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

AB 1791 (Weber)
Version: June 3, 2024
Hearing Date: June 18, 2024
Fiscal: Yes
Urgency: No
CK

SUBJECT

Digital content provenance

DIGEST

This bill requires social media platforms to redact "personal provenance data" from content uploaded on their platforms while retaining "system provenance data" that reveals information about the creation of the content. The bill requires platforms to redact both if inseparable, and to append a label with the latter data.

EXECUTIVE SUMMARY

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

This bill is focused on the provenance data attached to content posted on social media. It requires social media platforms to redact all "personal provenance data," which is the embedded information in content that reveals personal information or other data that can be associated with a particular user. But to ensure that synthetic content can be separated from what is authentic, it requires platforms to maintain "system provenance data," or the embedded data that serves to verify the content's authenticity, origin, or history of modification. Where the two cannot be separated, the platform must redact it all and provide a label revealing the latter.

This bill is author-sponsored. No timely support was received by this Committee. It is opposed by industry associations, including TechNet.

PROPOSED CHANGES TO THE LAW

Existing law:

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

2) Defines "digital content forgery" as the use of technologies, including artificial intelligence and machine learning techniques, to fabricate or manipulate audio, visual, or text content with the intent to mislead. (Gov. Code § 11547.5.)

3) Defines "digital content provenance" as the verifiable chronology of the original piece of digital content, such as an image, video, audio recording, or electronic document. (Gov. Code § 11547.5.)

4) Defines "personal information" to mean information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household. (Civ. Code § 1798.140(v).)

5) Defines "social media platform" as a public or semipublic internet-based service or application that has users in California and that meets both of the following criteria:
    a) A substantial function of the service or application is to connect users in order to allow users to interact socially with each other within the service or application. A service or application that provides email or direct messaging services shall not be considered to meet this criterion on the basis of that function alone.
    b) The service or application allows users to do all of the following:
        i. Construct a public or semipublic profile for purposes of signing into and using the service or application.
        ii. Populate a list of other users with whom an individual shares a social connection within the system.
        iii. Create or post content viewable by other users, including, but not limited to, on message boards, in chat rooms, or through a landing page or main feed that presents the user with content generated by other users. (Bus. & Prof. Code § 22675(e).)

This bill:

1) Requires a social media platform to redact personal provenance data from content uploaded to the platform by a user.

2) Prohibits a social media platform from redacting system provenance data from content uploaded to the social media platform by a user, except as provided.

3) Provides that if a social media platform is unable to redact personal provenance data from content without also redacting system provenance data from the content, a social media platform must redact both from the content and shall append a label to the content that meets all of the following criteria:
    a) The label prominently discloses any system provenance data that was redacted by the social media platform.
    b) The label remains appended to the content even if the content is shared, reposted, or otherwise replicated within the social media platform.
    c) The label does not disclose any personal provenance data.

4) Requires the social media platform, when content to which they have appended a label is downloaded, shared to an external internet website, or otherwise distributed in a manner that does not permit the platform to control how the content is displayed, to embed the information contained in the label into the distributed content or add it to the content's metadata.

5) Requires a social media platform to abide by relevant industry standards to the greatest extent possible when redacting provenance data, labeling content, embedding information into content, or adding information to metadata pursuant to this section.

6) Provides that a violation constitutes an unfair business practice punishable pursuant to the Unfair Competition Law, Section 17200 et seq., of the Business and Professions Code.

COMMENTS

1. Blurring reality: AI-generated content

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

networks to identify the patterns and structures within existing data to generate new and original content.

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

> Over the last year, generative AI tools have made the jump from research prototype to commercial product. Generative AI models like OpenAI's ChatGPT and Google's Gemini can now generate realistic text and images that are often indistinguishable from human-authored content, with generative AI for audio and video not far behind. Given these advances, it's no longer surprising to see AI-generated images of public figures go viral or AI-generated reviews and comments on digital platforms. As such, generative AI models are raising concerns about the credibility of digital content and the ease of producing harmful content going forward.
>
> Against the backdrop of such technological advances, civil society and policymakers have taken increasing interest in ways to distinguish AI-generated content from human-authored content.[1]

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

---

[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 9, 2024.

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

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

bodies. The deepfake images can be produced using a cellphone."[4] In February of this year, voters in New Hampshire received robocalls that are purported to have used an AI voice resembling President Joe Biden advising them against voting in the presidential primary and saving their vote for the November general election.[5] Recently, a former federal judge urged the federal judiciary's Advisory Committee on Evidence Rules to update evidentiary rules regarding the admissibility of evidence believed to be AI generated.[6] But, in addition to concerns about the potential for AI-generated material to be admitted into evidence is the reverse, false claims that authentic 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."[7]

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

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

> A variety of AI systems can generate large quantities of synthetic content that becomes increasingly hard for humans to distinguish from human-generated and authentic content. The wide availability and increasing capabilities of those systems have a significant impact on the integrity and trust in the information ecosystem, raising new risks of misinformation and manipulation at scale, fraud, impersonation and consumer deception. In light of those impacts, the fast technological pace and the need for new methods and techniques to trace origin of information, it is appropriate to require providers of those systems to embed technical solutions that enable marking in a machine readable format and detection that the output has been generated or manipulated by an AI system and not a human. Such techniques and methods should be sufficiently reliable, interoperable, effective and robust as far as this is technically feasible,

---

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

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

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

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

> 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), should also clearly and distinguishably disclose that the content has been artificially created or manipulated by labelling the artificial intelligence output accordingly and disclosing its artificial origin."

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

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

[8] *See* footnote 1.

