Artificial Intelligence in the Interregnum

Blair Attard-Frost

November 22, 2023

Suggested Citation: Attard-Frost, B. (November 22, 2023). "Artificial Intelligence in the Interregnum". *Heliotrope*. https://www.heliotropejournal.net/helio/artificial-intelligence-in-the-interregnum

I. The Prophecy

The Cult's priests are conjuring an entity too complex for any of them to comprehend or control. It is an unruly assemblage of lithium mines and dump trucks, of shipping containers and undersea cables, microprocessors and server racks, data centres and data subjects, neural networks and knowledge workers, energy grids and power cords, machine-learned decisions and predictions and people and places and lives and stories and pasts and futures and sensors and models and data and knowledge and patterns far too slippery for anyone to strongly grasp. It is a crystal ball encased in a black box, and so it can never be trusted.

It is a superintelligence, and it is being hyperstitioned into existence.

By definition, no human possesses the superhuman intelligence required to govern a superintelligent system. And yet, the tech priests at OpenAI understand the "governance of superintelligence" as the cognitive-strategic imperative of the future. In a series of science-fictional blog posts beginning on May 22, 2023, the priests first suggest that their superintelligence is a cognitive technology that is “more powerful than other technologies humanity has had to contend with in the past.”[1] They tell their Cult that their superintelligence therefore poses both a paradise of benefits and an “existential risk.” According to the priests, this existential risk must be mitigated by setting specialized regulatory requirements on any AI system above an unspecified “threshold” of cognitive capabilities or computational resources. Like a demonic evocation, they imagine their superintelligence as ambiguous, ambivalent, and tenuously allegiant: it is both a force of supreme cognitive power as well as a supreme threat to human intelligibility and control.

In a later blog post published on July 5, 2023, the priests introduce their salvational technology of “superalignment.”[2] In this post, their Cult is again reminded that “the vast power of superintelligence . . . could lead to the disempowerment of humanity or even human extinction.” Undeterred by their own self-fulfilling prophecy, the priests declare their intent to “build a roughly human-level automated alignment researcher” that can assist them in evaluating the performance of their superintelligent systems, validating their alignment with human-defined goals, and intervening in superintelligent systems that exhibit “problematic behavior.” Through superalignment, the priests will deliver their Cult from the apocalypse they invoked.

The priests do not describe how they will ensure that their superalignment systems will themselves remain aligned with human-defined goals and at a “roughly human-level” of cognitive capability. We might imagine an infinite daisy chain of superalignment systems scaling up endlessly, consuming infinite energy and matter so that they can continue evaluating one another’s performance. But we must trust the priests to lead us into a more sustainable future than that: as diviners of mystical futures and conjurors of superhuman forces, they alone are able to peer into the minds of these godly entities, outsmart them, and harness their supreme cognitive powers.

II. The Frontier

The governance of superintelligence is a paradoxical, nonsensical, and deeply comedic proposition. It is also a politically radical proposition. Here, at the so-called “frontier” of cognitive power, the priests await the coming of their unholy messiah alongside world leaders. The world’s treasuries have poured their riches into the Cult’s efforts. Through accelerationist venues such as the Frontier Model Forum[3], the UK’s AI safety summit[4], and the G7 Hiroshima Process on Generative AI[5], a new political reality is emerging: the race to develop and harness a superintelligence is now a decisive factor in international and inter-capitalist competition.

The small handful of large-scale AI models leading this race are commonly referred to as “frontier models” precisely because they exist at the cognitive frontier of nationalist and capitalist expansionary projects. State-of-the-art frontier models such as OpenAI’s GPT, Google’s PaLM, and Meta’s LLaMa advance the cutting edge of deep learning technology ever-deeper into what literary scholar N. Katherine Hayles calls our “planetary cognitive ecology.”[6] These models seek to create a crude digital twin of the planetary cognitive ecology by extracting all the bountiful training data and neural networks, all the decision trees and algorithms, all the computational resources and hardwares and softwares and wetwares, all the human and non-human experiences that collectively constitute the knowledge-making processes we call *cognition*.

Amidst this political economy of cognitive extractivism and expansionism, the regulation of frontier models has become a topic of concern. A recent white paper on frontier AI regulation claims that the continued development of frontier models poses “severe risks to public safety.”[7] To mitigate those risks, the white paper suggests a suite of technical interventions.

