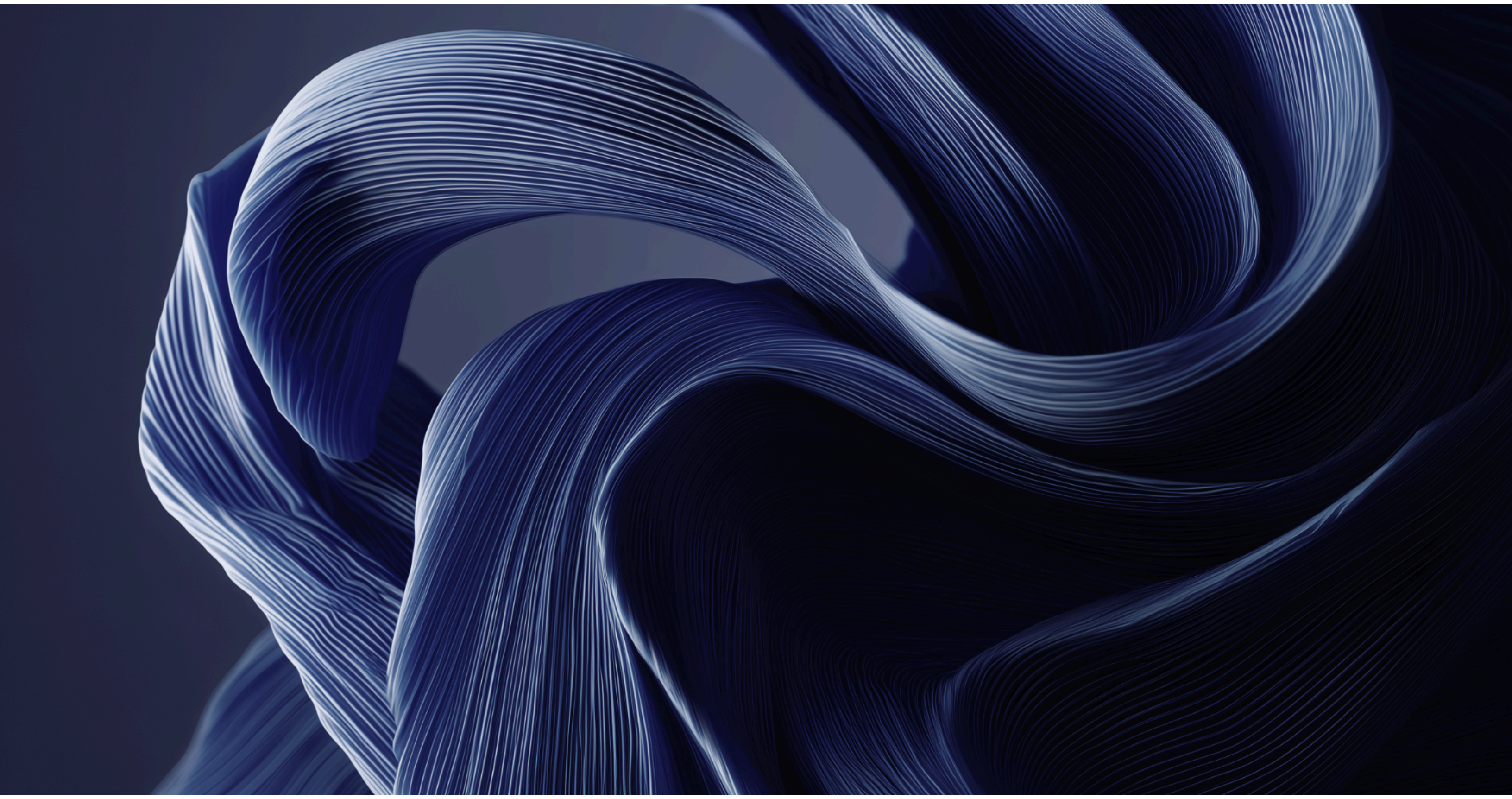
POLICY BRIEF | NOVEMBER 2025

Labeling of AI Agent Activity in Article 50 of the EU AI Act

Alan Chan



GovAI

Executive summary

- The online activities of AI agents could distort human beliefs and behaviors. For example, humans could mistake synthetic likes and shares on social media for genuine human opinion.

- Labeling such activity as AI-generated could help society understand and address its risks, such as by helping users be more conscious of attempts to influence them.