3. Ensuring sufficient tools to establish the provenance of synthetic and nonsynthetic content

This bill looks to address the issues of determining digital provenance of content posted to social media platforms while protecting against very real privacy concerns.

According to the author:

> AI generated images and video are becoming more easily accessible and convincing every day. There are serious consequences to deepfakes from our political dialogue, to rattling the stock market, and fraud. It is why being able to authenticate digital content is incredibly important. AB 1791 is allowing users who decide to opt-in to add transparency to their content, will not be removed by the platform.

The bill defines "provenance data" as data that is embedded into digital content, or that is included in the digital content's metadata, for the purpose of verifying the digital content's authenticity, origin, or history of modification. It then breaks that down into two separate categories. The first is "personal provenance data," defined as provenance data that contains personal information, as defined in the California Consumer Privacy Act; unique device, system, or service information that is reasonably capable of being associated with a particular user; and time-of-day information. "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 that provides proof of content authenticity.

The bill requires platforms to redact the personal provenance date and prohibits them from redacting the system provenance data. Where the platform is unable to separate the two, it must redact it all and label the content, disclosing the system provenance data without disclosing personal provenance data. The label must remain appended even when shared, reposted, or otherwise replicated within the platform.

When content to which a social media platform has appended a label is downloaded, shared to an external internet website, or otherwise distributed in a manner that does not permit the platform to control how the content is displayed, the platform is required to embed the information contained in the label into the distributed content or add it to the content's metadata.

The bill obligates the platforms to abide by relevant industry standards to the greatest extent possible when redacting provenance data, labeling content, embedding information into content, or adding information to metadata pursuant to this section.

The author recently removed the private enforcement mechanism in the bill. The bill now makes clear that a violation of it constitutes an unfair business practice punishable pursuant to the Unfair Competition Law.

4. Opposition concerns

A coalition of industry associations, including the California Chamber of Commerce, writes in opposition:

> Conceptually, we've agreed that online platforms should not remove what this bill now calls "system provenance data". We also previously noted that platforms should be allowed to remove certain personal information found in the metadata of user content to protect users' privacy and security. We have long agreed that users should ultimately have control over their personal information and believe in this case they should decide whether to remove or include personal information in their content.
>
> The new amendments alter those concepts to instead require the removal of personal information in all instances, now referred to as "personal provenance data", and to remove both if personal provenance data cannot be removed without removing the system provenance data. This seems to remove the ability for a user, say a photographer or digital artist, to keep their identification as part of the embedded provenance.

In response to these reasonable concerns about centering the consumer's wishes rather than requiring a one-size-fits-all approach, the author has agreed to amendments that provide platforms the ability to forego removing personal provenance data if they obtain the user's express consent to forgo removing that data from either a specific piece of content or for all content uploaded by the user.

The Recording Industry Association of America highlights several reasons for their opposition, some of which are addressed or at least mitigated by the above amendment. In addition, they argue: "Removal of such data harms our legitimate enforcement efforts to protect our members' copyrighted sound recordings, particularly with respect to sound recordings that have been stolen and uploaded to a social media platform before the sound recording has been commercially released." The association further argues that the "removal obligations contradict the edicts of 17 U.S.C. 1202, which makes it unlawful to remove such data, to the extent such data qualifies as copyright management information, in certain circumstances."

In response, the author has agreed to amendments that make clear where an obligation to remove any provenance data conflicts with 17 U.S.C. § 1202, the latter takes precedent.

Opposition also argues that the technology is not sophisticated enough to accomplish the redactions required by the bill. Similar concerns have been raised in response to other measures regarding the feasibility of various digital provenance requirements given the nascent stage of their development. To provide sufficient runway for the industry to develop the proper technology to accomplish the lofty goals of this bill, the author has agreed to amendments that push out the effective date of these requirements to January 1, 2028.

SUPPORT

None received

OPPOSITION

California Chamber of Commerce
Computer and Communications Industry Association
Netchoice
Recording Industry Association of America
Technet

RELATED LEGISLATION

Pending Legislation:

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

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

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

decision that an automated decision tool is being used to make, or be a controlling factor in making, the consequential decision and to provide that person with, among other things, a statement of the purpose of the automated decision tool. AB 2930 is currently in this Committee.

AB 3211 (Wicks, 2024) establishes the California Provenance, Authenticity and Watermarking Standards Act, which requires a generative AI system provider to take certain actions to assist in the disclosure of provenance data to mitigate harms caused by inauthentic content, including placing imperceptible and maximally indelible watermarks containing provenance data into content created by its systems. The bill requires a large online platform, as defined, to, among other things, use labels to prominently disclose the provenance data found in watermarks or digital signatures in content distributed to users on its platforms, as specified, and to use state-of-the-art techniques to detect and label inauthentic text content that is uploaded or distributed by individual users or networks of users. The bill also requires recording device manufacturers to enable options for embedding provenance data into recordings. AB 3211 is currently in this Committee.

Prior Legislation: SB 1216 (Gonzalez, 2022) required the Secretary of Government Operations, upon appropriation by the Legislature, to evaluate, among other things, the impact the proliferation of deepfakes has and the risks, including privacy risks, associated with the deployment of digital content forgery technologies and deepfakes on government, businesses, and residents of the state. It required the secretary to develop a coordinated plan to accomplish specified objectives, including investigating the feasibility of, and obstacles to, developing standards and technologies for state departments for determining digital content provenance. It requires the secretary, on or before October 1, 2024, to report to the Legislature on the potential uses and risks of deepfake technology to the state and businesses, as specified.

PRIOR VOTES:

Assembly Floor (Ayes 50, Noes 10)
Assembly Privacy and Consumer Protection Committee (Ayes 7, Noes 1)

****