SENATE JUDICIARY COMMITTEE
Senator Thomas Umberg, Chair
2025-2026 Regular Session

AB 853 (Wicks)
Version: May 23, 2025
Hearing Date: July 15, 2025
Fiscal: Yes
Urgency: No
CK

SUBJECT

California AI Transparency Act

DIGEST

This bill establishes requirements on large online platforms, capture device manufacturers, and generative AI (GenAI) system hosting platforms to embed and disclose provenance data in certain GenAI created or altered content.

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 created, 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."

This bill builds on the California AI Transparency Act (CAIT Act), passed last year. It places obligations on large online platforms to use a label to disclose certain provenance data that is accessible to users. To begin to create a paper trail of authenticity, manufacturers of devices that can record photographs, audio, or video content, like cameras and phones, must embed latent disclosures in content captured by them and allow users to include latent disclosures. Websites that make GenAI systems available for use are prohibited from making a system available unless it places disclosures into the content it creates or modifies that are permanent or difficult to remove.

This bill is sponsored by the California Initiative on Technology and Democracy (CITED). It is supported by several groups, including Consumer Reports. The bill is opposed by several industry associations, including the Consumer Technology Association.

PROPOSED CHANGES TO THE LAW

Existing law:

1) Establishes the California AI Transparency Act, which becomes operative on January 1, 2026. (Bus. & Prof. Code § 22757 et seq.)

2) Requires a "covered provider," a person that creates, codes, or otherwise produces a GenAI system that has over 1,000,000 monthly visitors or users and is publicly accessible within the geographic boundaries of the state, to make an AI detection tool available at no cost by which a person can assess whether content was created or altered by the provider's GenAI system. (Bus. & Prof. Code § 22757.2(a).)

3) Prohibits a covered provider from doing any of the following in carrying out the duties above:
    a) Collect or retain personal information when a person utilizes the covered provider's AI detection tool, except that it may collect and retain the contact information of a person who submitted feedback.
    b) Retain any content submitted to the AI detection tool for longer than is necessary to comply with this law. (Bus. & Prof. Code § 22757.2(c).)

4) Requires a covered provider to offer users the option to include in AI-generated image, video, or audio content created by its own generative AI system a manifest disclosure that meets specified criteria, including that it identifies the content as AI-generated content. (Bus. & Prof. Code § 22757.3(a).)

5) Requires a covered provider to include in AI-generated image, audio, and video content created by its generative AI system a latent disclosure that is detectable by the tool specified above and is, to the extent technically feasible, permanent or extraordinarily difficult to remove. (Bus. & Prof. Code § 22757.3(b).)

6) Provides that a covered provider that violates the above provisions is liable for a civil penalty in the amount of $5,000 per violation to be collected in a civil action filed by the Attorney General, a city attorney, or a county counsel. Each day that a covered provider is in violation shall be deemed a discrete violation. (Bus. & Prof. Code § 22757.4.)

This bill:

1) Requires a large online platform to do both of the following:
    a) Use a label to disclose any machine-readable provenance data detected in content distributed on the platform that meets all of the following criteria:
        i. The label indicates whether provenance data is available.

ii. The label indicates the name and version number of the GenAI system that created or altered the content, if applicable.
iii. The label indicates whether any digital signatures are available.
iv. The label is presented in a conspicuous manner to users.

b) Allow a user to inspect any provenance information in an easily accessible manner.

2) Prohibits a large online platform from doing the following:
   a) Stripping any system provenance data or digital signature from content uploaded or distributed on the platform.
   b) Retain any personal provenance data from content shared on the large online platform.

3) Defines "digital signature" as a cryptography-based method that identifies the user or entity that attests to the information provided in the signed section. "Large online platform" means a public-facing social media platform, content-sharing platform, messaging platform, advertising network, stand-alone search engine, or web browser that distributes content to users who did not create or collaborate in creating the content that exceeded 2,000,000 unique monthly users during the preceding 12 months.

