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

SB 942 (Becker)
Version: March 20, 2024
Hearing Date: April 16, 2024
Fiscal: Yes
Urgency: No
CK

SUBJECT

California AI Transparency Act

DIGEST

This bill places obligations on businesses that provide generative artificial intelligence (AI) systems to develop and make accessible tools to detect whether specified content was generated by those systems. These "covered providers" are required to include visible and imperceptible markings on AI-generated content to identify it as such. The bill requires providers to register with the California Department of Technology (CDT).

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 generative AI 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 works to ensure that providers of these generative-AI systems are equipping consumers with a tool to identify when specific content has been generated by their systems. This required tool must be made available to the public, as provided. Covered providers are also required to place markings, both visible and imperceptible, on this generated content to identify it as such and to be able to trace it back to the covered providers systems. Providers must register with CDT and work to prevent downstream users from removing the markings. This bill is author-sponsored. No support has been received by the Committee. The bill is opposed by a coalition of industry groups, including Netchoice. Should this pass out of this Committee, it will be referred to the Senate Governmental Organization Committee.

PROPOSED CHANGES TO THE LAW

Existing law:

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

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

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

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

5) Requires the Secretary of GovOps to develop a coordinated plan to accomplish all of the following:
    a) Investigate the feasibility of, and obstacles to, developing standards and technologies for state departments for determining digital content provenance.
    b) Increase the ability of internet companies, journalists, watchdog organizations, other relevant entities, and members of the public to

meaningfully scrutinize and identify digital content forgeries and relay trust and information about digital content provenance to content consumers.
c) Develop or identify mechanisms for content creators to cryptographically certify authenticity of original media and nondeceptive manipulations.
d) Develop or identify mechanisms for content creators to enable the public to validate the authenticity of original media and nondeceptive manipulations to establish digital content provenance without materially compromising personal privacy or civil liberties. (Gov't Code § 11547.5(c).)

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

7) Provides that the duties of the Director include:

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

This bill:

1) Establishes the California AI Transparency Act.

2) Requires a covered provider to create an AI detection tool by which a person can query the covered provider as to the extent to which text, image, video, audio, or multimedia content was created, in whole or in part, by a generative AI system provided by the covered provider that meets all of the following criteria:
a) The AI detection tool shall be publicly accessible and available via a uniform resource locator (URL) on the covered provider's internet website and through its mobile application, as applicable.
b) The AI detection tool shall allow a person to upload content or a URL.

   c) The AI detection tool shall support an application programming interface (API) that allows a person to invoke the AI detection tool without visiting the covered provider's website.
   d) The AI detection tool shall allow a person to provide feedback if the person believes the AI detection tool is not properly identifying content that was created by the provider.

3) Prohibits a covered provider from doing any of the following in carrying out the duties above:
   a) Reveal personal information that identifies who utilized the AI system to create AI-generated content that was submitted to the AI detection tool.
   b) Collect and 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.
   c) Retain any content submitted to the AI detection tool for longer than is necessary to comply with this law.

4) Requires a covered provider to include in AI-generated image, text, video, or multimedia content created by its own generative AI system a visible disclosure that meets all of the following criteria:
   a) The disclosure shall include a clear and conspicuous notice, as appropriate for the medium of the content, that identifies the content as generated by AI, such that the disclosure is not avoidable, is understandable to a reasonable person, and is not contradicted, mitigated by, or inconsistent with anything else in the communication.
   b) The disclosure shall, to the extent technically feasible, be permanent or difficult to remove.
   c) The output's metadata information shall include an identification of the content as being generated by AI, the identity of the tool used to create the content, and the date and time the content was created.

5) Requires a covered provider to include in AI-generated image, audio, video, or multimedia content created by its generative AI system an imperceptible disclosure that is machine detectable and is, to the extent technically feasible, permanent or difficult to remove.

6) Requires a covered provider to implement reasonable procedures to prevent downstream use of a generative AI system it provides without the disclosure required by this section, including the following:
   a) Contractually requiring third-party licensees of the generative AI system to refrain from removing a required disclosure.
   b) Terminating access to the generative AI system when the covered provider has reason to believe that a third-party licensee has removed a required disclosure.

7) Requires CDT, at least once every two years, to review these provisions and make recommendations to the Legislature regarding any amendments needed to account for changing technology and standards.

