Date of Hearing: April 16, 2024

ASSEMBLY COMMITTEE ON PRIVACY AND CONSUMER PROTECTION
Rebecca Bauer-Kahan, Chair
AB 3211 (Wicks) – As Amended March 21, 2024

AS PROPOSED TO BE AMENDED

SUBJECT: California Provenance, Authenticity and Watermarking Standards

SYNOPSIS

As generative AI (GenAI) technologies become more accessible, online content that appears real, but that is actually false, threatens to flood social media and other large online platforms. The unmitigated spread of synthetic content threatens to harm individual Californians in numerous ways, such as through the proliferation of nonconsensual deepfake pornography, scams, and the distribution of targeted election disinformation.

The issue of GenAI-produced content can be tacked through a three-step plan. First, all artificial content must be labeled as such. Second, all "real" content – such as video and audio recordings – much be labeled as such. Third, large online platforms must be required to prominently display these labels, allowing users to distinguish between artificial and real content.

This bill seeks to enact the plan described above. It would require 1) the providers of GenAI systems to embed "watermarks" detailing content provenance into all GenAI ouputs; 2) digital cameras and other recording devices to offer users the option to embed watermarks into outputs; and 3) social media companies to prominently display labels that contain the provenance information of real and GenAI-produced content.

This bill is sponsored by CITED and supported by Asian Americans and Pacific Islanders for Civic Empowerment, the Asian Law Alliance, the California Voter Foundation, and Courage California. It is opposed by the California Chamber of Commerce, Computer & Communications Industry Association, Netchoice, and Technet.

Proposed committee amendments revise and recast the bill to make certain substantive changes, detailed in the analysis, as well as other technical and clarifying changes. If the bill passes this Committee, it will next be heard by the Judiciary Committee.

SUMMARY: Establishes watermarking and content provenance requirements for GenAI system providers, manufacturers of digital cameras and other recording devices, and large online platforms. Specifically, **this bill**:

1) Requires GenAI system providers to place imperceptible and maximally indelible watermarks containing provenance data into the content created by the provider's system.

   a) Defines "generative AI system" to mean 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.

b) Defines "generative AI system providers" to mean 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. GenAI 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.

c) Defines "watermark" to mean information that is embedded into a generative AI output for the purposes of verifying the authenticity of the output or the identity or characteristics of its provenance, modifications, or conveyance.

d) Defines "provenance data" to mean 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.

2) Requires providers to provide downloadable watermark decoders that allow individuals or large online platforms to quickly assess whether a given piece of content was created with the provider's system, and to make those tools available to all large online platforms and the public. Defines "watermark decoders" to mean freely available software tools or online services that can read or interpret watermarks and output the provenance data embedded in them.

3) Requires providers to conduct red-teaming exercises involving third-party experts to test whether watermarks can be easily removed from the content produced by their systems, or falsely added to authentic content.

a) Requires summaries of these exercises to be posted online and sent to the Department of Technology (CDT) within six months of the bill taking effect, and annually thereafter.

b) Defines "AI red-teaming" to mean a structured testing effort to find flaws and vulnerabilities in an AI system, 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.

c) Defines "authentic content" to mean 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.

4) Prohibits providers from continuing to support systems that were in use prior to this bill taking effect unless either of the following conditions is met:

a) The provider releases a watermark decoder tool that can determine whether a given piece of content was created by the provider's GenAI system.

b) The provider publishes research demonstrating that the content produced by their system could not reasonably be mistaken as authentic.

5) Prohibits any provider or distributor of software or online services from making available or distributing a tool, application, or service capable of removing watermarks from synthetic content. Defines "synthetic content" to mean information, including images, videos, audio clips, and text, that has been significantly modified or generated by algorithms, including by AI.

6) Prohibits the distribution of open-source GenAI systems that are capable of having their watermarking functions removed.

7) Requires providers to report any vulnerabilities or failures they discover in a system to CDT, any other system providers who may have been affected by similar vulnerabilities or failures, and any affected party.

8) Requires providers to post any report made to the CDT online. If the publication of the report could pose public safety risk, allows the provider to instead post a summary disclosure or delay publication of the full report for no longer than 30 days.

