**Understanding Robots.txt and Its Implementation by Amazon, ESPNcricinfo, and Instagram**

The **robots.txt** file is a set of rules that guide web crawlers on which parts of a website they can or cannot access. It includes directives such as:

* **User-agent**: Specifies which bots the rules apply to.
* **Disallow**: Blocks access to specific URLs.
* **Allow**: Grants access to certain pages.
* **Sitemap**: Helps search engines discover website content.

While **robots.txt** can restrict compliant crawlers, it does not enforce strict security. Misconfigurations can impact **SEO** by unintentionally blocking important pages. Proper setup ensures efficient indexing while protecting server resources.

**Robots.txt Rules for Amazon, ESPNcricinfo, and Instagram**

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

Amazon’s robots.txt file includes multiple disallowed paths for all user agents, restricting access to areas such as **account pages, login portals, and certain product-related actions**. It also explicitly blocks several bots from accessing the site, including:

* **EtaoSpider**
* **GPTBot**
* **CCBot**
* **PerplexityBot**

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

ESPNcricinfo defines specific rules for different user agents. Notably, it:

* **Completely blocks GPTBot and Google-Extended** from crawling the site.
* **Allows all other bots** to access content.
* **Restricts access** to specific paths, such as print wrappers, error pages, and internal directories.

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

Instagram’s robots.txt file is extensive, restricting access to numerous paths for all user agents. It disallows areas like:

* **/ajax/**
* **/client\_error/**
* **/direct/**

Additionally, it **prohibits specific bots**, including:

* **Amazonbot**
* **Applebot-Extended**
* **ClaudeBot**
* **Google-Extended**
* **GPTBot**

Furthermore, Instagram explicitly states that **automated data collection is forbidden without written permission**.

**Conclusion**

Each of these companies structures their **robots.txt** file to balance **search engine visibility, user privacy, and protection against unauthorized data scraping**. While they allow indexing of public content, they also safeguard sensitive areas to maintain security and optimize server performance.