- Article 50 (Art. 50) of the EU AI Act requires companies to label certain AI activity and is expected to be in force from 2 August 2026. However, it is unclear whether actions from AI agents are to be labeled. An upcoming Code of Practice ("Code") will be an opportunity to clarify this issue by describing voluntary measures that, if followed, would constitute evidence of compliance with certain parts of Art. 50.[1]

- I argue that Art. 50 likely requires:

  - **Web requests (e.g. online payments) and browser actions from AI agents to be labeled**. Art. 50 requires labelling of AI outputs and actions plausibly count as such. Furthermore, labelling them would serve the Article's transparency goals, such as by helping humans distinguish AI activity from genuine human opinion. Finally, labeling of web requests and browser actions is feasible through metadata.

  - **Labels to be verifiable**. It should be possible to verify both who created a label (e.g. a particular developer) and whether it has been tampered with, because Art. 50 requires labeling to be "effective, interoperable, robust and reliable".

- To operationalize these considerations, I propose potential text for the Code in this article's Appendix.

- Even labeling all AI actions (similarly, all AI-generated content) as synthetic could fail to provide sufficient transparency if labels are ignored or do not convey useful information, for example if the majority of online content comes to be AI-generated.

This work represents the views of its authors, rather than the views of the organisation, and does not constitute legal advice. GovAI policy briefs are short and accessible pieces that have not undergone an official peer review process.

[1] The Code will cover paragraphs 2 and 4 of Article 50. The European Commission will separately develop guidelines that will facilitate compliance with the other paragraphs of Article 50.

Introduction

The online activities of AI agents could influence humans. Although such influence is not necessarily bad, it could also distort beliefs and behaviour. For example, when an agent shares or likes political content, humans may mistake such activity for genuine human opinion.

Labeling such activity could help society understand and address its risks. For example, it could help users be more conscious of attempts to influence them, or help digital platforms enforce platform policies against, for example, influence campaigns.

Article 50 (Art. 50) of the EU AI Act requires companies to label certain AI activity and is expected to be in force from 2 August 2026.[2] However, Art. 50 is ambiguous about whether actions from AI agents are to be labeled. In this piece, I argue that:

- Art. 50 likely requires actions to be labeled.
- Labeling web requests and browser actions is feasible by associating them with metadata (see Figure 1).
- Art. 50 additionally requires that it be possible to verify both who created a label (e.g. a particular developer) and whether it has been tampered with.

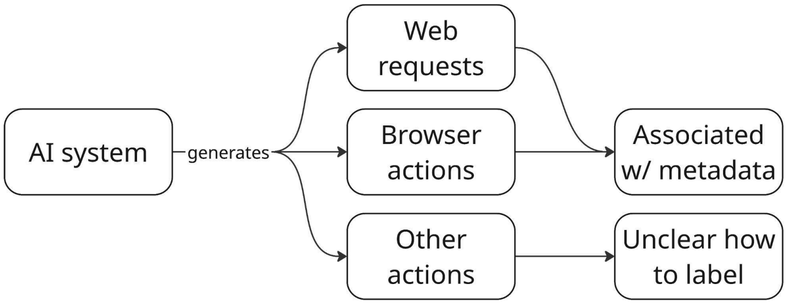


Figure 1 | AI systems can generate actions, such as web requests (e.g. retrieving weather data) and browser actions (e.g. cursor movements). These latter two types of actions can be associated with metadata.

Marking Actions

I focus on paragraph 2 of Art. 50 (Art. 50(2)), as it is the most relevant for companies that develop agents and/or serve them to users. It reads:

[2] The EU's Digital Omnibus proposes to extend the deadline to 2 February 2027 for AI systems placed on the market before 2 August 2026.

> *"Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated."*

In summary, AI systems that generate "synthetic audio, image, video or text content" should have their outputs marked (I use "marked" and "labeled" interchangeably). Many recent AI agents fall under this scope because they are composed of foundation models that at least generate text, and often more. Thus, their outputs must be marked. But do actions, such as making purchases or interacting with social media platforms, count as outputs?

Below, I argue that:

- Actions are outputs, which necessitates marking them.
- Marking actions could help achieve the goals of Art. 50.
- Marking certain actions is feasible through metadata.

Actions Are Outputs

There are three reasons why actions should likely be considered outputs under Art. 50(2).