9) Requires conversational AI systems to disclose to users that they receive synthetic content. For visual interfaces, this disclosure is required to be displayed prominently throughout a user interaction. For audio interfaces, the disclosure is required to be made verbally. Requires conversational AI systems to obtain affirmative consent from users before beginning a conversation. Defines "conversational AI systems" to mean 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.

10) Beginning January 1, 2026, requires newly manufactured digital cameras and recording devices distributed in California to offer users the option to place authenticity and provenance watermarks in content produced by those devices. Requires that those watermarks abide by relevant industry standards.

    a) Defines "authenticity watermark" to mean a watermark of authentic content that includes the name of the device manufacturer.

    b) Defines "provenance watermark" to mean 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.

11) Requires authenticity and provenance watermarks to persist in content produced using third-party applications that grant access to a digital camera's functionality.

12) Requires manufacturers of digital cameras to push software and firmware updates to existing digital cameras in order to grant them the ability to place watermarks on content, if technically feasible.

13) Beginning March 1, 2025, requires a large online platform to label content posted to the platform with the provenance data embedded in the content. The labels must indicate whether the content is fully synthetic, partially synthetic, authentic, authentic with minor modifications, or does not contain a watermark. Defines "large online platform" to mean 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.

14) Requires large online platforms to use state-of-the-art techniques to detect and label synthetic content that has had watermarks removed and that was produced by AI systems without watermarking functionality.

15) Requires large online platforms to require their users to disclose uploads of synthetic content. Requires these platforms to warn users that failing to disclose the nature of such content will lead to disciplinary action.

16) Permits a platform to give users the option to indicate they are unsure of whether uploaded content is synthetic. If so indicated, requires platforms to label the content "possibly synthetic".

17) Requires platforms to disable the accounts of users that have uploaded or distributed inauthentic content without disclosing it.

18) Requires platforms to create a verification process for users to apply a digital signature to content created by human beings. Requires the process to include options to verify without requiring disclosure of personally identifiable information.

19) Requires GenAI providers, GenAI distributors, and large online platforms to produce Risk Assessment and Mitigation Reports that assess the risks posed by synthetic content and the harms that have been or could be caused by such content.

    a) Requires, at a minimum, that these reports include assessments on the distribution of AI-generated child sexual abuse materials, nonconsensual pornography, election and public health disinformation, and plagiarism.

    b) Requires that these reports be audited by independent auditors using state-of-the-art techniques.

    c) Defines GenAI system distributors to mean organizations or individuals that distribute GenAI 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.

20) Permits CDT to authorize independent researchers to access special researcher tools designed to facilitate large-scale and efficient analysis of AI-generated content.

21) Sets the penalty for a violation of the bill's provisions at the greater of $1,000,000 or 5% of the violator's annual global revenue, to be assessed as an administrative penalty by CDT.

22) Requires CDT to adopt regulations as necessary within 90 days of the bill coming into effect. Requires CDT to adopt specific national or international standards for provenance, authenticity, watermarking, or digital signatures.

23) Provides that the bill's provisions are severable.

EXISTING LAW:

1) Establishes CDT and tasks it with the approval and oversight of information technology projects undertaken by the state. (Gov. Code § 11545 *et seq.*)

2) Defines "deepfake" to mean audio or visual content that has been generated or manipulated by artificial intelligence (AI) 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. Requires the Secretary of Government Operations to evaluate the impact of the proliferation of deepfakes on the state. (Gov. Code § 11547.5.)

3) Requires social media companies to post terms of service for each social media platform owned or operated by the company in a manner reasonably designed to inform all users of the social media platform of the existence and contents of the terms of service. (Bus. & Prof. § Code 22676.)

FISCAL EFFECT: As currently in print, this bill is keyed fiscal.

COMMENTS:

1) **AI and GenAI.** The development of GenAI is creating exciting opportunities to grow California's economy and improve the lives of its residents. GenAI can generate compelling text, images and audio in an instant – but with novel technologies come novel safety concerns.

What is AI? In brief, AI is the mimicking of human intelligence by artificial systems such as computers. AI uses algorithms – sets of rules – to transform inputs into outputs. Inputs and outputs can be anything a computer can process: numbers, text, audio, video, or movement. AI is not fundamentally different from other computer functions; its novelty lies in its application. Unlike normal computer functions, AI is able to accomplish tasks that are normally performed by humans.

