ASPEN POLICY ACADEMY
aspen institute

POLICY BRIEF

How NIST Can Reduce Information Integrity Risks from Synthetic Text with Community Guidance on Provenance and Fuzzy Provenance

Marilyn Zhang

EXECUTIVE SUMMARY

Synthetic text generated by artificial intelligence (AI) poses significant risks to information integrity; when users trust false content generated by synthetic means (especially for non-creative purposes), they may experience real harms, such as negative health outcomes. To combat these risks, NIST should develop community guidance that encourages platforms hosting text-based digital content to make accessible (in no more than 1 click) the provenance and "fuzzy provenance" of the piece of text, when available.

Provenance refers to information that enables a user to definitively determine whether the text was human or AI generated; because AI text is prone to falsehoods, knowing that text is AI generated can encourage users to treat it cautiously.

Fuzzy provenance refers to exact text matches on the internet; this gives users further information to help them make their own judgment about the trustworthiness of text. To expand the usefulness of fuzzy provenance, NIST should recommend that generative AI companies make their free models' records available to be crawled and indexed by search engines, so that text matches with generative AI model records can also be surfaced.

Together, this community guidance would set forth a reliable and effective framework to give users more information enabling them to determine whether a piece of text is trustworthy.

To learn more about this project, please visit aspenpolicyacademy.org

aspen institute

BACKGROUND

Synthetic Text and Information Integrity

"Synthetic" text, or text generated by AI models online, has been proliferating. As the NIST AI-600-1 report on the risks of synthetic content notes, this means that there is a lowered barrier of entry to generated text that may not distinguish fact from opinion or fiction or acknowledge uncertainties, or could be leveraged for large scale dis- and mis-information campaigns. For instance, generative AI has been found to be a disinformation amplifier because it has a tendency to produce false content.

Information integrity risks caused by synthetic text (especially generated for non-creative purposes) can cause real harm to users and society at large. For instance, people who encounter misleading or false health-related claims on social media are more likely to have negative health outcomes as a result. These challenges are widespread; a recent study found that at least 4 in 10 Americans have encountered harmful claims related to COVID-19, reproductive health, and gun violence.

Provenance Methods for Reducing Information Integrity Risks

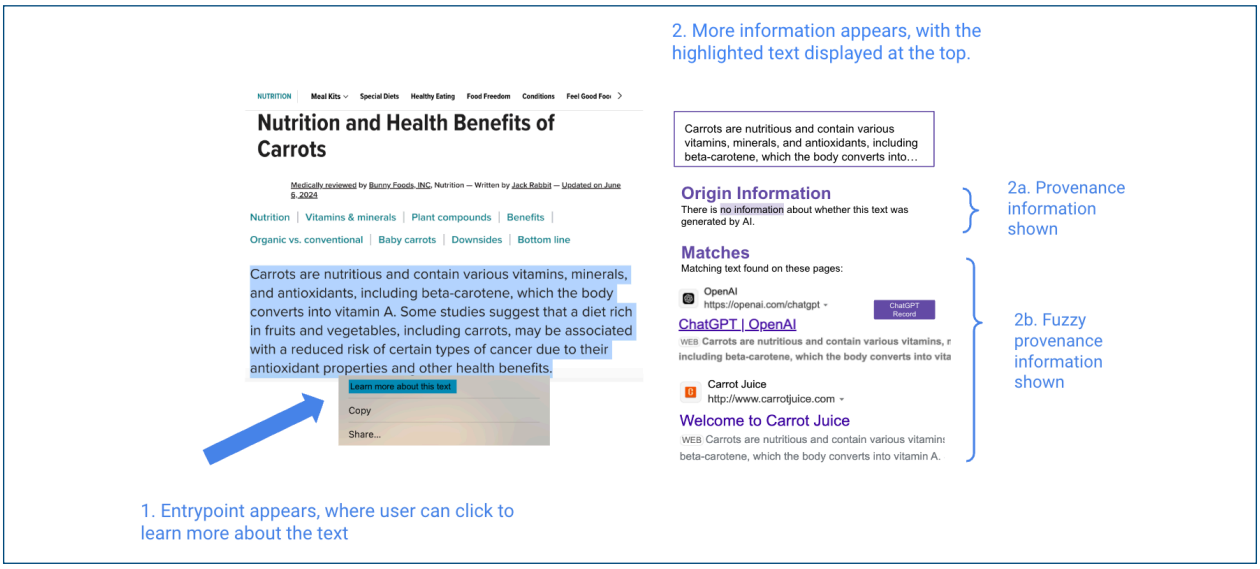
NIST has an opportunity to provide community guidance to reduce the information integrity risks posed by synthetic content. The main solution NIST is currently considering for reducing the risks of synthetic content in general is tracking provenance, which refers to whether a piece of content was generated by AI or a human. As NIST AI 100-4 describes, provenance is often ascertained by creating a non-fungible watermark, or a cryptographic signature for a piece of content like an image; the watermark is permanently associated with the piece of content. Where available, provenance information is helpful because knowing whether text was AI generated (and more liable to having false content) can help a user know whether to rely on the statements it contains. For example, an AI generated news report may be less trustworthy than a human news report because the former is more prone to fabrications.