8) Requires covered providers to register with CDT and provide a URL to any AI detection tool it has created. Requires CDT to create and display on its website a Generative AI Registry that displays the names of registered covered providers and a link to the covered provider's AI detection tool. CDT is authorized to charge a registry fee to covered providers to cover administrative costs.

9) Authorizes CDT to adopt regulations necessary to perform its duties hereunder.

10) 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 only by the Attorney General. Each day that a covered provider is in violation shall be deemed a discrete violation.

11) Defines the relevant terms, including:

    a) "Artificial intelligence" or "AI" means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments by using machine-based inputs and human-based inputs to perceive real and virtual environments, abstract its perceptions into models through analysis in an automated manner, and use model inference to formulate options for information or action.
    b) "AI-generated content" means any form of digital content that is created with deep learning or machine learning processes.
    c) "Covered provider" is a business that provides a generative AI system that has, on average over the preceding 12 months, over 1,000,000 monthly visitors or users and is publicly accessible within the geographic boundaries of the state.
    d) "Generative AI system" refers to deep learning models that can generate text, images, and other content based on the data they were trained on.

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 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 April 2, 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 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."[7]

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

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

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

---

[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.

> 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]

Recently, 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-

[8] *See* footnote 1.

generated."[9] A group of tech companies, including 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."[10] OpenAI announced that it will add C2PA metadata to images created with ChatGPT and the API for the DALL-E 3 model.

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

> President Biden [convened] seven leading AI companies at the White House [] – Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI – to announce that the Biden-Harris Administration has secured voluntary commitments from these companies to help move toward safe, secure, and transparent development of AI technology.
>
> Companies that are developing these emerging technologies have a responsibility to ensure their products are safe. To make the most of AI's potential, the Biden-Harris Administration is encouraging this industry to uphold the highest standards to ensure that innovation doesn't come at the expense of Americans' rights and safety.
>
> These commitments, which the companies have chosen to undertake immediately, underscore three principles that must be fundamental to the future of AI – safety, security, and trust – and mark a critical step toward developing responsible AI. As the pace of innovation continues to accelerate, the Biden-Harris Administration will continue to remind these companies of their responsibilities and take decisive action to keep Americans safe.[11]

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

> **Develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated, including robust provenance, watermarking, or both, for AI-generated audio or visual content**

---

[9] 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/.
[10] *Overview*, Coalition for Content Provenance and Authenticity, https://c2pa.org/.
[11] *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI* (July 21, 2023) The White House, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

> Companies making this commitment recognize that it is important for people to be able to understand when audio or visual content is AI-generated. To further this goal, they agree to develop robust mechanisms, including provenance and/or watermarking systems for audio or visual content created by any of their publicly available systems within scope introduced after the watermarking system is developed. They will also develop tools or APIs to determine if a particular piece of content was created with their system. Audiovisual content that is readily distinguishable from reality or that is designed to be readily recognizable as generated by a company's AI system—such as the default voices of AI assistants—is outside the scope of this commitment. The watermark or provenance data should include an identifier of the service or model that created the content, but it need not include any identifying user information. More generally, companies making this commitment pledge to work with industry peers and standards-setting bodies as appropriate towards developing a technical framework to help users distinguish audio or visual content generated by users from audio or visual content generated by AI.

3. Creating tools to identify synthetic content

This bill looks to implement some of these approaches to ensuring individuals can identify when content is AI generated by placing a series of obligations on "covered providers," businesses that provide a generative AI system that has met certain thresholds.

These covered providers are required to create AI detection tools that individuals can use to assess whether certain content was created by the provider's generative AI system. Individuals will be able to upload content to the tool, which must be publicly accessible through a URL on the provider's website. The bill places strict privacy limitations both for those that created the AI-generated content and those running queries on the tool, except as provided.

Using methods like those discussed in detail above, covered providers are required to include both clear and conspicuous and imperceptible disclosures in content generated by their systems to identify it as synthetic content. Despite the identified limitations for such markings, the author argues this process will help to create a digital paper trail for AI-generated content. Specifically, the bill requires the systems' outputs metadata to include not only identification of the content as synthetic, but also the date and time it was created and the system used to create it. These markings must be permanent or difficult to remove, to the extent technically feasible. To deter tampering with these disclosures, the bill requires providers to implement reasonable procedures to prevent downstream interference with the required disclosures, including requirements to

contractually obligate third-party licensees not to remove them and to terminate access when there is a reasonable basis to believe they have.