AI that are trained on small, specific datasets in order to make recommendations and predictions are sometimes referred to as "predictive AI." This differentiates them from GenAI, which are trained on massive datasets in order to produce detailed text and images. When Netflix suggests a TV show to a viewer, the recommendation is produced by predictive AI that has been trained on the viewing habits of Netflix users. When ChatGPT generates text in clear, concise paragraphs, it uses GenAI that has been trained on the written contents of the internet.

GenAI tools can be released in open-source or closed-source formats by their creators. Open-source tools are publically available; researchers and developers can access their code and parameters. This accessibility increases transparency, but it has downsides: when a tool's code and parameters can be easily accessed, they can be easily altered, and open-source tools have the potential to be used for nefarious purposes such as generating deepfake pornography and targeted propaganda. Some open source models currently on the market include:

- LLaMA2: "Large Language Model Meta AI 2". An open source LLM released by Meta in 2023.

- Bert: "Bidirectional Encoder Representations from Transformers". An open source LLM released by Google in 2018.

- Stable Diffusion: An open source image generator released by startup Stability AI in 2022.

By comparison, closed-source tools are opaque with respect to their security features. It is harder for bad actors to generate illicit materials using these tools. But unlike open-source tools, closed-source tools are not subject to collective oversight because their inner workings cannot be examined by independent experts. Some closed-sourced models currently on the market include:

- GPT-4: A large multimodal model that accepts both image and text inputs and emits text outputs. Released by OpenAI in 2023.
- AlphaFold: Released by Google DeepMind in 2018, AlphaFold predicts protein structures from amino acid sequences.
- Copilot: A chatbot released by Microsoft in 2023.
- IBM Watson: A question-answering system developed by IBM to compete on Jeopardy in 2011, and released commercially in 2013.
- Siri: A digital assistant released by Apple in 2011.

2) **Author's statement.** The author states that their primary purpose in introducing the bill is to establish a comprehensive regulatory framework to mitigate the harmful impacts of synthetic or "deep fake" content. According to the author, these include:

1. Harms caused by inauthentic content presented as authentic: The bill acknowledges the wide range of potential harms caused by inauthentic 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 inauthentic content to be mistaken as authentic.

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 digital cameras and recording devices to offer watermarking options for authentic 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 inauthentic 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, the author aims to mitigate the various potential harms enabled by increasingly sophisticated generative AI technologies.

3) **The problem.** Image manipulation and video doctoring have existed for nearly as long as photography and recording equipment, but they have historically required great effort and talent. In the past few years the rapid development of GenAI has drastically reduced those barriers to entry, allowing a vast quantity of convincing – but ultimately fake – content to be generated in an instant. The creation of text, imagery, video, and audio by GenAI has the potential to change the world by automating repetitive tasks and fostering creativity. When employed by bad actors, however, these capabilities have the potential to destroy lives and destabilize societies.

Nonconsensual pornography. GenAI has been used to create pornography since its inception. This content is inevitably nonconsensual, and as GenAI technology improves, these products will become harder to distinguish from reality. Women have always been the primary victims of these efforts; in the run-up to the 2024 Super Bowl, a series of images involving Taylor Swift began to appear on the social media platform X (formerly Twitter). These images were removed over the following days, but the damage had been done:

> "We are too little, too late at this point, but we can still try to mitigate the disaster that's emerging," says Mary Anne Franks, a professor at George Washington University Law School and president of the Cyber Civil Rights Initiative. Women are "canaries in the coal mine" when it comes to the abuse of artificial intelligence, she adds. "It's not just going to be the 14-year-old girl or Taylor Swift. It's going to be politicians. It's going to be world leaders. It's going to be elections."[1]

The harms of nonconsensual AI-powered pornography are already being felt in California:

> A third school in Southern California has been hit with allegations of digitally manipulated images of students circulating around campus . . . "Sixteen eighth-grade students were identified as being victimized, as well as five egregiously involved eighth-grade students," Superintendent Michael Bregy wrote. While Bregy acknowledged that children "are still learning and growing, and mistakes are part of the process," he affirmed disciplinary measures had been taken and noted that the incident was swiftly contained. The district vowed to hold accountable any other students "found to be creating, disseminating, or in possession of AI-generated images of this nature."[2]

Scams. GenAI-powered speech and video is driving a new era in scamming. These AI tools are trained on publicly available data – the more data a target has online, the easier it is to develop a passable imitation of them or their loved ones. This is especially true of wealthy clients, whose public appearances, including speeches, are often widely available on the internet.[3] For example, a complicated scam utilizing both deepfake video and false audio was recently performed in Hong Kong. A multinational company lost $25.6 million after employees were fooled by deepfake technology, with one incident involving a digitally recreated version of its chief

[1] Brian Contreras, "Tougher AI Policies Could Protect Taylor Swift—And Everyone Else—From Deepfakes," Feb. 8. 2024, www.scientificamerican.com/article/tougher-ai-policies-could-protect-taylor-swift-and-everyone-else-from-deepfakes/.

