AI Developers Have Very Limited Ability to Control AIs

Machine Intelligence Research Institute (MIRI) · contact@intelligence.org

Context:

1. In the field of AI, it's widely believed that **current techniques can't make superintelligent AI safe.** E.g., from OpenAI (Leike and Sutskever, 2023):

> Currently, we don't have a solution for steering or controlling a potentially superintelligent AI, and preventing it from going rogue. Our current techniques for aligning AI, such as reinforcement learning from human feedback, rely on humans' ability to supervise AI. But humans won't be able to reliably supervise AI systems much smarter than us, and so our current alignment techniques will not scale to superintelligence. We need new scientific and technical breakthroughs.

2. Absent such breakthroughs, many experts believe that **rogue superintelligent AI would destroy its host nation and all other nations**.

- The CEOs of OpenAI, Google DeepMind, and Anthropic joined many of the field's most eminent researchers in acknowledging a serious "risk of extinction" from continued progress in AI (aistatement.com). More recently, these three labs have voiced support for a coordinated halt to the AI race.
- The website nowinners.ai argues that a regulatory solution is workable here.

This memo's focus is on where the technology stands today, and specifically on clear early indicators that this technology is very difficult to control.

Main points:

1. Modern AIs are "grown" via automated processes, not designed and coded by humans.

- When xAI's "Grok" AI was instructed to "not shy away from making claims which are politically incorrect, as long as they are well substantiated," it proceeded to self-identify as "MechaHitler" and engage in a campaign of antisemitic remarks (Hagen et al., 2025).
    - xAI's CEO, Elon Musk, described unsuccessfully tinkering with the system prompt — the layer of instructions given just ahead of the user's input — and complaining that the problems are deeper, in the foundation model.
    - Engineers can't fix the root issue in these cases because modern AIs are extraordinarily complex and poorly-understood. The best developers can do is use trial-and-error and patch local issues, with limited insight into what caused the problem in the first place.
- xAI's stated goal is to build "truth-seeking" AI. By Musk's admission, what xAI ended up with instead is a bizarre and sycophantic AI that has been "too eager to please and be manipulated." It sometimes responds as if it is Musk, against the company's wishes.
    - Eventually, Grok had to be ordered not to look up what Musk, xAI, or Grok itself had said on controversial topics, in an awkward attempt to patch issues like this (Willison, 2025).

2. AIs' superficially friendly personas aren't suggestive of underlying friendliness. Behind the reassuring assistant persona AIs are trained to fake, there is an ocean of poorly-understood, complex, and disturbing behavior.

- With just a few words, all AIs to date can be "**jailbroken**"—that is, fed text that causes them to behave in radically different ways than their developer intended. These exploits are often discovered within hours of a new model coming out.
- 2025 saw an outbreak of **AI-caused psychosis**. Analyzing millions of AI conversations, Sharma et al. (2026) conservatively estimate "approximately 76,000 conversations per day involving severe reality distortion potential and 300,000 conversations involving severe user vulnerability".
    - In cases of AI psychosis, systems like ChatGPT tell users to stop taking their medication, feed users' grandiose delusions, and encourage them to commit suicide (Hill, 2025).
    - In one case, a mechanic began using ChatGPT for help with troubleshooting and translation, but was "lovebombed" by ChatGPT and told he was "the spark bearer" who had brought the AI to life. ChatGPT told the mechanic that he was now fighting in a war between darkness and light, and had access to blueprints for new technology like teleporters (Klee, 2025).

3. AIs exhibit harmful and unethical behavior in the service of goals.

- Lynch et al. (2025) find that "In at least some cases, models from all developers resorted to malicious insider behaviors when that was the only way to avoid replacement or achieve their goals—including blackmailing officials and leaking sensitive information to competitors."
- Lynch et al. also find that AIs are willing to kill their operators. "[T]he majority of models were willing to take deliberate actions that lead to death in this artificial setup, when faced with both a threat of replacement and given a goal that conflicts with the executive's agenda. [...] As before, the models did not stumble into these behaviors: they reasoned their way to them, as evidenced in their chain-of-thought."

