ChatGPT’s Infrastructural Ambitions: AI, Commodification, and the Commons

Fenwick McKelvey & Robert Hunt, Concordia University

OpenAI’s ChatGPT, the hybrid private company–nonprofit’s latest project, arrives just as the Government of Canada attempts to pass its own response to AI, the Artificial Intelligence and Data Act (AIDA). Amidst ongoing debates over ChatGPT and its growing connections to major platforms, we urge greater consideration of the information commons as a key policy frame to understand AI chatbots and the large-language models used to train them. ChatGPT could not exist without the collective production of resources to support and maintain these commons. Its exploitation of those commons will only continue as OpenAI and its competitors try to monetize chatbots.

Commons-based approaches respond to demands for stronger collective rights in the AIDA Bill. Currently, AIDA focuses largely on economic or psychological harm to individuals with only a gesture towards larger systemic issues. Critics of the bill have questioned this narrow focus on harms in contrast with the Office of the Privacy Commissioner's recommendations for a rights-based approach.

We take ChatGPT’s recent rollout of third-party plug-ins as an occasion to elaborate how Canadian AI policy can be informed by theories of the information commons and to call attention to AI’s persistent reliance on precarious and low-wage platform work. As lawmakers develop policy to regulate AI, they should consider how AI firms have already taken advantage of existing commons and how they often resort to precarious labour to tackle key policy concerns like content moderation. The connection between these two issues is on clear display in ChatGPT’s recent launch of plug-ins.

ChatGPT’s infrastructural ambitions on display

On 23 March 2023, OpenAI announced the arrival of plug-ins for ChatGPT that connected the experimental AI live to the internet. These plug-ins allow the bot to “access up-to-date information, run computations, or use third-party services.” OpenAI president and cofounder Greg Brockman illustrated the new products’ utility by tweeting a video demonstrating how ChatGPT could find a recipe online, calculate the dish’s calorie count, and order the ingredients from Instacart. The demo shows how ChatGPT’s conversational interface could be used to do more than generate text. It also reveals that its developers aspire to make the technology *infrastructural*. The plug-ins make clear that ChatGPT’s owners want it to become a—possibly the—key platform for accessing the internet and accomplishing everyday tasks, including those that depend on precarious human labour.

Beyond showcasing the chatbot’s new abilities and aspirations, the video also demonstrates the hidden, infrastructural work behind ChatGPT—indeed, most modern AI—from the data used to train the model to the labour required to pick items off grocery shelves. The chatbot’s capacity to deliver impressively human-like responses to users’ queries relies on a group of large language models (LLMs), which are trained on massive datasets of text to “learn” to predict natural sequences of words. These datasets were built in a variety of ways, including scraping public websites (e.g., Wikipedia), digitized books, and social media networks.

In other words, millions of internet users' content was converted into data that trained the models that became the infrastructure for ChatGPT and similar applications. All this relatively indiscriminate data harvesting normally requires human judgment and labour to filter out racist, abusive, or otherwise offensive text, relying on a global system of ghost work. But cleaning such a massive dataset before training would be tremendously difficult, so OpenAI hired low-paid workers in Kenya to annotate problematic text that could be used to train ChatGPT what *not* to say.

ChatGPT's back story has taken on particular significance as the chatbot, initially launched as a free tool, enters a new phase of commodification. With billions of dollars invested in generative AI, new revenue streams will inevitably be pursued in the future. OpenAI's strategy to develop and maintain an app store might seem novel at first, but it has become foundational to most modern platform firms' business models. For $20 (USD) per month, subscribers can access ChatGPT Plus; similar or rival chatbots are being incorporated into other subscription-based software products, such as Microsoft's 365 family of applications (which could themselves be understood as infrastructural to much contemporary labour).

This coming wave of products and services raises pressing questions about LLM-based chatbots and their implications for aspects of internet governance, such as copyright, freedom of expression, and data privacy. From our view, these new products shine a bright light on drawbacks to the openness of information commons like internet content. Though nascent in its functionality, ChatGPT exemplifies how some AI applications trouble established understandings of online commons, prompting fundamental questions about what these commons are, who should have access to them, and how they can be maintained and governed. That billions of individual acts of creation were treated like a collective pool of non-proprietary data to manufacture subscription-based products shows how well-intentioned efforts to build and maintain commons can be preyed upon by corporations who see them not as communal resources but as vast pools of free labour.

These attempts to become critical infrastructure have been a persistent concern in media and information policy, prompting greater interest in policy approaches informed by the commons.

How does ChatGPT trouble the commons?

ChatGPT and other LLM-based chatbots simultaneously require and undermine theories of the political economy of communication. Building on the work of Vincent Mosco, we see two strategies at work:

1. Extrinsic commodification, where firms harvest and operationalize historical common data resources under aggressive interpretations of copyright law;
2. Intrinsic commodification, where firms mine and revalue data collected in their everyday operations.

Like past theories of enclosure and commodification, these efforts undermine the reproduction of information commons, turning public resources into private assets.

Generative AI, in most forms, relies on extrinsic commodification, such as the Common Crawl dataset being used to train OpenAI's models. The nonprofit 501(c)(3) organization relies on a broad interpretation of fair dealing and fair use to collect images and text published on the web in its entirety.

Ostensibly, Common Crawl is a product of a commons-based production model. At its launch in 2011, Lisa Green, director of the organization, announced: "it is crucial [in] our information-based society that Web crawl data be open and accessible to anyone who desires to utilize it." Though loosely premised on openness, by allowing powerful corporations to fulfill their desires, Common Crawl functions more as an engine of processes of commodification that encodes public resources into proprietary AI models.

Platforms—especially platforms that rely on data to optimize their operations, or those that Nick Srnicek refers to as "lean" platforms—have increasingly reconsidered their transactional or business data as sources of training data. The result is a form of intrinsic commodification that seeks to extract value out of ongoing company activities. This form of commodification applies to a number of platforms, from Meta and Google using their free services as sources of information models to Microsoft reconfiguring its Office Suite as a source of training data for its partner OpenAI. These developments are a critical matter for Canadian communication policy as well.

Artificial intelligence raises deeper questions about the information commons. Most directly, large AI firms' seamless commodification of public data calls into question whether adhering to a principle of openness successfully maintains information commons. Commons-based projects like Creative Commons content licenses or the General Public License for software are grounded in values of sharing, citing, and collective benefit. Treating these efforts as merely facilitating reservoirs of free data for powerful corporations negates their relational and reciprocal nature.

Artificial intelligence, then, might require reconsidering the commons as a relational norm premised on care and maintenance rather than unrestricted use. However, the uncopyrightable nature of current AI-produced works raises a secondary issue: the potential pollution of the commons by AI-generated works. Identifying AI-generated text is already a concern for OpenAI's owners, who are worried that training new models on the output of past models might cause chaos in the system. As generative AI trains on its own creations, meaningful signals become lost in the deluge of automated content production—how might human users suffer if the global information commons of the internet is swamped with machine-made and possibly plagiarized, misleading, inaccurate, or defamatory content? We ask policy makers to take seriously the exploitation of our collectively built information commons and to take care of the networked labours that enable it.