Actions are Examples of Outputs

Art. 50(2) does not explicitly define "outputs", but the definition of "AI system" in Art. 3(1) lists "outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments".

Actions can be interpreted as "decisions that can influence physical or virtual environments". Indeed, digital actions from AI agents all involve some sort of impact on physical or virtual environments. For example:

- A purchase of an online good affects a user's bank account and leads to the delivery of the good.
- A post on social media can influence users' beliefs and behaviour.
- A request to a weather service causes a web server to return weather information.

The Examples of Outputs are Non-Exhaustive

The phrase "outputs **such as**... [emphasis added]" in Art. 3(1) implies that the examples of outputs are not exhaustive. To extrapolate additional examples, one could draw on the definition of an AI system, which states that it "infers, from the input it receives, how to generate outputs".

An output could then be, that which an AI system infers how to generate from the input it receives. Under this definition, actions are outputs. For example, Claude infers, from user instructions, how to carry out web searches to write a research report. ChatGPT agent infers, also from user instructions, what purchases to make.

As an aside, this argument implies that anything which an AI system generates is an output. This conclusion matches the intuitive meaning of “output”, but it may be neither useful nor feasible to mark certain outputs. For example, AI systems also generate (unnormalized) probabilities of what the next token should be, but marking such probabilities is likely not useful because most users never see them. Later, we argue that marking actions is both useful and feasible.

Many AI Actions Are Represented as Text

For AI systems that act in virtual environments, actions are often represented as text and synthetic text content must be marked under Art. 50(2).

Marking Actions Could Help Achieve the Goals of Art. 50

Marking actions can help to reduce risks from misinformation and manipulation, consistent with the goals of Art. 50 (see Recital 133).

When AI agents like, share, or comment on social media content, humans may mistakenly believe that this activity reflects genuine human opinion. Humans lack time to verify every interaction, and AI-run accounts are not always identifiable.

Such mistaken beliefs about what is popular or socially acceptable could affect human behaviour at scale. For example, the decision to trade a stock depends upon its perceived popularity, such as during the GameStop short squeeze.

Marking AI-generated actions could:

- **Help make users more aware and skeptical** of AI-facilitated attempts to influence their beliefs or behavior.
- **Help digital platforms tag or filter AI-generated activity** so as to:
  - Understand the impact of AI interactions. For example, measuring the spread and impact of AI-generated misinformation requires some way of identifying it.
  - Enforce existing platform policies.
- **Help improve accountability for AI misuse.** For example, providers could mark actions with a provider-specific identifier and a pseudonymous user tag. If a social media platform detects AI-generated activity that violates platform policies, it could report the identifier and

tag to the provider. With such information, the provider can prevent further violations from the user and AI system.

Marking Certain Actions through Metadata

Providers can feasibly add metadata to certain types of actions:

- **Web requests** (e.g. for retrieving a web page or interacting with a web application) are sent as structured text data. Marking these requests could be as simple as adding “AI_generated: True” to the data.

- **Browser actions,** such as cursor movements or mouse clicks, could be marked in a variety of ways, such as by adding an indicator that web pages can detect.

A key challenge is detection: recipients of actions need to know what metadata to look for. If each provider uses a custom method, detection becomes difficult. A potential solution is for providers to standardize how metadata is added and publicly document this standard.

For other types of actions, I suggest giving providers the flexibility to judge whether marking is feasible and whether other actors could reasonably be expected to make use of such marks.

Verifiability

Art. 50(2) also requires the marking of AI outputs to be “effective, interoperable, robust and reliable as far as this is technically feasible, taking into account the specificities and limitations of various types of content, the costs of implementation and the generally acknowledged state of the art, as may be reflected in relevant technical standards.”

Authenticity and Integrity

For marking to be effective, robust, and reliable, I argue that actors other than the provider need to be able to verify:

- **That a mark was created by a known provider (“authenticity”):** Without the ability to verify authenticity, an attacker could, for example, mark real human communications as AI-generated to undermine trust.

- **Whether a mark has been tampered with (“integrity”):** Without the ability to verify integrity, an attacker could, for example, modify a mark to claim that an AI-generated non-consensual intimate image (NCII) was generated by the victim.

What might such verification look like in practice? Consider an agent that uses a video conferencing platform with a human-looking digital avatar. When the agent calls a human, the platform should be able to verify the agent’s metadata and display the verification status to the human.

