Ⅲ期（一般）

令和７年度

武蔵野大学大学院　データサイエンス研究科　データサイエンス専攻　　入学試験問題（３月９日）

［ 専門に関する筆記試験（英語試験を兼ねる） ］

問: 次の文が指摘している問題点を英文で簡潔にまとめ、この問題への対策を考察し英文で述べよ。

We often take the internet for granted. It’s an ocean of information at our fingertips—and it simply works. But this system relies on swarms of “crawlers”—bots that roam the web, visit millions of websites every day, and report what they see. This is how Google powers its search engines, how Amazon sets competitive prices, and how Kayak aggregates travel listings. Beyond the world of commerce, crawlers are essential for monitoring web security, enabling accessibility tools, and preserving historical archives. Academics, journalists, and civil societies also rely on them to conduct crucial investigative research.

Crawlers are endemic. Now representing half of all internet traffic, they will soon outpace human traffic. This unseen subway of the web ferries information from site to site, day and night. And as of late, they serve one more purpose: Companies such as OpenAI use web-crawled data to train their artificial intelligence systems, like ChatGPT.

Understandably, websites are now fighting back for fear that this invasive species—AI crawlers—will help displace them. But there’s a problem: This pushback is also threatening the transparency and open borders of the web, that allow non-AI applications to flourish. Unless we are thoughtful about how we fix this, the web will increasingly be fortified with logins, paywalls, and access tolls that inhibit not just AI but the biodiversity of real users and useful crawlers.

To grasp the problem, it’s important to understand how the web worked until recently, when crawlers and websites operated together in relative symbiosis. Crawlers were largely undisruptive and could even be beneficial, bringing people to websites from search engines like Google or Bing in exchange for their data. In turn, websites imposed few restrictions on crawlers, even helping them navigate their sites. Websites then and now use machine-readable files, called robots.txt files, to specify what content they wanted crawlers to leave alone. But there were few efforts to enforce these rules or identify crawlers that ignored them. The stakes seemed low, so sites didn’t invest in obstructing those crawlers.

But now the popularity of AI has thrown the crawler ecosystem into disarray.　As with an invasive species, crawlers for AI have an insatiable and undiscerning appetite for data, hoovering up Wikipedia articles, academic papers, and posts on Reddit, review websites, and blogs. All forms of data are on the menu—text, tables, images, audio, and video. And the AI systems that result can (but not always will) be used in ways that compete directly with their sources of data. News sites fear AI chatbots will lure away their readers; artists and designers fear that AI image generators will seduce their clients; and coding forums fear that AI code generators will supplant their contributors.

Shayne Longpre: “AI crawler wars threaten to make the web more closed for everyone”, available via WWW, https://www.technologyreview.com/2025/02/11/1111518/ai-crawler-wars-closed-web/. (2025)