors to deceive users with unlabeled synthetic content. By doing so, [I aim] to mitigate the various potential harms enabled by increasingly sophisticated generative AI technologies.

a. GenAI providers: tracking synthetic content

The bill places a strong obligation on GenAI providers to create the trail for tracking what is GenAI produced or modified content. Providers are required to place an imperceptible and "maximally indelible" watermark into synthetic content produced or significantly modified by their systems. "Watermark," for these purposes, is information that is embedded into a GenAI system's output for the purpose of conveying its synthetic nature, identity, content, including image, audio, video, text, or computer code, for the purpose of communicating the provenance, history of modifications, modification, or history of conveyance. "Maximally indelible watermark" is defined as a watermark that is designed to be as difficult to remove as possible using state-of-the-art techniques and relevant industry standards. "Synthetic content" simply means information, including images, videos, audio, and text, that has been produced or significantly modified by a generative AI system.

The bill provides for alternative methods of embedding provenance data where the content is too small. Ultimately, the bill requires providers to design watermarks, to the greatest extent possible, to communicate information that identifies the content as synthetic and identifies the provider. To make this information useful for consumers, the bill further requires GenAI providers to make a watermark decoder freely available to the public that is easy to use and adheres to relevant industry standards. The providers are also required to conduct red-teaming exercises, essentially structured testing of their watermarking technology, and to make the results public. GenAI providers must provide a full report to CDT and notify them of any identified vulnerabilities or failures in the system related to the erroneous or malicious inclusion or removal of provenance information or watermarks, along with other reporting requirements.

If their GenAI system is capable of producing "potentially deceptive content," providers must ensure the systems can generate and store, in a searchable online database in a manner that can be retrieved by a viewer of the content, a digital fingerprint of and provenance data[9] for any piece of such content that they produce. "Potentially deceptive content" means synthetic content that is so similar to nonsynthetic content that it could reasonably be mistaken as nonsynthetic content.

To ensure wider protections for this approach, the bill prohibits making available systems or tools that are designed to remove watermarks and GenAI hosting platforms from making available systems that are not compliant with the watermarking requirements laid out above.

The bill also requires conversational AI systems to clearly and prominently disclose that they produce synthetic content, as provided. At the outset of a user's interaction, such systems must obtain the user's affirmative consent acknowledging the user has been so informed.

The above provisions become operative on February 1, 2025.

b. Recording device manufacturers and provenance data

On the flip side of ensuring that synthetic content is properly marked to decipher it, is establishing the ability to identify authentic content. This bill requires recording devices sold, offered for sale, or distributed in California to offer users the option to place a watermark, compatible with industry standards, into content produced by that device. Users would have the option to customize the type of provenance data that is so included, but personally identifiable information must be excluded by default. The devices are required to inform users of the watermark settings and shall indicate when watermarks are being applied.

[9] For privacy purposes, this provenance data shall not include personally identifiable information.

These requirements take effect on January 1, 2026 and apply to newly manufactured devices. However, the bill requires manufacturers, if technically feasible, to offer a software or firmware update enabling a user of a recording device made before that date and purchased in California, to place a watermark on the content it creates and to decode the provenance data.

c. Large online platforms

The bill also includes a number of responsibilities for online platforms, where Californians are most likely to interact with synthetic content. It defines a "large online platform" as a public-facing internet website, web application, or digital application, including a social network,[10] video-sharing platform, messaging platform, advertising network, or search engine that had at least 1,000,000 California users during the preceding 12 months and can facilitate the sharing of synthetic content.

The platforms are required to label content distributed thereupon disclosing its provenance data. The bill places a number of requirements on the labels, including that they must clearly identify whether content is synthetic, nonsynthetic or somewhere in between.

Platforms are required to use state-of-the-art techniques to detect and label synthetic content that has had watermarks removed or that was produced by GenAI systems without watermarking functionality and potentially deceptive content uploaded by users.

If content does not include provenance data, users must be required to disclose whether it is synthetic and it must be labeled as of unknown provenance. Users must be provided some functionality for users to apply a digital signature to nonsynthetic content.

A large online platform that can detect potentially deceptive content that does not contain watermarks that comply with applicable industry standards shall generate and store, in an online database to be shared and privately accessible by all other online platforms and CDT, digital fingerprints and any associated provenance data for these images, excluding personally identifiable information. CDT is authorized to share access to these databases with coordinating bodies acting to facilitate more rapid and computationally efficient detection and labeling of synthetic content.

These requirements take effect March 1, 2025.