4) Requires a capture device manufacturer, with respect to any capture device the capture device manufacturer produces for sale in the state, to do all of the following:
   a) Provide a user with the option to include a latent disclosure in content captured by the capture device that, to the extent that it is technically feasible and reasonable, conveys specified information, including identifying information for the manufacturer and the device, as well as the time and date of the content's creation or alteration. This option must be available for the capture device's default capture application and third-party applications, as specified.
   b) Embed latent disclosures in content captured by the device by default.
   c) Clearly inform users of the existence of settings relating to provenance data upon a user's first use of the recording function on the capture device.
   d) When any capture device application is in use, contain a clear indicator when provenance data is applied.
   e) Include in the capture device's default capture application the ability for a user to opt out of the inclusion of provenance data based on guidelines or specifications promulgated by an established standard-setting body in the user's captured content.
   f) Make secure hardware-based provenance data capture available to third-party applications.

5) Defines “capture device” as a device that can record photographs, audio, or video content, including video and still photography cameras, mobile phones with built-in cameras or microphones, and voice recorders.

6) Prohibits a GenAI system hosting platform from making available a GenAI system that does not place disclosures pursuant to existing Section 22757.3 that are permanent or extraordinarily difficult to remove into content created or substantially modified by the GenAI system.

7) Defines “GenAI hosting platform” as a website that makes a GenAI system available for use by a resident of the state, regardless of whether the terms of that use include compensation.

8) Prohibits a provider or a distributor of software or online services from making available a system, application, tool, or service that is designed for the primary purpose of removing latent disclosures applied pursuant hereto.

9) Subjects those in violation of these provisions to the liability currently imposed by the California AI Transparency Act.

10) Includes a severability clause.

COMMENTS

1. Blurring reality: AI-generated content

GenAI can create new content, including text, images, code, or music, by learning from existing data. GenAI models can produce realistic and novel artifacts that resemble the data they were trained on, but do not copy it. For example, GenAI can write a poem, draw a picture, or compose a song based on a given prompt or theme. It enables users to quickly generate new content based on a variety of inputs.

The world has been in awe of the powers of this new technology but 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]

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, last year, Taylor Swift was the victim of sexually explicit, nonconsensual GenAI deepfake images that were widely spread across social media platforms.[2]

More recently, the impact of GenAI-created content has hit closer to home during the protests in Los Angeles:

> The spread of falsified information, especially images, is emerging as a troubling issue in recent civil disturbances and demonstrations — such as the George Floyd protests of 2020. Back then, outdated videos of explosions from different countries, or arrests made months before in a different city, made the rounds on social media, stoking fear of extreme violence.
>
> The introduction of AI, however, marks new territory. The latest rollout of accessible AI video generators presents unique challenges to the truth — and public perception of it — because videos are "more powerful as a medium in terms of convincing people of reality," said Jamie Cohen, an assistant professor at CUNY Queens College who studies internet literacy.
>
> "Pictures are easily manipulated," he said. "That idea has been there. But when it comes to videos, we've just been trained as an individual society to believe videos. Up until recently, we haven't really had the opportunity to assume videos could be faked at the scale that it's being faked at this point."

[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 23, 2025.

[2] 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/.

> It's not just the sheer amount of slop that is filling social media feeds that's posing a problem. It's the ability to visually manipulate scenes to fit the creator's political agenda that is making it so much harder to decipher the truth during the L.A. protests. Consider these completely fake AI-generated videos posted over the past few days: one of a hypocritical protester who preaches peace and then throws a molotov cocktail. Or another of a man screaming "Viva Mexico," but then cowering away from an officer who says he will take him to Mexico. These clips aren't just delivering conservative talking points either: This one emanates from the left, featuring a young man delivering a heartfelt speech about standing with his community and fighting injustice.
>
> Until recently, the concern with manipulated images has been related to their ability to misinform a crowd and sway public opinion. But a growing body of research proves that not much can change people's minds — not even AI misinformation. What the technology can do, however, is reinforce preexisting beliefs, leaving people impervious to actual facts.[3]