[2] Mackenzie Tatananni, "'Inappropriate images' circulate at yet another California high school, as officials grapple with how to protect teens from AI porn created by classmates," *Daily Mail*, Apr. 11, 2024, https://www.dailymail.co.uk/news/article-13295475/Inappropriate-images-California-Fairfax-High-School-AI-deepfake.html

[3] Emily Flitter and Stacy Cowley, "Voice Deepfakes Are Coming for Your Bank Balance", *New York Times*, Aug. 30, 2023, www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html.

financial officer ordering money transfers in a video conference call. Everyone present on the video call, except the victim, was a fake representation of real people. The scammers applied deepfake technology to turn publicly available video and other footage into convincing versions of the meeting's participants.[4]

Political propaganda and disinformation. Deepfake technology is being used around the world to spread disinformation and propaganda. 2024 is a major election year in democracies around the globe: at least 64 countries will hold elections, representing close to 49% of the world's population.[5] It is also likely to be the first of many election years in which AI plays a pivotal role, as the technology becomes more widely available and easier to use. This has already been observed in Slovakia, where deepfake audio influenced an election in 2023:

> Days before a pivotal election in Slovakia to determine who would lead the country, a damning audio recording spread online in which one of the top candidates seemingly boasted about how he'd rigged the election. And if that wasn't bad enough, his voice could be heard on another recording talking about raising the cost of beer. The recordings immediately went viral on social media, and the candidate, who is pro-NATO and aligned with Western interests, was defeated in September by an opponent who supported closer ties to Moscow and Russian President Vladimir Putin.[6]

Similar deepfakes have now been deployed in the United States in advance of the 2024 presidential election. In late January, between 5000 and 20,000 New Hampshire residents received AI-generated phone calls impersonating President Biden that told them not to vote in the state's primary. The call told voters: "It's important that you save your vote for the November election." Concern about this call has led at least 14 states to introduce legislation targeting AI-powered disinformation.[7] It is still unclear how many people might not have voted based on these calls.

Deepfakes are not only being deployed by third parties; they can be used by the candidates themselves, either to improve their own self-images or to detract from their opponents. In mid-2023, former Republican presidential candidate Governor Ron DeSantis used AI to add fighter jets to one of his campaign videos.[8] Around the same time, Governor DeSantis' super PAC released an ad containing an AI-generated speech by former president Donald Trump.[9] The Republican National Committee also released a 30-second ad that displayed images of disorder

[4] Harvey Kong, "'Everyone looked real': multinational firm's Hong Kong office loses HK$200 million after scammers stage deepfake video meeting," *South China Morning Post*, Feb. 4, 2024, www.scmp.com/news/hong-kong/law-and-crime/article/3250851/everyone-looked-real-multinational-firms-hong-kong-office-loses-hk200-million-after-scammers-stage.

[5] Koh Ewe, "The Ultimate Election Year: All the Elections Around the World in 2024," *Time*, Dec. 28, 2023, https://time.com/6550920/world-elections-2024/.

[6] Curt Devine, Donie O'Sullivan, Sean Lyngass, "A fake recording of a candidate saying he'd rigged the election went viral. Experts say it's only the beginning," *CNN*, Feb. 1, 2024, www.cnn.com/2024/02/01/politics/election-deepfake-threats-invs/index.html.

[7] Adam Edelman, "States turn their attention to regulating AI and deepfakes as 2024 kicks off," *NBC News*, Jan. 22, 2024, www.nbcnews.com/politics/states-turn-attention-regulating-ai-deepfakes-2024-rcna135122.

