[redacted]

MP

[redacted]

[redacted]

[redacted]

Michael Prorock <mprorock@mesur.io>
mesur.io
2024-08-01

Position Paper for AI-CONTROL Workshop: Addressing the Limitations of Robots.txt in Controlling AI Crawlers

1. Introduction

The emergence of Generative AI and the surrounding ecosystem has introduced new challenges for the internet, highlighting the limitations of the Robots Exclusion Protocol (RFC 9309). The current mechanisms for controlling automated access are inadequate for both AI system operators and content creators. This paper explores the deficiencies of the robots.txt approach and proposes considerations for a more robust solution.

2. Current Limitations of Robots.txt

2.1 Global Nature and Site-Wide Scope

Robots.txt operates on a global scale, applying rules across an entire site. While this broad application is useful, there is a growing need for more granular control. Content-specific directives and overrides, particularly concerning provenance and authenticity, are necessary to meet the nuanced requirements of both AI systems and content creators. Incorporating approaches like C2PA (Coalition for Content Provenance and Authenticity) or from other areas related to well described JOSE/COSE/LAMPS patterns for provenance could address these needs.

2.2 User Agent Specification

Robots.txt relies on user agents to identify themselves, which is increasingly insufficient. Distinguishing between classes of activity, such as search indexing versus AI data ingestion, is critical. A more sophisticated mechanism is required to specify appropriate use, enable a publisher to identify corresponding licenses, and provide direct limitations on hosted data (e.g. X path is available for training, Y path for agent traversal or interaction, Z path for crawl and search indexing).

2.3 Licensing and Contact Information

The current protocol does not facilitate the identification of licensing terms or the contact information for content creators. This gap hampers the ability to manage and negotiate content use effectively. Providing the ability for a publisher to specify a path to an automated registration flow allowing content authentication and licensing details would enhance transparency and control.

2.4 Lack of Digital Signing and Provenance, Content Origin Specification

Robots.txt lacks provisions for digital signing and authenticity verification, which are crucial for ensuring the integrity and provenance of content. The inclusion of these features would help verify the source and authenticity of content, reducing misuse and unauthorized access. Additionally, as we look at issues like model collapse, misinformation propagation, and other items that are directly impacted by AI generated content, identifying if data is

that are directly impacted by AI generated content, identifying if data is human or model created (and if so, which model) will become increasingly important.

2.5 Rate Control for Agents and other AI driven actors

The protocol does not support crawler and agent specific rate control mechanisms, instead relying on HTTP methods which would for traditional usage patterns, but not for emerging ones. Additional specification in this area is essential for managing the load on servers from automated crawlers.

2.6 Insufficient Definition of Crawler

The definition of a crawler in robots.txt is too narrow, failing to encompass agents using headless browsers and other current and common techniques. A broader and more precise definition is necessary to cover the full spectrum of automated agents and AI specific crawlers accessing content.

3. Proposed Enhancements

3.1 Granular Control and Provenance

Introduce content-specific controls and provenance features to allow more detailed management of automated access.

3.2 Activity Class Distinction

Develop mechanisms to distinguish between different classes of automated activities. This distinction will enable more tailored access rules and better alignment with content creators' intentions.

3.3 Automated Registration and Licensing

Define an automated registration process that includes digital signing, authenticity verification, and licensing terms. This process will provide clear guidelines for content use and facilitate compliance with legal and ethical standards.

3.4 Rate Control and Content Origin Specification

Incorporate rate control features and the ability to specify content origination and the nature of authorship (human, AI, AI-assisted) within the protocol. These additions will help manage server load and ensure appropriate use of different content types.

3.5 Comprehensive Crawler Definition

Expand the definition of a crawler to include all forms of automated agents, including those using headless browsers. This broader definition will ensure comprehensive coverage and better regulation of automated access.

4. Conclusion

The current robots.txt protocol is inadequate for the evolving needs of AI systems and content creators. Enhancements in granularity, provenance, activity distinction, licensing, rate control, and crawler definition are necessary to develop a robust mechanism for controlling AI crawlers. By addressing these limitations, we can create a more effective and fair system for managing automated access to online content.

This position paper aims to contribute to the ongoing discussion at the AI-CONTROL Workshop, fostering the development of a comprehensive and adaptable framework for AI crawler regulation.