The bill creates an oversight role for CDT. Covered providers must register with and pay a fee to CDT and provide a URL to their detection tools. CDT is to create a registry that displays the providers with basic identifying information and links to their detection tools. CDT is authorized to issue regulations as necessary.

The Attorney General is charged with enforcement, authorizing civil penalties of $5,000 per violation per day.

The author states:

> Generative Artificial Intelligence (AI) is advancing at an unprecedented pace, revolutionizing industries and everyday life. However, this rapid progress has also sparked a pressing need for regulatory frameworks to govern its use. While AI offers innovative solutions that drive economic growth and efficiency, it simultaneously presents complex challenges related to misinformation, manipulation, bias, and ownership of generated content.
>
> One of the most concerning aspects of unregulated AI models is their potential to produce "deepfake" content, such as manipulated images and videos, which can be used to deceive or manipulate individuals and society at large. Deepfake pornography and fabricated political messages have already emerged as serious threats, highlighting the urgency of implementing effective regulations.
>
> As AI-generated content approaches a level of realism that makes it indistinguishable from genuine content, the need for transparency becomes paramount. It is crucial for industry leaders to take responsibility for clearly labeling AI-generated content and developing tools that can accurately identify such content, especially when there is no visible disclosure. This transparency is essential for maintaining trust in AI technologies and ensuring that users can make informed decisions about the content they consume.
>
> SB 942 stands as a critical legislative response in our ongoing regulatory arms race with artificial intelligence.

Concerns have been raised in response to this and other measures attempting to implement identification protocols or detection tools. Many existing tools suffer from inaccuracies and false positives. Various technical methods for synthetic content transparency, such as watermarking, fingerprinting, or metadata, are being developed

and are evolving but are far from foolproof, as individuals can, for instance, sometimes find ways to remove or obscure these disclosures from content. A coalition in opposition, including Technet, writes:

> Many of our companies and platforms are at the forefront of developing content provenance and watermarking technology, which is still in its early stages. However, SB 942 enacts requirements for a technology that is still under development and rapidly evolving. For example, there isn't a program that can watermark text, making the bill's requirements to do so impossible to comply with. We believe references to text watermarking should be removed to reflect this reality.
>
> Furthermore, content provenance and watermarking is still incredibly unreliable and in many cases easy to break. Researchers at the University of Maryland were able to break all the currently available watermarking methods. Some can be avoided by simple cropping, resizing, or screenshotting an image. More concerning, these researchers were able to insert fake watermarks and credentials into images, creating false positives.
>
> In its standards for large online platforms, SB 942 should more clearly delineate between 1st party and 3rd party content. 1st party content would be images, videos, or audio that is generated using a covered provider's generative AI tools and is then posted or distributed on that platform. In this instance, a covered provider can control the creation of a content provenance or watermark into the content. As mentioned, many of our companies are already working to incorporate this type of technology to increase transparency around AI-generated content. It is currently technically infeasible to accurately and reliably detect content that is created using a different platform's AI tools. As noted above, considering the current ease with which current watermarks can be broken, a legal requirement and mandate for 3rd party content isn't appropriate.

The author responds:

> While studies have shown the vulnerability of existing third-party AI-detection tools and industry watermarking efforts to bad actors, we don't believe it to be crucial enough to abandon transparency efforts. We recognize that watermarking efforts will not prevent all forms of abuse, but it will provide a measure of protection against harm. Our bill also recognizes the need for these safeguards to be continually adapted and refined as adversaries find novel ways to weaponize the latest technologies.