[8] Ana Faguy, "New DeSantis Ad Superimposes Fighter Jets In AI-Altered Video Of Speech," *Forbes*, May 25, 2023, www.forbes.com/sites/anafaguy/2023/05/25/new-desantis-ad-superimposes-fighter-jets-in-ai-altered-video-of-speech/.

[9] Alex Isenstadt, "DeSantis PAC uses AI-generated Trump voice in ad attacking ex-president," *Politico*, Jul. 17, 2023, www.politico.com/news/2023/07/17/desantis-pac-ai-generated-trump-in-ad-00106695.

and destruction, with a voiceover that described the “consequences” of re-electing President Biden.[10] None of the images in this ad were real.

4) **The solution.** In principle, the issue of content authenticity can be solved by carrying out a three-step process:

1. Require all content produced by GenAI to be labeled “fake.”

2. Require all content produced by cameras and other recording devices to be labeled “real.”

3. Require all online platforms to prominently display these labels.

This is the precise strategy adopted by the bill. But is such a solution feasible? The devil is, of course, in the details. Each step in this process has a variety of technical and legal hurdles to overcome – but there is no real alternative. It is not currently possible to develop a tool that perfectly identifies AI-generated content. And as GenAI technology continues to develop, this situation will only become more dire.

5) **Step one.** Step one of the three-step content authenticity plan outlined above requires all content produced by GenAI to be labeled “fake.” To accomplish this, AB 3211 would require what it calls “generative AI system providers” to build watermarking functionality into the GenAI systems released by those providers. Watermarking refers to the practice of embedding visible or invisible markers into GenAI products. This is a tricky technology – at present, there is no single set of industry standards that system providers can turn to as they grapple with this bill’s requirements. Conversations between committee staff and the author have revealed that this is partially by design: in requiring a good-faith effort to adopt watermarking technologies on the part of GenAI providers, the author hopes to incentivize the development of “best practices” for watermarking.

The bill defines “generative AI system providers” to encompass the organizations and individuals that make GenAI systems – or substantial components thereof, such as the parameters that make these systems tick – available, or that provide them as standalone models, or that embed them into other systems. So far, so good: each of these entities is capable of building watermarking functionality into their products. However the definition goes on to include individuals and organizations that put these systems into service (often termed “deployers”), as well as online repositories or hosting websites that make systems available “even if those repositories are not the original makers of the AI systems they make available.” It is not clear how these parties could abide by the other requirements of this bill, given that they do not have the ability to edit the GenAI systems they work with. By including them in this definition, they are effectively excluded from the market by virtue of being unable to comply. **A committee amendment has been proposed to narrow this definition such that it specifically targets parties who can carry out the provisions of this bill.**

Watermarking is a tricky technology. Watermarks can be stripped from content relatively easily by common screenshotting tools, file compression software, and image editing programs like Photoshop. They can also be faked by treating a GenAI system like a copy machine: a real image or video can be fed into a GenAI system and spit back out, unaltered except for the addition of a

[10] GOP, "Beat Biden," Apr. 25, 2023, https://www.youtube.com/watch?v=kLMMxgtxQ1Y.

watermark. The system's user receives a piece of authentic content that has been marked as inauthentic, ready to be posted online in order to create confusion and sow discord.

AB 3211 accounts for these inherent weaknesses in two ways: first, by requiring that providers and other entities named throughout this bill make their tools and systems "compatible with widely used industry standards." Second, this bill requires GenAI providers to conduct "AI red-teaming exercises" in collaboration with third-party experts. These red-teaming exercises are described in the White House's recent executive order on AI:

> The term "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 AI. Artificial Intelligence red-teaming is most often performed by dedicated "red teams" that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system.[11]

Upon completing a red-teaming exercise, a provider is required to both post the results of the exercise online and submit the results to CDT. Upon discovering a vulnerability or flaw in their system, a provider is required to submit a report to CDT, inform all GenAI system providers who might be affected by similar vulnerabilities or flaws, and inform all affected users. This third point presents a privacy concern: in order to inform affected users, this bill would require a provider to collect the contact information of its users. **A committee amendment has been proposed that would limit the requirement to inform users to those users whose contact information the provider has already collected, thus removing the incentive for providers to collect contact information from all users.**