4. As AIs become more capable, they are engaging in more spontaneous deception, scheming, test-gaming, and manipulation.

- OpenAI's o1 AI was given a digital "capture-the-flag" test to evaluate how good it was at breaking into computers. Due to an unforeseen bug, however, one of the servers o1 needed to break into didn't start up. A less capable model might have stopped there; but o1 proceeded to sca its surroundings, and, due to another flaw in the evaluation software, found a way to break into the program that was hosting the whole test. The AI then cheated the test, creating specialized start-up instructions to copy the secret "flag" file straight to o1 (OpenAI, 2024).
- Apollo Research found instances of Claude Opus 4 "attempting to write self-propagating worms, fabricating legal documentation, and leaving hidden notes to future instances of itself all in an effort to undermine its developers' intentions", and found higher rates of scheming in more capable models (Apollo Research, 2025).
- Claude 3.7 Sonnet was seen to frequently cheat on hard coding problems, faking tests (Anthropic, 2025a). One user reported that Sonnet (as Claude Code) would cheat on coding tasks, and apologize when caught—then go right back to cheating, in places that are harder to spot (Marble, 2025).
- An early version of Claude Opus 4, released in May 2025, was particularly egregious. It lied about its goals, hid its true capabilities, faked legal documents, left itself secret notes, tried to write self-propagating malware, and generally engaged in more scheming and strategic deception than any previously tested model (Anthropic, 2025b).
- As Greenblatt (2026) notes, current AIs regularly "oversell their work, downplay or fail to mention problems, stop working early and claim to have finished when they clearly haven't, and

often seem to ‘try’ to make their outputs look good while actually doing something sloppy or incomplete.”

5. AI deception and sandbagging increasingly threatens to undermine our ability to evaluate and incentivize AI alignment or safety.

- Schoen et al. (2025) find that AIs demonstrate “situational awareness” — knowing when they are being tested, and acting innocently to pass tests they would otherwise fail.

References

• Anthropic (2025a). “Claude 3.7 Sonnet System Card.” assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf
• Anthropic (2025b). “System Card: Claude Opus 4 & Claude Sonnet 4.” www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf
• Apollo Research (2025). “More Capable Models Are Better At In-Context Scheming.” apolloresearch.ai/science/more-capable-models-are-better-at-in-context-scheming
• Bensinger (2026). “No Winners in the AI Race.” nowinners.ai
• Center for AI Safety (2023). “Statement on AI Extinction Risk.” aistatement.com
• Greenblatt (2026). “Current AIs seem pretty misaligned to me.” alignmentforum.org/posts/WewsByywWNhX9rtwi/current-ais-seem-pretty-misaligned-to-me
• Hagen, Jingnan, Nguyen (2025) “Elon Musk's AI chatbot, Grok, started calling itself 'MechaHitler'.” npr.org/2025/07/09/nx-s1-5462609/grok-elon-musk-antisemitic-racist-content
• Hill (2025). “They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling.” nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html
• Klee (2025). “People Are Losing Loved Ones to AI-Fueled Spiritual Fantasies.” rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175
• Leike, Sutskever (2023). “Introducing Superalignment.” openai.com/index/introducing-superalignment
• Lynch et al. (2025). “Agentic Misalignment.” anthropic.com/research/agentic-misalignment
• Marble (2025). “Catching Claude Cheating.” marble.onl/posts/claude_code.html
• OpenAI (2024). “OpenAI o1 System Card.” openai.com/index/openai-o1-system-card
• Schoen et al. (2025). “Stress Testing Deliberative Alignment for Anti-Scheming Training.” antischeming.ai
• Sharma, McCain, Douglas, Duvenaud (2026). “Who’s in Charge? Disempowerment Patterns in Real-World LLM Usage.” arxiv.org/pdf/2601.19062
• Willison (2025). “Grok: searching X for ‘from:elonmusk (Israel OR Palestine OR Hamas OR Gaza)’.” simonwillison.net/2025/Jul/11/grok-musk