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.[4] 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 creates to challenge the authenticity of real content, a phenomena coined the "liar's dividend."[5]

2. <u>Taking action to identify synthetic content and address its usage</u>

Last year the European Union AI Act was enacted. 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-

[3] Catherine Kim, *The LA protests are drowning in AI slop* (June 11, 2025) Politico, https://www.politico.com/newsletters/politico-nightly/2025/06/11/the-la-protests-are-drowning-in-ai-slop-00401401.

[4] 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.

[5] 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.

> 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.

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), [to] 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. Meta has committed to "label images that users post to Facebook, Instagram and Threads when we can detect industry standard indicators that they are AI-generated."[6] They stated their goal:

[6] 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/.

> We want people to know when they see posts that have been made with AI. Earlier this year, we announced a new approach for labeling AI-generated content. An important part of this approach relies on industry standard indicators that other companies include in content created using their tools, which help us assess whether something is created using AI.

A group of tech companies, including OpenAI, Adobe, Google, and Microsoft, has established the Coalition for Content Provenance and Authenticity (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."[7] Recently, OpenAI announced advances in this space with regard to one of their newest models:

> OpenAI, a leading research organization in artificial intelligence, has taken a significant step forward by integrating advanced image generation capabilities into its latest model, GPT-4o. This update not only expands the functionalities of GPT-4o but also addresses the growing concern of digital authenticity and transparency in AI-generated content.
>
> The new feature allows GPT-4o to generate high-quality images that can be used for various applications, from content creation to visual design. However, to ensure that these images are not mistaken for real-world content, OpenAI has implemented a crucial security measure. All images generated by GPT-4o will include C2PA (Content Authenticity Initiative) metadata. This metadata serves as a digital signature, clearly indicating that the image is AI-generated and not a genuine photograph.[8]

3. Requiring tools to identify synthetic content

Last year, the Legislature responded to these issues by passing the California AI Transparency Act (CAIT Act), SB 942 (Becker, Ch. 291, Stats. 2024), which is set to become operative on January 1, 2026. The CAIT Act imposed obligations on "covered providers," persons that create, code, or otherwise produce a GenAI system that has over 1,000,000 monthly visitors or users and is publicly accessible within the geographic boundaries of the state. It requires such providers to make an AI detection tool available at no cost by which a person can assess whether content was created or altered by the provider's GenAI system.

The CAIT Act also regulates AI-generated images, video, or audio that are created by a GenAI system. Covered providers are required to include a latent disclosure in such

[7] *Overview*, Coalition for Content Provenance and Authenticity, https://c2pa.org/.
[8] *OpenAI Launches Advanced Image Gen in GPT-4 with C2PA Metadata* (March 26, 2025) Visive.ai, https://www.visive.ai/news/openai-launches-advanced-image-gen-in-gpt-4-with-c2pa-metadata.

content that is detectable using the above tool, and that is, to the extent technically feasible, permanent or extraordinarily difficult to remove. This latent disclosure must identify the provider, the tool, and the time and date of the content's creation or alteration.

Covered providers are also required to provide users making such content with their system with the option to include a manifest disclosure that identifies it as AI-generated content.

A covered provider that violates the CAIT Act is liable for a civil penalty in the amount of $5,000 per violation to be collected in a civil action filed by the Attorney General, a city attorney, or a county counsel. Each day that a covered provider is in violation is a discrete violation.