[10] The author has agreed to an amendment that makes reference to "social media platforms," as that term is defined in existing law, rather than "social network."

d. Impact assessments

The bill further requires, beginning January 1, 2026, and annually thereafter, GenAI providers and large online platforms to produce a Risk Assessment and Mitigation Report that assesses the risks posed and harms caused by synthetic content generated by their GenAI systems or hosted on their GenAI hosting platforms. These reports are required to include a host of specified information, including assessments of the illegal distribution of GenAI child sexual abuse material and nonconsensual intimate imagery (or "deepfake" porn). It shall also include information about known harms caused by synthetic content generated by their systems or hosted on their platforms. Audits of these reports by independent entities are also required to validate the claims made.

e. Enforcement

CDT is authorized to assess massive administrative penalties of up to $1 million or five percent of the violator's annual global revenue, whichever is greater, for every violation. CDT is given 90 days from the effective date of the bill to promulgate regulations to implement and carry out the law.

4. Concerns with the bill

A number of stakeholders have raised issues with the sheer breadth of the bill and the inability to sift through its many provisions in time for this Committee's hearing, especially given the significant amendments recently taken. Specific to the substance, concerns have also been raised in response to this and other measures regarding the feasibility of watermarking and other digital provenance requirements given the nascent stage of their development. Many existing tools suffer from inaccuracies and false positives. Various technical methods for synthetic content transparency, such as watermarking, fingerprinting, or metadata, are being developed and are evolving but are far from foolproof, as individuals can, for instance, sometimes find ways to remove or obscure these disclosures from content.

A coalition in opposition, including TechNet and NetChoice, argue the bill's "prescriptive requirements on content provenance and watermarking are technologically premature." They state:

> Many of our companies and platforms are at the forefront of developing content provenance and watermarking technology, which is still in its early stages. However, AB 3211 enacts incredibly prescriptive requirements for a technology that is still under development and rapidly evolving. For example, there isn't a program that can watermark text, making the bill's requirements to do so impossible to comply with. We believe references to text watermarking should be removed to reflect this reality.

> Furthermore, content provenance and watermarking is still incredibly unreliable and in many cases easy to break. Researchers at the University of Maryland were able to break all the currently available watermarking methods. Some can be avoided by simple cropping, resizing, or screenshotting an image. More concerning, these researchers were able to insert fake watermarks and credentials into images, creating false positives. Provenance and watermarking tools tend to help good faith actors act virtuously, but they have limits on stopping bad actors. No provenance solution that's been created so far, including watermarking or metadata, stops bad actors from simply 'stripping' provenance elements and posting a fake piece of content as authentic.
>
> In its standards for large online platforms, AB 3211 should more clearly delineate between 1st party and 3rd party content. 1st party content would be images, videos, or audio that is generated using a large online platform's generative AI tools and is then posted or distributed on that platform. In this instance, a platform can actually control the creation of a content provenance or watermark into the content. As mentioned, many of our companies are already working to incorporate this type of technology to increase transparency around AI-generated content. It is currently technically infeasible to accurately and reliably detect content that is created using a different platform's AI tools. As noted above, considering the current ease with which current watermarks can be broken, a legal requirement and mandate for 3rd party content isn't appropriate.

Despite the identified limitations for such markings, the author argues the bill is the first step and is not intended to require perfection:

> While no system is perfect, robust watermarking combined with behavioral analysis, red-teaming, and required disclosures will make deception far more difficult and increase the risk calculus of bad actors. The bill mandates ongoing improvements to standards as technology evolves to stay ahead of actors looking to circumvent these measures to cause harm.

Relevant here, Meta has recently committed to "label images that users post to Facebook, Instagram and Threads when we can detect industry standard indicators that they are AI-generated."[11] A group of tech companies, including Adobe, Google, and Microsoft, has established the Coalition for Content Provenance and Authenticity

[11] Nick Clegg, *Labeling AI-Generated Images on Facebook, Instagram and Threads* (February 6, 2024) Meta, https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/.

(C2PA) to address "the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content."[12] OpenAI announced that it will add C2PA metadata to images created with ChatGPT and the API for the DALL-E 3 model.

Just last month, Google announced new advances in its watermarking technology, including the capability to watermark GenAI created text:

> Among Google's swath of new AI models and tools announced today, the company is also expanding its AI content watermarking and detection technology to work across two new mediums.
>
> Google's DeepMind CEO, Demis Hassabis, took the stage for the first time at the Google I/O developer conference on Tuesday to talk not only about the team's new AI tools, like the Veo video generator, but also about the new upgraded SynthID watermark imprinting system. It can now mark video that was digitally generated as well as AI-generated text.
>
> Watermarking AI-generated content will matter increasingly as the technology gains prevalence, especially when AI gets used for malicious purposes. It's already been used to spread political misinformation, claim someone said something they haven't, and create nonconsensual sexual content.
>
> SynthID was announced last August and started as a tool to imprint AI imagery in a way that humans can't visually decipher — but can be detected by the system. The approach is different from other aspiring watermarking protocol standards like C2PA, which adds cryptographic metadata to AI-generated content.
>
> Google had also enabled SynthID to inject inaudible watermarks into AI-generated music that was made using DeepMind's Lyria model. SynthID is just one of several AI safeguards in development to combat misuse by the tech, safeguards that the Biden administration is directing federal agencies to build guidelines around.[13]

[12] *Overview*, Coalition for Content Provenance and Authenticity, https://c2pa.org/.

