Anthropic	**Stated AI Control Plans**	No control plan found. The Responsible Scaling Policy says that when AI can fully automate the work of junior AI researchers, they "will develop an affirmative case that (1) identifies the most immediate and relevant risks from models pursuing misaligned goals and (2) explains how we have mitigated these risks to acceptable levels. The affirmative case will describe [..] mitigations (such as monitoring [..] [Responsible Scaling Policy, 2025].
	Implemented safeguards	None found
DeepSeek	**Stated AI Control Plans**	No public control plan found
	Implemented safeguards	None found
Google DeepMind	**Stated AI Control Plans**	Early sketch of a plan without technical details. **Details:** Relevant section of Frontier Safety Framework: "An initial mitigation approach focuses on detecting when models might develop a baseline instrumental reasoning ability at which they have the potential to undermine human control, assuming no additional mitigations were applied. When models reach this capability level, we believe applying an automated monitor to the model's explicit reasoning (e.g., chain-of-thought output) is an effective mitigation. Once a model is capable of effective instrumental reasoning in ways that cannot be monitored, additional mitigations may be warranted—the development of which is an area of active research." ▪ The Framework states that developers should implement 'control evaluations' [Greenblatt, 2023] and construct a safety case showing: "That the likelihood of deceptive alignment risk would be low enough for safe deployment, even if the model were trying to meaningfully undermine human control." [Google Deepmind, 2025].
	Implemented safeguards	None found
Meta	**Stated AI Control Plans**	No public control plan found
	Implemented safeguards	None found
OpenAI	**Stated AI Control Plans**	No public control plan found Voluntary commitment in a safety framework requires "safeguards meeting a High standard" if the relevant capability threshold is crossed. Exact safeguard not specified, but several are suggested. **Details:** Preparedness Framework states that a "High" risk classification for "Cybersecurity" domain triggers requirement for ".. safeguards meeting High standard against misalignment (Appendix C.2) for large-scale internal deployment". Appendix C.2 contains illustrative safeguards against misaligned models. Sub-section on "Reliable and Robust System Oversight" lists 4 potential measures:" ▪ Monitor AI supervises agent actions to detect agents in the case of trying to take harmful actions, scheming, or similar ▪ System forces human approval of agent actions (either by default or by escalation of the monitor) ▪ Model's reasoning is provided faithfully and interpretably to humans to review ▪ Agent actions are all logged in an uneditable database, and asynchronous monitoring routines review those actions for evidence of harm" [OpenAI, 2025]
	Implemented safeguards	None found
x.AI	**Stated AI Control Plans**	No public control plan found
	Implemented safeguards	None found
Zhipu AI	**Stated AI Control Plans**	No public control plan found
	Implemented safeguards	None found