4. <u>Expanding the scope of the CAIT Act</u>

This bill seeks to bolster the CAIT Act by establishing similar transparency requirements on large online platforms, capture device manufacturers, and GenAI system hosting platforms.

a. Large online Platforms

This bill lays out a set of obligations and prohibitions on large online platforms. The bill requires platforms to use a conspicuous label to disclose any machine-readable provenance data detected in content distributed on the platform. The label must indicate whether provenance data or any digital signatures are available and the name and version number of the GenAI system that created or altered the content, if applicable. Users must be able to inspect any provenance information in an easily accessible manner. The bill prohibits platforms from stripping any such system provenance data or digital signature from the content uploaded or distributed on the platform, but also prohibits platforms from retaining any personal provenance data from content shared.

"Digital signature" is defined as a cryptography-based method that identifies the user or entity that attests to the information provided in the signed section. Under existing law, "system provenance data" is provenance data that is not reasonably capable of being associated with a particular user and that contains information regarding the type of device, system, or service that was used to generate a piece of digital content or information related to content authenticity. "Personal provenance data" is provenance data that contains either personal information or unique device, system, or service information that is reasonably capable of being associated with a particular user, but excludes any information contained in a digital signature.

b. Capture device manufacturers

This bill also places obligations on "capture device manufacturers." "Capture device" means 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. Manufacturers of these devices must provide users with the option to include latent disclosures in content that is captured by the device by default. Users must be informed of the provenance data settings and an indicator must alert a user when provenance data is being applied while a capture device application is in use. The latent disclosure must indicate the name of the manufacturer and the device and the time and date of the content's creation or alteration, to the extent that it is technically feasible and reasonable.

Concerns have been raised about the technical feasibility of some of these requirements. While certain latent disclosures have been embedded in such devices for years, technology allowing for such disclosures to be embedded into other recordings is still developing. Given these realities the author has agreed to an amendment that delays the operative date of these provisions by two years.

c. GenAI system hosting platforms

Finally, the bill prohibits a GenAI system hosting platform from making available a GenAI system that does not place disclosures pursuant to existing Section 22757.3, discussed above, that are permanent or extraordinarily difficult to remove into content created or substantially modified by the GenAI system. "GenAI hosting platform" means a website that makes a GenAI system available for use by a resident of the state, regardless of whether the terms of that use include compensation. To ensure the integrity of these latent disclosures, the bill prohibits specified parties from making available a system, application, tool, or service that is designed for the primary purpose of removing such disclosures.

5. <u>Stakeholder positions</u>

According to the author:

> New and emerging developments of generative AI (GenAI) tools have made it easier to create, edit, and doctor images, video, and audio. AI technologies can create and manipulate content to look realistic and convincing, which allows bad actors to create harmful content and spread disinformation. AB 853 will help mitigate some of the harmful impacts of AI-generated content and provide more transparency of content in the digital information ecosystem by ensuring that large online platforms and capture devices provide more information for users to understand the source of content.

CITED, the sponsor of the bill, explains the need for it:

> Generative artificial intelligence (GenAI) technologies are powerful tools capable of creating all manners of images, audio, video, and text content from simple prompts. The breakneck speed at which these tools have evolved has meant human beings are increasingly unable to tell the difference between authentic, human-generated content, and synthetic content generated by AI.[9]
>
> The impact of this increasingly blurry line between authentic and synthetic digital media is already being felt by our society. From supercharging online scams,[10] to using child sexual abuse material to generate non-consensual intimate imagery,[11] to the proliferation of public safety[12] and political disinformation,[13] GenAI tools have contributed to the steady erosion of trust in our information ecosystem. Without adequate tools to help differentiate between human-generated authentic content and AI-generated synthetic content, the truth decay[14] already happening in our society will only accelerate.

Writing in opposition, the Consumer Technology Association argues:

> AI is not static—it is dynamic infrastructure that evolves at exponential velocity. Regulating at the application layer, before consensus has formed on the tooling, provenance standards, or technological feasibility, risks freezing innovation in place. That is the essential flaw of AB 853: it attempts to legislate before the tools for compliance are widely available, widely adopted, or even fully developed.
>
> CTA supports meaningful transparency and accountability in AI-generated content. In fact, we've published ANSI/CTA-2125, a standard

[9] Nightingale & Farid, *AI-synthesized faces are indistinguishable from real faces and more trustworthy, Proceedings of the National Academy of Sciences* (Feb. 14, 2022), https://www.pnas.org/doi/full/10.1073/pnas.2120481119.

