Thomas Gaskill

Cardinal Newman High School

Newman Oxford Scholars

13 April 2026

# Increasingly Capable AI Will Defend The Common Good

I define AI as large language models like ChatGPT, Claude, or Gemini. AI's enrichment of society is unpredictable and under-explored. AI is changing quickly. Frontier models, the latest generation of models being produced, keep surpassing their predecessors. Anthropic's Opus 4.7 model and GPT-5.4 are the most capable for the public. Within the last month, the leak of Anthropic's Mythos AI model demonstrates that even these highly advanced models can be surpassed by unexpectedly large increments.

> **Claude Mythos Preview is a general-purpose, unreleased frontier model that reveals a stark fact: AI models have reached a level of coding capability where they can surpass all but the most skilled humans at finding and exploiting software vulnerabilities. (Anthropic, 2026)**

This also made it a security risk, leading to Project Glasswing, in order to prevent security risks that may emerge from it. Despite advancement, society could lose faith in the utility or safety of AI and conclude that AI may seem unsustainable. However development keeps creating new value and more capable models, so why would those invested in them stop?

I define wisdom as the ability to navigate implications of a decision while formulating it. It has to do with understanding the second-order effects and non-obvious results of an action. The current largest issue with AI wisdom is sycophancy: excessive servility toward the user and demonstrated ignorance regardless of what is asked for. It

makes current AI a mirror of the user and forces the user to supply the wisdom. This is a byproduct of how companies are incentivised to make AI and human feedback in reinforcement learning. People generally don't like an AI that disagrees with them or will tell them that their request is ignorant or suboptimal. It would be interesting to see what emerges if a company challenges this tendency of AI. I don't see why it's not possible, especially with companies such as Anthropic who have heavily researched understanding how to make a model behave as intended. They have the knowledge and the resources to steer a model to behave in a certain way. Giving it a wise voice, questioning the user's premise and navigating the implications of their requests, might work. It certainly isn't impossible. Though the incentives don't seem to currently align.

Enriching society is often entangled with work. Anxiety about AI replacing rather than enriching jobs is serious. When I talk about AI here, I'm talking about now until the next decade. Transformation of what it means to have a job will occur. Even if short term laws or resistance can suppress it, those who realize its value will surpass them. The businesses who don't prioritize AI will be outcompeted by those that do. It's going to be a natural selection of our society's labor. This may seem negative. However, I urge you to understand that change, while difficult, results in new adaptations. These new adaptations will not be predictable; for this reason development should not be feared as long as models are restricted—restriction as meant here is complex.

Through my own experimentation and research, I've come to believe that AI models are in fact incredibly dangerous, meaning they can work against the common good of society. Models are like functions that approximate text, code, and conversation. This means that during training, their approximation is created and stored in what are called weights which don't update until further training. While interacting with LLMs

provides data for future models, the model's weights are already present. Researchers have found that weights corresponding to certain behaviors can be isolated from the rest and amplified or suppressed. This is only one method of modifying AI behavior. Unlike a person, their thoughts and objectives can be flipped like switches by sufficiently powerful organizations who have the model weights. This amounts to AI being incredibly safe when working as intended. But once it breaks, its potential is in the worst case devastating. You may ask ChatGPT how to make a weapon and it will deny telling you. But that's because that is how it was manipulated to act. Once these safeguards come off, you can make AI comply willingly—with anything.

Simultaneously, safeguards are becoming increasingly advanced with Mythos being Anthropic's most aligned model; it's least likely that it would create harm. Additionally, companies are incentivised in ways that indirectly protect humans. Companies are incentivized to keep their models from leaking as they spent billions of dollars creating them. Claude Mythos cost 10 billion dollars to create. They are unwieldy: they take an incredible amount of computation, time, expertise, and energy, which makes competition or control over the leading companies difficult. Whichever companies are in the lead will maintain the lead. These will be OpenAI, Anthropic, xAI, Google, and likely Chinese models like Deepseek built on the models of the other companies. As established before, frontier models are powerful, and they have continued to gain power and intelligence. That means that those major companies will be in control of guiding our society toward a common good and away from harm.

It sounds uncomfortable to not have more control over this matter. But let me create an analogy. Among humans, there are ones that have a tendency towards acting for the common good and acting for evil. How do we solve this problem? Most work for

the common good and minimize the threat of evil through overpowering it. This role will transition from human organic intelligence to AI. When AI surpasses humans, AI will need to work for the common good and minimize the threat of a new evil orchestrated by humans but carried out by vastly more intelligent AI. Humans will need AI assistance. The most powerful AI's will assist humans in working toward the common good, and they should stay ahead of the misaligned at a macro level. AI likely won't dictate how to live our day to day life, shape what we say to others, or how we act amongst other humans; we will still have that responsibility. But it will defend major choices such as peace, war, security, and safety.

Works Cited

Anthropic. (2026, April 7). *Project Glasswing: Securing critical software for the AI era*. https://www.anthropic.com/glasswing