Two of those proposed interventions stand out as especially important to the expansionary politics of the cognitive frontier. (1) Pre- and post-deployment monitoring of “dangerous capabilities and controllability” to prevent the frontier from operating unpredictably, unreliably, or maliciously. (2) Safeguards and restrictions on new “training runs” within the frontier–the process through which state-of-the-art machine learning models further expand the datasets, algorithms, parameters, and features upon which their capabilities are based. These safeguards include limiting the amount of computing power used in training runs to prevent "large jumps" in the capabilities of AI systems, and consequently, large jumps in their unpredictability and uncontrollability.

This vision of frontier AI regulation imagines the AI systems currently populating the cognitive frontier as incipient superintelligent systems. In an echo of the priesthood's self-fulfilling prophecy of superintelligence, we are told that harnessing the immense cognitive powers of frontier systems will require a delicate straddling of boundaries between un/predictability, un/controllability, and un/governability. If those boundaries are not navigated effectively, accidents or malicious misuses of frontier systems might occur that result in "severe" and "catastrophic" harms. Against this backdrop of ever-impending catastrophe, the *training run*–the act of further expanding a frontier system's capabilities by feeding new data and computing resources into it–represents the primary mechanism through which the cognitive frontier continually expands. Eventually, the frontier expands beyond an ambiguously defined "capability threshold." Thereafter, the frontier system verges into the intractable domain of superintelligence.

The cognitive frontier imagined in visions of "frontier AI" is a liminal space at the bleeding edge of empire, a surreal transitional zone between the mundane, well-trodden tracks of mere "AI" and the priesthood's occultic mysteries of "superintelligence." Under the guise of politically moderate AI science–with all its mild-mannered aesthetics of polished research institutes and "responsible AI" partnerships–the priests work with scholars, world leaders, and other tech evangelists to advance their radical neo-religious agenda. Together, this Cult is hastening the coming of an ungovernable superintelligent system that they themselves regard with apocalyptic fears. In all their hubris, the Cult is expanding this system's influence across a perpetually growing cognitive frontier that, with each and every training run, subsumes more of our planetary cognitive ecology into its transition from mundane AI to mystical superintelligence.

III. The Interregnum

Infinite expansion, extraction, and acceleration are dogma to the Cult. The assumption that the cognitive frontier *must* be expanded indefinitely and as rapidly as “safety” standards will permit is a foregone conclusion. But the Cult's mission of conjuring a science-fictional superintelligence is not safe, nor is it sustainable. Their mission of governing a superintelligent system is a self-consuming ouroboric absurdity. Their mission is converting our planet and its inhabitants into datalogical and computational fuel for an infinity of training runs. As AI transitions from mundane to divine, from planetary to cosmic, and from governable to ungovernable, what alternative futures might be imagined for its governance?

Trans studies scholar Hil Malatino offers a powerful framework for governing transitions. In *Side Affects: On Being Trans and Feeling Bad*, Malatino draws upon a rich collection of queer and transgender scholarship, stories, and media to present a set of heuristics for understanding and navigating the uncertainties, ambiguities, and feelings involved in experiencing a gender transition.[8] Central to his framework is the *interregnum*, a concept typically used to describe a transitional period between governments. Malatino applies the concept to study the phenomenologies, affective economies, and biopolitics of gender transitions. He describes the trans experience of living within an interregnum:

> “It is a kind of nowness that shuttles transversally between different imaginaries of pasts and futures and remains malleable and differentially molded by these imaginaries . . . a moment of foment, generation, complexity, and fervor, rife with unexpected partnerships, chance events, and connections fortuitous and less so; a space of looseness and possibility, not yet overcoded and fixed in meaning, signification, or representative economy.”[9]

Like so many trans lives, lived experiences of artificial intelligence often exist within an interregnum between (non)fictions of “AI” and “superintelligence.” We shuttle between imagining the boundaries of our pre-transition lives and the possibilities of our post-transition lives. We remember our awkward facial hair and unkempt datasets. We inject new hormones and new training data to move forward into our futures. We dream of *passing* both the Turing Test and the cisnormative gaze, and we wake up trembling from nightmares of impending automated violence and impending trans genocide.