Writing in a support if amended position, Oakland Privacy also focuses in on the nascent stage of digital provenance tools and raises issues with this portion of the bill:

> The second part of the bill, and the part which we do not support, relates to a requirement for the providers to place an AI detection tool on their websites and requires the Department of Government Operations to set up a registry of providers and the URLs for their detection tools. We acknowledge that the bill is meant to be forward-thinking and that AI detection tools will improve over time, but the current state of development of these tools is so nascent and so currently flawed that state sanction of these tools as accurate and helpful tools is misguided and inappropriate.
>
> How bad are AI detection tools? Pretty bad. Content-At-Scale, one of the largest AI detection tools identified sections of the United States Constitution as highly likely to be AI generated. The most developed tools address text and rely on certain indicators including predictability, variety of sentence length and structure, word repetition, unnatural word usage, and inconsistent verb tense. It goes without saying that many of these are also characteristics of poorly written human-generated text. The tools are dealing in probabilities and while they may effectively flag bad writing that *could* be AI-generated, they are simply not accurate enough to be relied on as evidence of anything. We would also express a fairly high level of certainty based on these indicators that they would frequently target text written by non-native speakers of English, introducing yet another possible ramification of bias in AI tools.

The bill largely models legislation introduced at the federal level by Senators Schatz and Kennedy, the AI Labeling Act of 2023. Also at the federal level, President Biden issued an executive order late last year on AI, to set standards for safety and security and to ensure responsible AI development and deployment in the United States. The order defined generative AI as "the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content." The author may wish to incorporate this definition into the bill to ensure greater cohesion with federal efforts and to create efficiencies for compliance in the industry.

The EO calls for the development of a risk management framework for generative AI, modeled after the AI Risk Management Framework created by the National Institute for Standards and Technology (NIST). NIST reports that it is contributing to a report on authenticating, labeling, or detecting synthetic content. The report will identify existing, and potential future, standards, tools, methods, and practices for:

- Authenticating content and tracking its provenance.

- Detecting and labeling synthetic content, with techniques such as watermarking.
- Preventing generative AI from producing child sexual abuse material or any material containing non-consensual intimate imagery of real individuals.
- Testing software used for the above purposes.
- Auditing and maintaining synthetic content.[12]

Such guidance will assist covered providers in identifying best practices in carrying out their obligations pursuant to this bill.

4. Defining AI

Given the immense potential but attendant challenges and dangers of advancing AI technologies, the Legislature is currently considering dozens of bills on the subject of regulating and fostering AI. The first challenge is determining exactly what we mean by the term. A more thorough discussion of the various definitions that have been crafted by national and international entities for AI can be found in this Committee's analysis of SB 1047 (Wiener, 2024). In order to gain both the benefit of the expertise and compromise that went into formulating those definitions and the efficiencies that come with harmonization, the Committee, with a variety of stakeholders, including the author, have come up with the following definition to begin this process and to amend into the bill:

> "Artificial intelligence" means an engineered or machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs that can influence physical or virtual environments and that may operate with varying levels of autonomy.

SUPPORT

None received

OPPOSITION

California Chamber of Commerce
Computer and Communications Industry Association
Netchoice
Technet

[12] *Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence: Synthetic Content*, NIST, https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence-2.

RELATED LEGISLATION

Pending Legislation:

SB 970 (Ashby, 2024) ensures that media manipulated or generated by artificial intelligence (AI) technology is incorporated into the right of publicity law and criminal false impersonation statutes. The bill requires those providing access to such technology to provide a warning to consumers about liability for misuse. The bill also requires Judicial Council to create screening procedures to identify written evidence altered or created by AI technology and to provide educational materials to court users and personnel on identifying such materials. SB 970 is currently in the Senate Public Safety Committee.

SCR 17 (Dodd, 2023) affirms the California Legislature's commitment to President Biden's vision for a safe AI and the principles outlined in the "Blueprint for an AI Bill of Rights" and expresses the Legislature's commitment to examining and implementing those principles in its legislation and policies related to the use and deployment of automated systems. SCR 17 is currently in the Assembly Privacy and Consumer Protection 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 the Assembly Privacy and Consumer Protection Committee.

AB 3211 (Wicks, 2024) establishes the California Provenance, Authenticity and Watermarking Standards Act, which requires a generative AI system provider, as defined, to take certain actions to assist in the disclosure of provenance data to mitigate harms caused by inauthentic content, including placing imperceptible and maximally indelible watermarks containing provenance data into content created by an AI system that the generative AI system provider makes available. AB 3211 also 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. The bill would require a large online platform to use state-of-the-art techniques, including, but not limited to, analysis of user behavioral signals indicating usage of synthetic content, to detect and label inauthentic text content

that is uploaded or distributed by individual users or networks of users. AB 3211 is currently in the Assembly Privacy and Consumer Protection Committee.

Prior Legislation:

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

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

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

****