Google

Thought starters for public comment period

Alignment

Challenge	Opportunity & key question	Possible solution space
Alignment: A number of options have emerged, which give various directions on what content can be crawled and for what purposes (e.g. creating AI models, generating content, etc.). **Requiring additional instructions for each individual crawler entails additional implementation complexity for site owners.**	How could we as a community develop a **common control solution for web content** and standardize taxonomies for web crawling purposes?	The web and AI communities could drive **more alignment around the various new options for blocking crawlers** (i.e. new user agents being proposed or introduced by various companies and groups). The web and AI communities could create **a taxonomy of crawl purposes**, which crawlers would support via enhancements in robots.txt.

Transparency

Challenge	Opportunity & key question	Possible solution space
Transparency: There are **limited means for web publishers to have transparency into the ownership and purpose of crawlers** accessing their sites.	How could we as a community develop **methods to ensure web publishers benefit from visibility into the ownership and purpose of web crawlers** that access their sites? How can we **improve transparency through metadata or a registry**, enabling publishers to make more informed decisions about crawlers accessing their content?	Crawlers could **clearly & uniquely identify themselves**. Crawlers could provide **user-readable documentation** with details like: who is crawling (and why), how site owners can opt-out. Crawlers could provide **machine-readable, regularly refreshed lists of IP addresses or IP address ranges**. Crawlers could have a way to **automatically verify their authenticity**.

Granularity

Challenge	Opportunity & key question	Possible solution space
Granularity: Today's solutions offer a limited ability to control the usage of content published on the web. Site owners may want **more granular controls**; however, new solutions must continue to be **machine-readable in order to be usable at scale**.	How could we as a community develop **more refined controls for web content**, so that web publishers may control how their content is used? How can we agree on **reasonable options to offer publishers greater control** of the use of their content, while still allowing for innovation and respecting the rights of end users?	Site owners could **address groups of crawlers according to the primary product or service they support**, such as "search engines" and "generative AI applications." This would reduce complexity of updating robots instructions every time a new crawler is announced. The Robots Exclusion Protocol could accommodate **wildcard strings for user-agent addressing**. The community could evolve conventions around which strings indicate which types of crawlers.

Adoption

Challenge	Opportunity & key question	Possible solution space
Adoption: Not all crawler operators respect controls such as robots.txt, as **compliance is voluntary**. There is also no easy and publicly visible means of knowing identifying all available crawlers.	How could we as a community align with platforms and publishers on **mechanisms / disclosures to improve respect for protocols like robots.txt**, while also mitigating the challenges of detection and enforcement? How can we incentivize **industry-wide adoption of shared standards, best practices & behavior**? What tools can be developed help easily activate these controls at scale?	Organizations that run crawlers could self-certify via a **"Crawler code of conduct" or a common set of rules** for open web access. Adoption of publisher controls could be facilitated by a **web-based tool that helps website owners create and manage control files** that define behavior via business intent rather than by user-agents.

What's next

Stay informed & involved.	Engage with us in-person and/or virtually	Learn more about robots.txt.
Encourage others to sign up through our mailing list. goo.gle/pubcontrols	Stay tuned for discussion forums and opportunities to engage via in-person conferences, Github, and additional virtual sessions.	Check out "Robots.txt Introduction and Guide". goo.gle/robotstxt