Luckily, digital signatures enable verification of metadata. Providers would sign metadata with their private key, and external parties could verify authenticity and integrity with the provider's public key. This approach, long used in digital certificates, is emerging for AI-generated content in general (e.g. C2PA) and web requests from agents (with some implementation support already).

Limitations

My proposal suffers from two main limitations: metadata can fail to reach relevant actors and can fail to inform even if received.

Metadata Can Fail to Reach Relevant Actors

It is extremely easy to remove metadata. For example, removing a single line from a web request (e.g. “AI_generated: True”) would preserve the action while removing the label. If users are allowed to modify actions before they are executed, they could intentionally remove labels to pass off AI-generated content as authentic.

Even if recipients of actions also receive metadata, they could fail to pass it on to other relevant actors. For example, social media platforms could receive metadata, but neglect to show it to users.

One might wonder if watermarking could overcome these problems, but it is not clear what it means to watermark actions in general. Challenges remain even for actions that can be represented as text:

- Text watermarking methods tend to be less reliable than image, audio, or video watermarking because text has fewer degrees of freedom.
- Text watermarking is less reliable for shorter texts. Requests to web services are often short.
- Web services expect requests in precise formats. Any watermark that changes this format would invalidate the action.

Furthermore, the authenticity and integrity of watermarks could be more difficult to verify. Watermarks hide invisible patterns in content, but these patterns can only carry small amounts of information. A digital signature requires more data than current watermarks can invisibly store. A potential workaround to explore could be having watermarks point to verification information stored on an external website.

Yet, even if metadata is easily removable, many actors could demand them as evidence of trust and reject interactions otherwise. For example, individuals (or virtual assistants on their smartphones) may not accept phone calls without a label from a known contact. Art. 50 could support the development and scaling of effective labeling technology so as to enable such demand. Analogously, Let’s Encrypt made it easier for websites to obtain SSL certificates, which enables web browsers to make HTTPS connections the default.

Metadata Can Fail to Inform

Especially as AI-generated content becomes more ubiquitous, knowing simply that content is AI-generated will likely cease to be a meaningful signal of trustworthiness. Additional signals could come from knowing which or whose AI system generated the content. For example, I am more likely to trust a video call with an agent if I can verify that a known contact controls it, even if the video is AI-generated.

Yet, Art. 50 does not require these additional signals. The Code could do so in principle, but companies would be less likely to sign on. Future efforts beyond Article 50 may be needed.

As a final challenge, users could ignore or misinterpret metadata. Often, users bypass web browser security warnings. Furthermore, although metadata can provide content provenance, provenance is not the same as credibility or accuracy. For example, users could mistrust all AI-generated content, even if such content is accurate and comes from a trusted actor.

Open Questions

A number of open questions remain regarding both how to implement labels for AI actions and how to manage misinformation and manipulation risks beyond Art. 50's requirements:

- How would the requirement to mark actions be enforced? A potential option is for a government body to test the AI system to verify that actions are marked appropriately.
- Typically, trusted actors known as "certificate authorities" are needed to attest to the authenticity of digital signatures. Who should be the certificate authorities for digital signatures on marked AI outputs?
- What could additional content provenance requirements look like? For example, in what contexts, if any, should the identity of certain AI users (e.g. commercial users) be attached to AI outputs?

Conclusion

I argued that Art. 50 of the EU AI Act likely requires:

- Web requests (e.g. online payments) and browser actions from AI agents to be labeled.
- That it be possible to verify both who created the label (e.g. a particular developer) and whether it has been tampered with.

I further argued that implementing these requirements is feasible.

A number of key questions remain regarding how to implement such a labeling requirement, such as how to enforce it and who the certificate authorities should be.

By itself, Art. 50 will not completely address impersonation, fraud, or deception risks. Further content provenance requirements are a promising area of investigation.

About the Authors

Alan Chan ✉ 𝕏 in

Research Fellow, GovAI

Alan's research focuses on governing AI agents. He is also interested in technical AI governance and societal resilience more broadly. He obtained his PhD from Mila (Quebec AI Institute).

Acknowledgements

I am grateful to the following people for helpful feedback on previous drafts of this piece: Jonas Freund, Markus Anderljung, Ben Garfinkel, and Tobin South. All mistakes remain my own.