[13] Umar Shakir, *Google's invisible AI watermark will help identify generative text and video* (May 14, 2024) The Verge, https://www.theverge.com/2024/5/14/24155927/google-ai-synthid-watermark-text-video-io; *see also* Karissa Bell, *Google expands digital watermarks to AI-made video and text* (May 14, 2024) Yahoo!news, https://au.news.yahoo.com/google-expands-digital-watermarks-to-ai-made-video-175232320.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAI9FrY1zdqIO5ttk0ewE40_6KKMExr2AatuMW87CU3d1m989vvDptiMDa208Ejy3PQ0NA1-2KuSNH1Po8zormuGzG2sldsFq4QBcl3vqj5EAw1TOmsAr25yuQwrak0rFONQQMNRT51cBTTamk8ox5fgtZPvQxfk9hnNEC2QnQETC.

Many companies have already voluntarily committed to follow specified guidelines. As described in the White House fact sheet:

> President Biden [convened] seven leading AI companies at the White House [] – Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI – to announce that the Biden-Harris Administration has secured voluntary commitments from these companies to help move toward safe, secure, and transparent development of AI technology.
>
> Companies that are developing these emerging technologies have a responsibility to ensure their products are safe. To make the most of AI's potential, the Biden-Harris Administration is encouraging this industry to uphold the highest standards to ensure that innovation doesn't come at the expense of Americans' rights and safety.
>
> These commitments, which the companies have chosen to undertake immediately, underscore three principles that must be fundamental to the future of AI – safety, security, and trust – and mark a critical step toward developing responsible AI. As the pace of innovation continues to accelerate, the Biden-Harris Administration will continue to remind these companies of their responsibilities and take decisive action to keep Americans safe.[14]

The most relevant commitment is focused on earning the public's trust by ensuring individuals are aware of when content is AI generated:

> **Develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated, including robust provenance, watermarking, or both, for AI-generated audio or visual content**
> Companies making this commitment recognize that it is important for people to be able to understand when audio or visual content is AI-generated. To further this goal, they agree to develop robust mechanisms, including provenance and/or watermarking systems for audio or visual content created by any of their publicly available systems within scope introduced after the watermarking system is developed. They will also develop tools or APIs to determine if a particular piece of content was created with their system. Audiovisual content that is readily distinguishable from reality or that is designed to be readily recognizable as generated by a company's AI system—such as the default voices of AI assistants—is outside the scope of this commitment. The watermark or

[14] *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI* (July 21, 2023) The White House, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

> provenance data should include an identifier of the service or model that created the content, but it need not include any identifying user information. More generally, companies making this commitment pledge to work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI.[15]

Similar legislation has been introduced at the federal level by Senators Schatz and Kennedy, the AI Labeling Act of 2023. Also at the federal level, President Biden issued an executive order late last year on AI, to set standards for safety and security and to ensure responsible AI development and deployment in the United States. The order defined generative AI as "the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content."

The EO calls for the development of a risk management framework for generative AI, modeled after the AI Risk Management Framework created by the National Institute for Standards and Technology (NIST).

The many recent developments in this space are encouraging. However, given the expansive obligations imposed by the bill and the questions regarding feasibility for all of the GenAI providers, platforms, and manufacturers that must implement this technology, the author has agreed to amendments that delay the effective date of the various sections of the bill to July 1, 2026. This allows time for the technology to advance and also gives the Legislature a chance to make adjustments next year should they see fit.

5. Additional stakeholder positions

A coalition of groups in support, including SEIU California, writes:

> Generative AI technologies are increasingly able to generate inauthentic images, audio, video, and text content—sometimes called "deepfakes"—that appear authentic even to experts. Fake content that is presented as real can include many things, including:
>
> - Disinformation impacting free and fair elections, which we have already seen with deepfakes undermining or influencing national elections in Bangladesh, Slovakia, and Argentina, and local elections around the United States;
> - Fake digital media that undermine national security, such as fake photos of an attack on the Pentagon, which were disseminated by a

[15] *Ibid*.

