The robots.txt files of Amazon, ESPN crucify, 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 Etao Spider, GPT Bot, CC Bot, and Perplexity Bot, from accessing any part of the site.

**ESPN crucify (espncricinfo.com/robots.txt):**

ESPN crescendo’s robots.txt file provides tailored instructions for different user agents. It disallows GPT Bot 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. cite🚢turn0search1🚶

**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 Amazon bot, Apple bot-Extended, Claude Bot, Google-Extended, GPT Bot, 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 the need for search engine visibility with the protection of sensitive areas and prevention of unauthorized data scraping.