**ROBOTS.TXT**

The robots.txt file is a set of rules for web crawlers, guiding them on which parts of a website they can or cannot access. It uses directives like User-agent to specify bots, Disallow to block URLs, and Allow to grant access to specific pages. The file can also include a Sitemap directive to help search engines discover content. While robots.txt prevents compliant bots from crawling restricted areas, it does not enforce strict security. Misconfigured rules can impact SEO by blocking important pages. Properly setting up robots.txt ensures efficient indexing while protecting server resources.

The robots.txt files of Amazon, ESPNcricinfo, and Instagram define specific rules for web crawlers, indicating which parts of their websites are accessible or restricted.

**Amazon** (amazon.com/robots.txt):

Amazon's robots.txt file includes directives that disallow various paths for all user agents, such as account access, login pages, and certain product-related actions. Notably, it explicitly disallows specific bots, including

* EtaoSpider
* GPTBot
* CCBot
* PerplexityBot from accessing any part of the site.

**ESPNcricinfo** (espncricinfo.com/robots.txt):

ESPNcricinfo's robots.txt file provides tailored instructions for different user agents. It disallows GPTBot and Google-Extended from accessing the site entirely, while allowing all other user agents to access all content. Additionally, it specifies disallowed paths related to print wrappers, error pages, and certain internal directories.

**Instagram** (instagram.com/robots.txt):

Instagram's robots.txt file is comprehensive, disallowing numerous paths for various user agents. It prohibits access to paths like /ajax/, /client\_error/, /direct/, and others for all user agents. Specific bots, including Amazonbot, Applebot-Extended, ClaudeBot, Google-Extended, GPTBot, and others, are entirely disallowed from accessing the site. The file also includes a notice stating that automated data collection is prohibited without express written permission from Instagram.

In summary, these robots.txt files reflect each company's approach to managing web crawler access, balancing search engine visibility with protecting sensitive areas and preventing unauthorized data scraping.