> twitter account of RT, a Russian-backed media company, and which caused a US stock market plunge;
>
> - Non-consensual intimate imagery, demonstrated by the obscene sexual imagery of Taylor Swift that was published and circulated widely around the time of the Super Bowl; and
> - Many other forms of problematic content, including scams and fraud, child sexual abuse material, and plagiarism.
>
> . . .
>
> AB 3211, the California Provenance, Authenticity and Watermarking Standards (PAWS) Act, attempts to solve these problems. The bill will require AI content generators to provide the public with tools to distinguish between authentic and fake content across digital media. AI content generators will be required to embed maximally-indelible and privacy-protecting content provenance data into the content that they generate— whether AI-generated or authentic—and all large online platforms will be required to display clearly understandable labels on content that alert users to its provenance. While these requirements alone will not fully eliminate the harms caused by fake online content, they can dramatically reduce those harms by making clear which content is fake and which is real.

Oakland Privacy writes in opposition with concerns about the expansive scope of the bill, urging the author to consider focusing on particularly troublesome content rather than the broad blanket requirements imposed by the bill. They also advocate for more technology neutral language.

The Coalition in opposition also raises concerns about the enforcement mechanism:

> Considering the prescriptive nature and technical infeasibility of some requirements and the technical impossibility of others, we believe the penalties for noncompliance of $1 million or 5% of global annual revenue are unjustifiable. As mentioned above, platforms cannot watermark text content. The technology to apply watermarks to audio, images, and videos are much further along but vary in their resistance to removal or inadvertent breaking.

The penalties imposed by the bill for *any* violation are extremely excessive. To more appropriately hone the penalty provision, the author has agreed to amendments that limit penalties to up to $50,000 per violation but provide for heightened penalties of up to $500,000 for violations that are intentional or that result from grossly negligent conduct.

SUPPORT

California Initiative for Technology & Democracy (sponsor)
Accountable Tech
Bay Rising
California Voter Foundation
Catalyst California
Center for Countering Digital Hate
Check My Ads
Chinese Progressive Association
City and County of San Francisco, Board of Supervisors
Hmong Innovating Politics
Move (Mobilize, Organize, Vote, Empower) the Valley
NextGen California
Partnership for the Advancement of New Americans
SEIU California
Techequity Action

OPPOSITION

California Chamber of Commerce
Computer and Communications Industry Association
Netchoice
Oakland Privacy
Technet

RELATED LEGISLATION

Pending Legislation:

SB 942 (Becker, 2024) places obligations on businesses that provide generative artificial intelligence (AI) systems to develop and make accessible tools to detect whether specified content was generated by those systems. These "covered providers" are required to include visible and imperceptible markings on AI-generated content to identify it as such. SB 942 is currently in the Assembly Privacy and Consumer Protection Committee.

SB 970 (Ashby, 2024) would have ensured that media manipulated or generated by AI technology is incorporated into the right of publicity law and criminal false impersonation statutes. SB 970 requires those providing access to such technology to provide a warning to consumers about liability for misuse. SB 970 was held in the Senate Appropriations Committee.

AB 1791 (Weber, 2024) requires social media platforms to delete "personal provenance data" related to a user's identity from content uploaded to the platform, while retaining provenance data related to the system or service used to generate the content. AB 1791 is currently in this Committee.

AB 2930 (Bauer-Kahan, 2024) requires, among other things, a deployer and a developer of an automated decision tool to, on or before January 1, 2026, and annually thereafter, perform an impact assessment for any automated decision tool the deployer uses that includes, among other things, a statement of the purpose of the automated decision tool and its intended benefits, uses, and deployment contexts. The assessments must be provided to the Civil Rights Department within 7 days of a request. AB 2930 requires a deployer to, at or before the time an automated decision tool is used to make a consequential decision, notify any natural person that is the subject of the consequential decision that an automated decision tool is being used to make, or be a controlling factor in making, the consequential decision and to provide that person with, among other things, a statement of the purpose of the automated decision tool. AB 2930 is currently in this Committee.

Prior Legislation:

SB 444 (Umberg, 2019) would have requested the Regents of the University of California (UC) to enact a resolution authorizing the law schools at UC Berkeley and UC Irvine to participate in a pilot project to develop AI or machine-learning solutions to address access to justice issues faced by self-representing litigants in their respective courts. The bill died in the Assembly Higher Education Committee.

AB 1576 (Calderon, 2019) would have required the Secretary of GovOps to appoint participants to an AI working group to evaluate the uses, risks, benefits, and legal implications associated with the development and deployment of AI by California-based businesses. The bill was held on the Senate Appropriations Committee suspense file.

SJR 6 (Chang, Res. Ch. 112, Stats. 2019) urged the President and the Congress of the United States to develop a comprehensive AI Advisory Committee and to adopt a comprehensive AI policy.

PRIOR VOTES:

Assembly Floor (Ayes 62, Noes 0)
Assembly Appropriations Committee (Ayes 14, Noes 0)
Assembly Judiciary Committee (Ayes 10, Noes 0)
Assembly Privacy and Consumer Protection Committee (Ayes 10, Noes 0)

****