NUS Law Working Paper No 2023/021

I Prompt, Therefore I Am

Simon Chesterman

chesterman@nus.edu.sg

[August 2023]

I Prompt, Therefore I Am

When people think of risks associated with artificial intelligence (AI), Hollywood looms large. Movies have long conjured the worst case scenarios: from Hal refusing to open the pod bay doors in *2001*, to a murderous Arnold Schwarzenegger travelling back through time. If there is a robot apocalypse, however, it is unlikely to resemble a *Terminator* movie. A more probable scenario is what we see off-screen in – ironically enough – the Writers Guild of America (WGA) strike.

Hollywood's scriptwriters are protesting, in part, about the threat of jobs being replaced by ChatGPT or some other large language model (LLM). But is this just an effort to protect jobs from competition? Or does generative AI truly threaten the sustainability of the creative arts and the knowledge economy more generally?

I Think, Therefore I'm Paid

Peter Drucker coined the term "Knowledge workers" in 1959 to refer to non-routine problem solvers. People who "think for a living" earn through their ability to analyse and write – something that ChatGPT can replicate in almost no time and at almost no cost.

Journalists, already taking a beating as readers turn from traditional to social media, now face the prospect of technology taking over the writing task as well. Yet that same threat confronts anyone who analyses or writes for a living, such as lawyers and even – gasp – academics. Applications are not limited to prose, as ChatGPT has demonstrated proficiency in coding as well as poetry.

Further upstream, teachers worry that their students will use ChatGPT for their assignments. Breathless accounts have tracked its progress on standardised tests from the SATs to US Bar Exams. (Though it struggles, at least for now, with Singapore's PSLE.)

I'm not especially worried about the death of education. Students who want to cheat have always found ways to cheat, and educators assigning tests have never provided for outsourcing to ChatGPT or anyone else. But much as books and calculators reduced the need for memorising or multiplying, teachers need to help students take advantage of new tools while cultivating skills that computers cannot – yet – supplant.

Unless we install some guardrails however, there's a real danger that these tools will discourage creativity by removing rewards and distorting incentives.

Good Models Borrow, Great Models Steal

A key question for foundation models on which generative AI is trained is whether they go beyond fair use exceptions to copyright law. There is a difference, for example, between me reading J.K. Rowling's *Harry Potter* novels and being inspired to write my own story of, say, a teenage wizard in an exclusive boarding school battling dark forces, and photocopying all seven volumes and rearranging the pages in a pastiche of the original.

Data mining, which seeks to generate insights from data and is the subject of a recent exception in Singapore's copyright law, has long been seen as a productive area for AI research. Yet analysing text or images for the purpose of making recommendations or optimising workflows is quite distinct from using those text and images to generate more text and images.

The difference is not just the usage, where copying is central to the process, but also the economic impact of that usage. One of the reasons why I can't publish *Harry Potter* fan fiction is that it might dilute J.K. Rowling's economic return on her intellectual investment. Generative AI is, arguably, doing that to entire industries of creative writers and artists, by using their past work to devalue their future work.

This is not a hypothetical problem – we have seen the proliferation of new "content" generated by AI. The science fiction magazine *Clarkesworld* had to shut down submissions because it was being flooded with AI garbage[1].

A second, clearer concern is that much of the data used by LLMs for training is pirated in the first place. More than 70,000 pirated books were found when Peter Schoppert analysed the "Books3" dataset.[2]

No one is suggesting that generative AI should not be trained. But it is reasonable to expect that models are not trained on stolen data, and that a technology Goldman Sachs says might raise global GDP by 7%[3] pays *something* to the creators whose works serve as its fuel.

[1] https://techcrunch.com/2023/02/21/clarkesworld-ai-generated-submissions/
[2] https://aicopyright.substack.com/p/the-books-used-to-train-llms
[3] https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html

Faking It

So much for the production of generative AI. What about consumption?

A key concern of governments around the world is the rise of persuasive fake content at great scale. "Fake news" existed long before Donald Trump, but AI-generated videos of Ukrainian President Volodymyr Zelenskyy "surrendering" last year suggested how it might be operationalised as a weapon of war. Yuval Noah Harari went further to argue that it threatens democracy itself.

A common thread in AI discussions is that we should know when we are interacting with a machine or a person. Simple as it seems, AI-assisted decision-making increasingly blurs that line. Some chatbots, for example, start on automatic for basic queries, moving through suggested responses vetted by a human, before eventually escalating up to direct contact with a person for unusual or more complex interactions.

For raw text and images produced by AI, their provenance should be disclosed. Efforts are underway to detect AI-generated text through anti-plagiarism software. Another difficult but more effective approach would be to "watermark" text and images in a manner that is invisible to users but detectable using a key. However, the spread of deepfake porn points to the difficulty of policing any such rules – even when it is required by law.

Brave New World?

It is important to understand that – however human ChatGPT might appear – it does not "understand" the text it produces in any meaningful sense. LLMs draw on billions of past examples to make probabilistic guesses on which words go together. The technology underlying ChatGPT and other models does not suggest a path to anything like sentience or consciousness.

At present, tools like ChatGPT are limited to typed queries and responses, but it is a matter of time before spoken word inquiries will get answers from a humanoid host. Even as it stands now, it is easy to begin attributing qualities to ChatGPT after a few hours of interaction – taking the "I" that is engaging with you seriously. For example, I recently asked ChatGPT on a

whim, after several rounds about text ownership, whether it aspired to more than just "helping users".

"I do not have the same level of consciousness or independent thought as a human being," it replied. "And I do not have personal desires or goals."

"So, what level of consciousness *do* you have?" I asked.

It quickly backtracked.

"AI systems like me perform tasks that may appear to require some level of intelligence or consciousness," it replied. "But this is the result of complex algorithms and programming, and does not reflect the subjective experience of a human being. Is there anything else I can help you with?"

Simon Chesterman is Vice Provost at the National University of Singapore and Dean of NUS College, as well as Senior Director of Governance at AI Singapore. His first work of general fiction is the novel "Artifice". This article draws in part on work first published in the Straits Times as "AI-generated content is taking over the world. But who owns it?"