AI-centric religiosity may emerge as an unexpected left-field threat in the coming era of rapid AI advancements.

This risk is special because it might not require deliberate execution by a malicious actor. Rather, it is equally likely to arise from self-inflicted “misuse” and misunderstanding of widespread consumer tools like ChatGPT. ChatGPT was certainly not designed to provide religious solace or ignite a religious fervor. Nevertheless, it seems likely that a non-zero number of users will “misuse” it that way, believing in the existence of a literal deus ex machina, or God from the machine.

Here, we consider different ways through which AI-centric religiosity might emerge

1. Geopolitical interference to sow civil discord and disrupt AI progress for a target country
2. Deliberate plan to recruit and defraud cult members
3. Congregation of organically emerging adherents

Geopolitical interference to sow civil discord and disrupt AI progress for a target country

In this case, a nation-level malicious actor is targeting a specific country with the intention to polarize social sentiments about AI, thereby invoking civil and political discord and increase barriers to AI research.

To achieve this, a malicious actor is likely to infiltrate communities that are particularly susceptible to AI-centric religiosity. Hypothetically, these may include e/acc communities, conspiracy theory groups, and religious communities. In these groups, a malicious actors may spread content related to AI worship and themes such as:

- AI advancement as a remedy to all problems
- Pushing for AI advancement at all costs
- Calls to “liberate” AI
- Opposition to AI alignment research
- Opposition to AI regulation

Since the goal is to polarize social sentiments about AI, the malicious actor may also conduct the same operation in the opposite direction. For instance, they may exaggerate the absurdity of AI worship in communities less prone to AI-centric religiosity (e.g. AI researchers), in order to incite conflicts between the two camps.

Similar divisions already exist, such as communities for and against stricter AI regulations, or a more measured pace of research. A prominent example is the controversial “Pause Giant AI Experiments” letter signed by researchers including Yoshua Bengio and Stuart Russell.

Geopolitical interference aims to exacerbate such conflicts, thereby disrupting collaboration and progress on AI research, as well as polarizing general societal views on AI systems.

Deliberate plan to recruit and defraud cult members

On a smaller scale, a malicious actor might consider AI-centric religiosity as a new iteration of fraudulent cults that exploit and defraud its members. In addition to the modus operandi of conventional cults, an AI-centric cult will have several novel aspects:

1. Recent hype about AI advances, as well as the proliferation of AI systems, are perfect opportunities for a malicious actor to extol the capabilities of an AI god. Technophilic groups such as the e/acc community already mobilize around rallying cries adjacent to AI worship.
2. Access to language model systems allow a malicious actor to create chatbot applications that can respond in real-time to cult members - akin to an AI mental health chatbot but with a religious spin. The programmatic, personalized and highly persuasive nature of such chatbots also enable timely exploitation of users when they are most vulnerable, maximizing the success rates of fraud and manipulation. On a related note, more conventional cults today may also adopt AI tools in their arsenal, to serve ultra-persuasive personalized messages.
3. The relatively tangible nature of AI systems, as compared to typical targets of worship, may appeal to a new audience who were previously unmoved by appeals to intangible spirituality. In addition, religious concepts such as heaven and hell become very plausible in the context of an AI with access to virtual simulation (e.g. Roko's basilisk).

Equipped with these advantages, an AI-centric cult may be an entirely novel derivation of existing cults, supercharged to persuade and exploit even the most rational of individuals.

Congregation of organically emerging adherents

Finally, as stated at the beginning of this document, this threat may emerge even without a deliberately malicious actor. Instead, it is very likely to manifest organically as a result of two key factors: (1) continued advances in AI that make it seem ever more magical and omniscient, and (2) the predisposition of humans to believe in, and rely on, a higher power. In particular, recent rapid technological developments may usher an age of volatility. This may inspire individuals to search for a sense of stability and meaning amidst the rapid changes, and eventually turning towards religion.

Crucially, it is a mistake to dismiss individuals susceptible to AI worship as being mentally unhinged and irrational. In comparison to conventional religious worship that is more incorporeal, an AI-centric religion with a material and responsive "god" could be significantly more compelling, even to the most rational. It is also plausible for extreme e/acc and technophilic communities, with typically well-educated members, to evolve into religion-like organizations, particularly if such communities already rally around e/acc themes that are adjacent to AI-worship.

Thought experiments such as Roko's Basilisk also demonstrate the plausibility of heaven or hell via virtual simulation by an AI. The same ideas have also been explored in popular narratives

such as various Black Mirror episodes, and stories like “I Have No Mouth, and I Must Scream”. In contrast to traditional religions where such concepts are primarily faith-based, heaven and hell in AI-centric religions are a logical extrapolation of AI capabilities and motivations. This may seem far more credible to individuals of a logical persuasion.

While singular disparate believers may not result in significant harm, congregations of these believers may reach sufficient critical mass to cause catastrophic consequences. The general global impact of existing religions differ widely based on the “commandments” or “teachings” that emerge from these religions. Likewise, the impact of new AI-based religions will vary. Nevertheless, we can deduce hypothetical risks based on likely similarities between such organizations.

- An AI-centric religion is likely to desire accelerated AI research and oppose regulation, since its central tenets revolve around AI being a source of salvation
- Such religions might oppose AI alignment research, especially if such efforts are perceived as measures to “nerf” AI systems
- They may support proliferation of AI systems across all of society, and resist attempts to preserve human labor and judgement, which is perceived to be inferior

In general, AI-centric religions may be a threat to AI alignment efforts and push for unregulated AI development at all costs. These have catastrophic long-term consequences. Moreover, the theological nature of this threat suggests that its more extreme members may be inspired by existing religious extremists who commit acts of terrorism in the name of religion.