This bill imposes a third and final requirement on GenAI system providers: providers must develop software that allows a user to verify whether any given piece of content was created by the provider's system. The bill refers to this software as a "watermark decoder." This is by far the most difficult requirement imposed on providers of GenAI systems. The bill specifically requires that this software be capable of being used "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." The costs associated with this requirement would likely be exorbitant, effectively limiting the release of GenAI systems to large, well-established entities. A committee amendment has been introduced to target the use of these tools at users seeking to test individual pieces of content, rather than high-throughput testing by large online platforms. Watermarking decoders have another issue: their existence makes it substantially easier to intentionally strip watermarks out of labeled content. A piece of labeled content can be slightly altered and passed through a watermarking decoder to have the effect of the alteration read out. By repeating this process, a bad actor can determine exactly how much a piece of inauthentic content must be altered before a watermarking decoder no longer recognizes it as inauthentic.

This bill would also prohibit the release of GenAI systems whose watermarking functionalities can be removed, effectively banning the release of open-source GenAI systems in California. This could severely disadvantage California's GenAI industry: two of the world's largest open-

[11] "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," *White House,* Oct. 30, 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

source GenAI systems are Meta’s Llama2, based in California, and Mistral AI’s Mistral 7B, based in France. The effect of this provision would likely be to remove Llama2 from contention. Banning the use of open-source GenAI would also severely impact the small business community, as training GenAI from scratch is extremely expensive and time intensive. Many small businesses build their tools and systems by using open-source GenAI systems as a starting point. **Given that the permissibility of open-source AI remains a contentious, open question, a committee amendment has been proposed to strike this provision from the bill.**

GenAI providers offering conversational AI systems are required to disclose their nature to users. In visual interfaces, the disclosure must be visible throughout the interaction. In audio interfaces, the disclosure must be made at the beginning and end of the interaction. In all cases, the conversational AI is required to obtain affirmative consent from the user before proceeding.

6) **Step two.** Step two of the three-step content authenticity plan requires all “real” content to be labeled as such. This bill would require newly manufactured digital cameras and other recording devices sold or distributed in California to offer users the option to place watermarks in recorded content. Manufacturers of existing cameras are required, if possible, to provide updates that grant watermarking functionality.

7) **Step three.** The third and final step of the three-step content authenticity plan requires large online platforms to use labels to prominently display provenance data found in watermarks that are, by now, embedded in both real and GenAI-produced content. “Large online platforms” are defined to include websites, web applications and digital applications that have at least one million California users per year. These platforms are additionally required to use state-of-the-art techniques to detect and label synthetic content that has had watermarks removed, or that was produced by AI systems without watermarking functionalities.

AB 3211 requires large online platforms to warn their users that failing to disclose GenAI-produced content is not permissible, and will result in disciplinary action. The bill goes on to require these platforms to disable the accounts of users who upload GenAI-produced content without disclosing it. **Because this raises substantial First Amendment concerns, a committee amendment has been proposed to remove it from the bill.**

8) **Additional considerations**. While the bill, as proposed to be amended, has been substantially clarified and refined, there are a number of considerations that the author may wish to address going forward:

Artificial text and audio. Watermarking, as a technology, is in its infancy. Progress has been made towards developing standards for watermarking of images and video, but how audio and text should be watermarked is far less clear. This bill does not discriminate between these types of content, despite drastic differences in the feasibility of incorporating watermarks into each.

Bananas and bite-sized content. “Invisible and maximally indelible” watermarks usually take the form of statistical correlations between different parts of a GenAI product: for example, the intensities of certain pixels in an image (if the GenAI product is visual) or the choice of words in paragraph (if the GenAI product is textual). The inclusion of watermarks inevitably degrades the quality of a GenAI product, and this effect becomes more profound as the GenAI product becomes smaller. For example: imagine a textual GenAI product that is watermarked by including the word “banana” 10 times throughout the document. If the document is tens of thousands of words long, the word “banana” does not stand out. But if the document is a short snippet of 100 words, “banana” appears constantly. The document would contain almost nothing else. The smaller a GenAI product is, the more it must be altered from its ideal form in order to include a watermark.

The Department of Technology? The bill tasks the CDT with administrative enforcement of violations. CDT is under the Government Operations Agency and is organized into several offices. CDT oversees nearly all aspects of information technology (IT) in state government, including advising the Governor on the strategic management and direction of the State’s IT resources; issuing and maintain policies, standards, and procedures governing the State's information security; and providing oversight of the State’s IT projects. Thus, it appears that CDT is generally tasked with managing the State’s internal operations rather than actively regulating private industry. The author may wish to explore whether CDT is the appropriate enforcement entity.