However, compared with images, videos, and audio media, provenance methods do not work well for reducing the information integrity risks for synthetic text in particular. As NIST AI 100-4 Sec. 3.1.3.2 details, "All provenance data tracking techniques discussed in this report when applied to *text* have limitations and can be vulnerable to tampering" (emphasis added). Users interact with text differently than they do with other types of media like images, video, or audio: even if a piece of text is originally AI generated with a watermark (i.e., by generating words following a specific distribution), people can easily copy a piece of text by paraphrasing, without transferring the original watermark. In contrast, when an image with a watermark is screenshotted or cropped to make a copy, the watermark is transferred with the pixels. This difference also makes text watermarks more vulnerable to adversarial attacks, where bad actors can spoof text that is detected as having a watermark, when no genuine watermark was actually present.

RECOMMENDATION

To capture the benefits of provenance, while avoiding some of its weaknesses (like scalability), NIST's community guidance should recommend that platforms make available to users both provenance and "fuzzy provenance," or exact text matches on the internet. NIST also should recommend that generative AI companies make their free models' records available to be crawled and indexed by search engines, so that fuzzy provenance information would show text matches with generative AI model records. Only model-generated outputs (not user inputs) should be made available, and only after personally identifiable information (PII) is stripped. Making both provenance and fuzzy provenance information available (in no more than 1 click) would give users more information to enable them to determine whether a piece of text is trustworthy and reduce information integrity risks.

Combined Provenance and Fuzzy Provenance Approach



The above image captures what an implementation of the combined provenance and fuzzy provenance guidance might include. When a user highlights a piece of text that is sufficiently long, they can click "Learn more about this text" to find more information.

Provenance information would be shown under the heading "Origin Information" and would include whether there is conclusive metadata (like watermarks) on whether the text was generated by AI, according to C2PA standards, where available.

Fuzzy provenance information would be shown under the heading "Matches" and would include other websites that have an exact text match to the highlighted text, similar to the results that come up when

using a search engine. Generative AI companies that follow NIST's guidance would have their free models' records available to be crawled and indexed by search engines. This would enable these records to appear in the results if they contain an exact text match. These records would be clearly labeled (e.g., "ChatGPT Record") and ranked at the top.

Benefits of the Combined Approach

Showing both provenance and fuzzy provenance information provides users with critical context to evaluate the trustworthiness of a piece of text. Between provenance and fuzzy provenance, users would have access to information about most pieces of high impact text — especially claims that could be particularly harmful for individuals, groups, or society at large, such as medical information. Making all this information immediately available where users encounter text also reduces friction for them.

Provenance information can be very helpful to users. For instance, knowing that a construction site's description was AI generated may encourage users to check other sources (like reviews) to see if the company is a real entity (and AI was used just to generate the description), or a fake entity entirely, before giving a deposit to hire the company (User Journey 1).

Where clear provenance information is not available, fuzzy provenance information can provide helpful context to help users fill that gap. First, sometimes the lack of an exact text match on the internet can help ascertain that the content is original (although it can be either AI generated or human generated), which is relevant to the trustworthiness of certain documents like reports (User Journey 2). Second, when presented with a misleading claim like "ginger is 10,000 times more effective than chemotherapy," seeing that the text matches are from fact check sites and unreliable sources such as other social media posts can encourage users to further investigate that claim (User Journey 3). Third, absent any text matches from reliable sources, knowing that a claim (for instance, about carrots and cancer) matched an AI model's record may prompt a user to realize that the text might be false content generated by AI (User Journey 4).

Here is a summary of how provenance and fuzzy provenance can provide useful information to evaluate the trustworthiness of a piece of text:

		Fuzzy Provenance Information			
		No Matches	**Only AI Record Matches**	**Only Non-AI Record Matches**	**Matches with Both AI Record and Other Sites**
Provenance Information	**Not AI Generated**	Likely human generated original content	AI likely plagiarized originally human generated content not found online	Likely human generated; investigate trustworthiness of match sources	AI may have plagiarized from human or human copied AI generated text; investigate further
	AI Generated	Likely AI generated original content; be extra careful of false content	AI generated, potentially multiple times; be extra careful of false content	Likely AI may have plagiarized from human or human copied AI generated text; investigate further	AI may have plagiarized from human or human copied AI generated text; investigate further
	Unknown	Likely human or AI generated original content	AI may have plagiarized from human or human copied AI generated text; investigate further	Investigate trustworthiness of match sources	AI may have plagiarized from human or human copied AI generated text; investigate further

Fuzzy provenance is also effective because it shows context and gives users autonomy to decide how to interpret that context. Academic studies have found that users tend to be more receptive when presented with further information they can use for their own critical thinking than they are when shown a conclusion directly (like a label), which can even backfire or be misinterpreted. That is why users may trust contextual methods like crowdsourced information even more than provenance labels.

Finally, fuzzy provenance methods are generally feasible at scale, since they rely only on exact text matching with other sources on the internet. This also makes fuzzy provenance methods work without needing coordination among text producers or compliance from bad actors.

BUDGET

Since NIST is nonregulatory, no enforcement budget is needed. Rather, this guidance would serve as a stepping stone framework to make it easier (based on this proof of concept and initial learnings) for an agency like the Federal Communications Commission or Federal Trade Commission to enforce in the future with an expanded mandate (and budget).

APPENDIX

For more information on this project, please see the below documents:

1. User journeys explainer
2. Sample operational plan with specifications and mocks

ABOUT THE POLICY ACADEMY

The Aspen Policy Academy offers innovative training programs to equip leaders across sectors – from tech to climate, science to social impact – with the practical policy skills to craft solutions for society's most pressing challenges.

ASPEN POLICY ACADEMY
aspen institute

aspen institute