ASPEN POLICY ACADEMY
aspen institute

How NIST Can Reduce Information Integrity Risks from Synthetic Text

Marilyn Zhang

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

aspen Institute

EXECUTIVE SUMMARY

The proliferation of text generated by artificial intelligence (AI), also known as "synthetic" text, can harm information integrity. This project advises that the National Institute of Standards and Technology (NIST) should develop community guidance encouraging platforms hosting text-based content to make it clearer when text is known to be generated by AI ("provenance"), and to allow users to see where else the same text appears if its origins are unknown ("fuzzy provenance"). If users can more easily figure out whether humans or AI wrote the text, they will better be able to determine whether it should be trusted.

PROBLEM

Information integrity risks caused by synthetic text are widespread. NIST reports that generated text enables dis- and mis-information campaigns by mixing facts with opinions and glossing over uncertainties. Falling for this false content online can lead to real world harm. For example, the World Health Organization has found that people who encounter false health-related claims on social media may experience negative health outcomes. In the United States, a recent survey found that at least 4 in 10 Americans have encountered false claims related to major public health topics, including COVID-19, reproductive health, and gun violence.

The increase in AI-generated content makes it all the more important to mitigate information integrity risks now.

SOLUTION

This project argues that NIST's community guidance should recommend making both provenance and fuzzy provenance available across text-based sites. Provenance refers to information describing a text's source that allows users to definitively determine whether the text was AI generated, while fuzzy provenance refers to exact text matches on the internet that help users better guess whether text is synthetic. To access this provenance information, users could highlight a piece of text and click "Learn more about this text" to find the text's origin and view text matches. Having provenance more clearly accessible will enable users to better decide what text to trust.

Moreover, this project advises NIST to recommend that generative AI companies allow search engines to crawl and index their provenance information. This data sharing would increase the usefulness of fuzzy provenance text matches, helping grow a reliable framework for determining whether a piece of text is trustworthy.

For more information about this proposal, see: (1) a policy memo *explaining the benefits of providing provenance and fuzzy provenance;* (2) an operational plan *describing how digital content hosts can make both forms of provenance available;* (3) a user journeys *video modeling how to access provenance on a website.*

ASPEN POLICY ACADEMY
aspen institute

ABOUT THE POLICY ACADEMY

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

aspen institute