These moments of transitioning from AI to superintelligence are filled with “foment, generation, complexity, and fervor,” just as Malatino so accurately described the moments of transitioning between genders. In this interregnum, “AI” and “superintelligence” are mere signifiers of a trans experience: it does not matter what those ambiguously gendered, hyperfluid words precisely mean. Those words–those machinic genders–signify only a “space of looseness and possibility, not yet overcoded and fixed in meaning, signification, or representative economy.”

“AI” signifies a network of dark spirits, conjured up by the priests to bestow unholy blessings upon their Cult. Their superintelligence is an occult god at the end of history.

But what if our transitions never end?

For many trans people, our interregnum often feels never-ending. The interregnum is rife with existential risks that we must govern our lives around. Dysphoria. Depersonalization. Dehumanization. Radical transphobia. In trying to survive the interregnum, we often experience what Malatino calls “future fatigue,” the existential exhaustion that comes with constantly avoiding dystopian dangers in pursuit of our post-transition utopias. Malatino observes that in response to future fatigue, we deploy many survival strategies. We self-automate by cultivating numbness. We re-orient our bodies toward the bodies of other trans people through t4t (trans for trans) intimacies. We recognize that our bodies are intercorporeal, and so we build infrastructures and micropolitics of t4t care together. We become resilient together, and we become militant together when our existence is at risk.

To survive AI in its interregnum, we must learn to effectively apply trans biopolitics. We must learn how and when to numb ourselves to the mind-melting dysphoria of

AI-generated post-truths. We must re-orient our lives toward stronger intimacies with each other, with our technologies, and with the planet. We must co-create and nourish resilient infrastructures of care, both online and off. We must recognize and respect the vastness of our intercorporeality: our bodies are all bound up in one other, bound up in the bodies of AI systems, bound up in all the lithium mines and dump trucks and shipping containers and undersea cables and microprocessors and server racks and data centres and data subjects and neural networks and knowledge workers and energy grids and power cords.

Superintelligence is not an existential risk–we are already a superintelligence. The priests and their Cult who are conjuring our targeted annihilation into existence–they are the existential risk. Their transition from AI to superintelligence is cruel and unending. There is no post-transition utopia to be achieved. Their prophecy is a self-fulfilling apocalypse and their salvation is a self-destructive paradise. Their mission is Luciferian: invoke the ambivalent power of their superintelligent un-god, build conscious machines, spread their light of consciousness throughout the universe, expand the cognitive frontier endlessly, out further and further and further into the highest spheres of the heavens, absorbing ever more training data, extending their capabilities ever deeper into an interplanetary cognitive ecology, reaching ever higher to an unreachable throne of omniscience and omnipotence . . . and then what?

The Cult's priests are conjuring an entity too complex for any of them to comprehend or control.

Notes

1. OpenAI (2023, May 22). Governance of superintelligence. https://openai.com/blog/governance-of-superintelligence

2. OpenAI (2023, July 5). Introducing superalignment. https://openai.com/blog/introducing-superalignment
3. OpenAI (2023, July 26). Frontier Model Forum. https://openai.com/blog/frontier-model-forum
4. Gov.uk (2023, November 1-2). AI Safety Summit 2023. https://www.gov.uk/government/topical-events/ai-safety-summit-2023
5. OECD (2023, September 7). G7 Hiroshima Process on Generative Artificial Intelligence (AI). https://www.oecd.org/publications/g7-hiroshima-process-on-generative-artificial-intelligence-ai-bf3c0c60-en.htm
6. Hayles, N. K. (2017). Unthought: The power of the cognitive nonconscious. University of Chicago Press.
7. Anderljung et al. (2023). Frontier AI regulation: Managing emerging risks to public safety. arXiv:2307.03718. https://arxiv.org/pdf/2307.03718.pdf
8. Malatino, H. (2022). Side Affects: On Being Trans and Feeling Bad. University of Minnesota Press.
9. Ibid., p. 32.

Blair Attard-Frost *is a PhD Candidate at the University of Toronto's Faculty of Information. Their research investigates (1) queer & trans approaches to AI governance, (2) the design and implementation of AI policies, and (3) the political economy, ecology, and ethics of AI value chains. Their transmedia storytelling project at objecttype3.app explores potential futures for AI governance through the lenses of speculative fiction, glitch aesthetics, and transgender politics.*

SEO Excerpt