UTC Proposal: Watermark Symbols for AI Training Consent and Text Provenance

16 September, 2025

Author:

Stephen Casper
scasper@mit.edu
PhD Candidate
Massachusetts Institute of Technology, USA
https://stephencasper.com/

Sponsors:

Rishi Bommasani
nlprishi@stanford.edu
Senior Research Scholar
Stanford University, USA
https://rishibommasani.github.io/

Anka Reuel
PhD Candidate
Stanford University, USA
www.ankareuel.com

Jessica Dai
jessicadai@berkeley.edu
PhD student
University of California, Berkeley, USA
https://www.jessicad.ai/

Shayne Longpre
slongpre@media.mit.edu
PhD student
MIT, USA
shaynelongpre.com

Luke Bailey
ljbailey@stanford.edu
PhD Student
Stanford University, USA
https://lukebailey181.github.io/

Kayo Yin
kayoyin@berkeley.edu
PhD student
University of California, Berkeley, USA
kayoyin.github.io

Proposal Summary

We propose two new characters for AI training consent and AI content provenance. Both would take the form of a zero-width non-breaking space. As such, they would be undetectable in rendered text. They would be identical in form but different in interpretation from U+2060.

1. **Training Non-Consent Indicator**
    - **Abbreviation:** TNCI
    - **Form:** A zero-width non-breaking space
    - **Intended usage:** For document authors to explicitly express non-consent to AI systems being trained on their text. The suggested default usage is for authors to insert one indicator per sentence in a random location, but individual authors could use any insertion rule according to their preferences.

2. **AI-Generated Text Indicator**
    - **Abbreviation:** AGTI
    - **Form:** A zero-width non-breaking space
    - **Intended usage:** For AI deployers to indicate that a piece of text was generated by an AI system. The suggested default usage is for AI model deployers to insert one indicator per sentence in a random location, but individual deployers could use any insertion rule according to their preferences.

Additional Details

Block: We recommend inclusion in the *General Punctuation* block of Unicode, but we have no strong preferences.

Font resource and font embedding are not applicable for this submission: Both of the proposed characters are identical in form to U+2060. This would make them both undetectable in rendered text. For example, every other letter in this sentence is U+2060. This can be verified at https://www.soscisurvey.de/tools/view-chars.php.

Motivation

Summary of motivation: AI is posing new challenges in the usage, tracing, and study of digital media. As AI researchers, we often study challenges with data provenance, consent, and ecosystem monitoring. Two recurring themes are the ongoing crisis of consent involving AI training data sourcing (Longpre et al., 2024) and the challenge of studying the (mis)uses of AI-generated media in the digital sphere (Reuel et al., 2024; Bengio et al., 2025). In response to these challenges, we propose two new Unicode characters. The first would offer authors a mechanism to express non-consent to AI training on their text. The second would offer AI system deployers a mechanism to indicate that text is AI-generated.

TNCI – Offering a unique tool to express non-consent for training on text: Currently, state-of-the-art AI text processing systems are trained on extremely large amounts of Internet text (Bengio et al., 2025). This text is often sourced relatively indiscriminately and without obtaining the consent of the original author, giving rise to a "crisis of consent" in the sourcing of data (Longpre et al., 2024). For example, there are currently over 40 lawsuits across the United States relating to AI and copyright. Our proposed character would offer authors an optional tool for indicating nonconsent to AI systems training on their text. Currently, there are some existing conventions for expressing author preferences about their content, such as in `robots.txt` files for web crawlers. However, unlike other solutions, the proposed TNCI character would allow for non-consent to be encoded at the text level, allowing it to travel with the text when copied. Finally, we note that the symbol is intended as a tool to aid in digital consent, but it can be inserted into text by non-authors and will not always be used. Its presence, or lack thereof, therefore cannot be treated as a *certain* sign that consent has or has not been given.

AGTI – Offering a unique tool for identifying AI-generated text and studying its uses in the wild. Today, AI-generated text is appearing all around us. There are many instances in which it is crucial to determine if text is generated by a human or an AI system, including in education, law, and the study of AI and society. For example, the AI content generation market was recently estimated to be valued at over $2 billion. It remains a persistent challenge to reliably detect AI-generated text in the wild (Fraser et al., 2025). The proposed AGTI character would offer AI system deployers an optional tool to indicate that text from their system is AI-generated. Unlike other mechanisms for doing so, such as header or footer text, this character would be encoded at the text level, allowing it to travel with the text when copied. Finally, we note that the symbol is intended as a tool to aid in text provenance, but it can be inserted into non-AI-generated text and will not always be used. Its presence, or lack thereof, therefore cannot be treated as a *certain* sign that consent has or has not been given.

The insufficiency of U+2060: Both of the new proposed characters are identical in form, but different in interpretation from U+2060. It is necessary to have two additional characters in this case because their utility depends on both form and definition.