[10] Bob Violino, *AI tools such as ChatGPT are generating a mammoth increase in malicious phishing emails*, CNBC (Nov. 28, 2023), https://www.cnbc.com/2023/11/28/ai-like-chatgpt-is-creating-huge-increase-in-malicious-phishing-email.html.

[11] Thiel, David, *Identifying and Eliminating CSAM in Generative ML Training Data and Models*, Stanford Digital Repository (Dec. 20, 2023), https://purl.stanford.edu/kh752sm9123.

[12] Shannon Bond, *Fake viral images of an explosion at the Pentagon were probably created by AI*, NPR (May 22, 2023), https://www.npr.org/2023/05/22/1177590231/fake-viral-images-of-an-explosion-at-the-pentagon-were-pro bably-created-by-ai.

[13] Alex Isenstadt, *Desantis PAC uses AI-generated Trump voice in ad attacking ex-president*, Politico (Jul. 17, 2023), https://www.politico.com/news/2023/07/17/desantis-pac-ai-generated-trump-in-ad-00106695.

[14] J. Kavanagh & M. Rich, *Truth Decay: An Initial Exploration*, Rand, (Jan 16, 2018), https://www.rand.org/pubs/research_reports/RR2314.html.

> which can address content provenance and assurance by providing a foundation for detecting and labeling AI-generated media in a consistent way. But like all standards, it requires time, industry convergence, and implementation capacity. AB 853 ignores this timeline, imposing requirements without the supporting ecosystem.

Writing in support, Consumer Reports states:

> Last year, SB 942 was enacted to ensure provenance information will be embedded into AI-generated content that will allow users to identify its origins. AB 853 complements this effort by adding two additional tools to make provenance information more useful:
>
> - At the point of content creation, AB 853 enables provenance markings on authentic, human-generated content by requiring that recording devices sold in California include the option to embed such information.
> - At the point of content dissemination, the bill requires social media and online platforms to display the source of content shared on their platforms, leveraging provenance data.
>
> Together with the foundation laid by SB 942, AB 853 empowers consumers to distinguish between AI-generated and human-created content, helping to slow the tide of misinformation. It equips individuals with the tools they need to make informed decisions about the trustworthiness of the media they encounter. It also would accelerate the adoption of and build upon voluntary provenance standards that major tech companies are currently developing, such as those proposed by the Coalition for Content Provenance and Authenticity (C2PA).

SUPPORT

California Initiative on Technology and Democracy (sponsor)
Consumer Reports
Tech Oversight Project dba Tech Oversight California
TechEquity Action
Transparency Coalition.ai
Truepic

OPPOSITION

California Civil Liberties Advocacy
Computer & Communications Industry Association
Consumer Technology Association
Technet

RELATED LEGISLATION

Pending Legislation: None known.

Prior Legislation:

SB 942 (Becker, Ch. 291, Stats. 2024) *See* Comment 3.

AB 3211 (Wicks, 2024) would have established the California Digital Content Provenance Standards Act, which requires a GenAI provider to, among other things, take certain actions to assist in the disclosure of provenance data. It would have required an online platform, as defined, to, among other things, use labels to disclose provenance data found in synthetic content, as specified. It also would have required recording device manufacturers to enable options for embedding provenance data into recordings. AB 3211 died on the Senate Floor.

PRIOR VOTES:

Assembly Floor (Ayes 58, Noes 2)
Assembly Appropriations Committee (Ayes 11, Noes 0)
Assembly Judiciary Committee (Ayes 9, Noes 0)
Assembly Privacy and Consumer Protection Committee (Ayes 11, Noes 1)
****