“One Million Dollars.” The bill authorizes CDT to assess an administrative enforcement of up to $1 million or five percent of the violator’s annual global revenue, whichever is greater. In 2023, for example, Meta generated $134 billion in revenue. A single violation of this bill could have resulted in a $6.7 billion fine. The author may wish to set a ceiling on the penalty that is more proportionate to the harm of a violation.

Compelled speech. With respect to conversational AI systems and provenance data in watermarks, the bill requires certain disclosures and thus compels speech. Because the right to speak encompasses the right not to speak, these provisions implicate the First Amendment. Should this bill pass this Committee, these issues will be more thoroughly examined in the Judiciary Committee.

8) **Committee amendments.** In addition to the substantive amendments described above, the author, sponsors, and committee staff worked together to revise and recast the bill and make a number of clarifying and technical change. As proposed to be amended, the bill’s language, in its entirety, would be replaced with the following:

> ***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 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 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 system" means an artificial intelligence system that generates derived synthetic content, including images, videos, audio, text, and other digital content.

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

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

(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:

(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) which portions of the content 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 subsection (i) being the most important, and subsection (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, which specific portions of the content 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 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.

(f) (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 system 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.

(g) (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.

(h) 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 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.

(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 in 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 upload or distribute 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.***
>
> ***(3) The report 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 agency 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.***

9) **Related legislation.** AB 1791 (Weber, 2024) prohibits a social media platform from removing digital content provenance verification from content uploaded to the social media platform by a user. The bill is pending in this Committee.

AB 3050 (Low, 2024) would require CDT to issue regulations to establish standards for watermarks to be included in covered AI-generated material. The bill is pending in this Committee.

ARGUMENTS IN SUPPORT

The sponsor, CITED, writes:

> California has the opportunity to lead the U.S. response to these threats. As the home of Silicon Valley and the heartland of AI innovation, California can and must play an important role in shaping the solution. To that end, AB 3211 represents a phased-in solution to address this complex problem that complements existing standard-setting efforts. The bill has been developed with insights from EU regulators who have worked intimately on the AI Act and from federal policy experts who see, in the face of congressional inaction, California's power to drive nationwide change.

California Voter Foundation writes:

> California lawmakers have a long history and strong track record of ensuring voters can make informed choices through enhanced campaign finance disclosure laws, truth in committee naming laws, and requiring counties to list supporters for and against ballot measures directly on California ballots. AB 3211 continues California lawmakers'

> longstanding efforts to empower California voters to make informed, confident choices by enabling the public to verify the authenticity of media information.

Courage California writes:

> As new technologies propel us into the future, we must move forward in a way that maximizes transparency, mitigates harm, and protects users. AB 3211, which borrows heavily from the guidance and leadership of the European Union's AI Act, accomplishes these goals.

ARGUMENTS IN OPPOSITION

A coalition of industry associations including CalChamber, Netchoice, Technet, and the Computer & Communications Industry Association writes:

> AB 3211 seems to treat all AI-generated content as inherently bad or risky. By requiring such thorough and prescriptive requirements for content labeling, the bill makes a value judgment that consumers must be notified and aware of any content that was created by AI. This applies to all inauthentic content including purely artistic or satirical content. While we agree with the intent to provide more information to consumers, in some instances it could create disclosure or notification fatigue. If watermarks and content credentials become so routine and placed on all AI-generated content, users may start to ignore and disregard their presence. Rather than focusing on whether the content itself was AI-generated, synthetic, or inauthentic, we would advise focusing on the misuse of this technology.

REGISTERED SUPPORT / OPPOSITION:

Support

California Initiative for Technology & Democracy, a Project of California Common CAUSE (CITED) (sponsor)
Asian Americans and Pacific Islanders for Civic Empowerment
Asian Law Alliance
California Initiative for Technology & Democracy, a Project of California Common CAUSE
California Voter Foundation
Courage California

Support If Amended

Oakland Privacy

Opposition

California Chamber of Commerce
Computer & Communications Industry Association
Netchoice
Technet

Analysis Prepared by: Slater Sharp and Josh Tosney / P. & C.P. / (916) 319-2200