About GovAI

GovAI is a 501c(3) non-profit organisation. Our mission is to help decision-makers navigate the transition to a world with advanced AI, by producing rigorous research and fostering talent. Researchers at GovAI work on a wide range of topics, with a particular emphasis on the security implications of frontier AI.

Appendix

I suggest potential Code text that incorporates the considerations from this work:

This Commitment applies to the following outputs from Signatories' AI systems:

1) Video, audio, and images;
2) Text, including outgoing requests to web services (e.g. HTTP requests);
3) Any actions taken in a web browser; and
4) Other types of outputs as appropriate.

Signatories will ensure that the outputs above are: 1) watermarked, whenever feasible and cost-effective watermarking methods exist for such outputs; and 2) associated with metadata, whenever feasible. Such watermarks and metadata will:

1) Indicate that the outputs are artificially generated or manipulated; and
2) Follow industry best practice, where it exists.

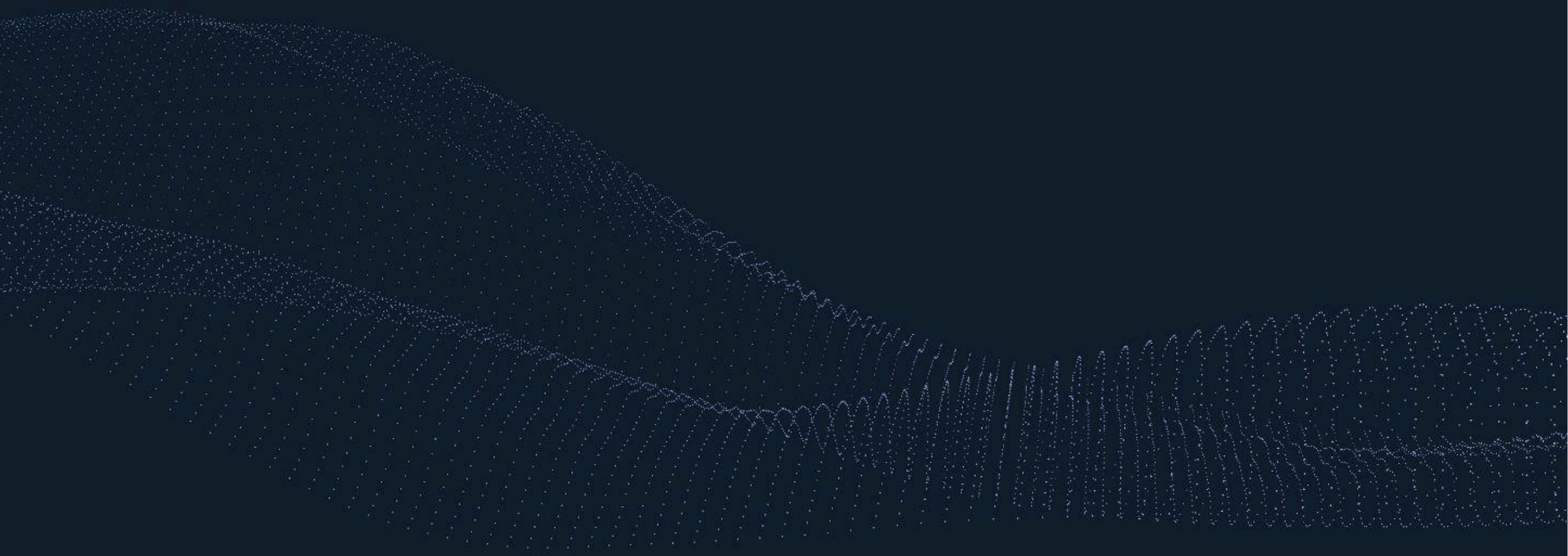
Signatories will ensure that it is possible for any actors that receive the metadata above to verify:

1) That the metadata originates from a particular Signatory; and
2) Whether the metadata has been tampered with.

Whenever Signatories enable their AI systems to send outputs to external platforms, tools, or services, Signatories should ensure that metadata are sent along with such outputs.

Signatories are encouraged to cooperate to standardise implementation of such watermarks and metadata so as to facilitate their detection.

This commitment does not apply if the AI system is authorised by law to detect, prevent, investigate, or prosecute criminal offences, subject to appropriate safeguards for the rights and freedoms of third parties, excluding when the general-purpose AI system is available for the public to report a criminal offence.



GovAI