Anthropic	OpenAI	Google DeepMind	Meta	xAI	DeepSeek	Z.ai	Alibaba Cloud
No public 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].	**No public control plan found.** 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 agent 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]	**Early sketch of a plan exists, without technical details.** From **Frontier Safety Framework** (v3.0): "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 previous iteration of the Framework (v2.0) 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."	**No public control plan found.**	**No public control plan found.**	**No public control plan found.**	The company has indicated in its survey response that it maintains control interventions around emergency response and has demonstrated internal monitoring readiness, **although no formal or publicly available plan has been disclosed.** **(1) Control interventions** The company maintains multiple mechanisms designed to enable rapid containment and mitigation of safety incidents, including i) technical capability to rapidly roll back a deployed model to a previous version globally (within 12h), ii) technical capability to rapidly tighten model safeguards and restrict specific capabilities (e.g., web-browsing) globally **(2) Monitoring readiness** It has i) conducted at least one full live emergency response drill/ simulation in the past 12 months, and has ii) created a formal and documented emergency response plan for AI safety incidents that delineates trigger threshold, named incident commander, and 24*7 duty roster.	**No public control plan found.**