Company Strategy		Quantitative Safety Plan (quantitative bounds on control/ alignment failure risk)
Anthropic	**No explicit strategy found that explains how they will ensure AGI control or alignment, but evidence below that they have regularly updated their research and planning around the issue.** **Updates** No notable AGI strategy updates since May 2025. **Recap from Summer 2025** **Foundational philosophy & Long-term scenarios** In "Core Views on AI Safety" (2023), Anthropic has laid out three possible futures (optimistic, intermediate, and pessimistic) depending on how tractable alignment proves to be. It also identified 6 long-term research pillars: Mechanistic Interpretability, Scalable Oversight, Process-Oriented Learning, Understanding Generalization, Testing Dangerous Failure Modes, and Societal Impact Evaluation. **Foundational governance structure** Anthropic has continuously updated its Responsible Scaling Policy, including the most recent updates in May 2025, to publicize its commitment to pausing model training or deployment if systems reach predefined Capability Thresholds without safety and adequate safeguards. The policy institutionalizes internal oversight through a Responsible Scaling Officer and the Board, mandatory risk assessments, and incident readiness exercises. **Research Agenda** The team has continued to emphasize research effort to manage rapidly advancing model capabilities. In "The Urgency of Interpretability" (2025), CEO Dario Amodei positions interpretability research as a race against accelerating intelligence, aiming by 2027 for tools that can "reliably detect most model problems." Complementing this, Sam Bowman's "Putting up Bumpers" (2025) advances an engineering-based alignment approach built on continuous testing and overlapping safety mechanisms.	No public-facing quantitative safety plan found
OpenAI	**No explicit strategy found that explains how they will ensure AGI control or alignment, but evidence below that they have regularly updated their research and planning around the issue.** **Update** In "Security on the Path to AGI," [1], OpenAI has shared their security initiatives on advancing to AGI, including an expanded Cybersecurity Grant Program and Bug Bounty Program, partnerships for continuous adversarial red teaming, deployment of AI-powered cyber defense systems, stronger safeguards for advanced AI agents such as Operator and Stargate, and adoption of zero-trust, hardware-backed infrastructure to scale security alongside advancing model capabilities. **Research Agenda** The company believes in avoiding optimization that encourages obfuscation: Developers should exercise caution when applying optimization pressures to model reasoning, especially when removing 'undesired reasoning', to prevent fostering deceptive behavior. In the company survey, the company stated that "We've published research and joined a broader working paper urging against optimizing on chains of thought: As we noted in the GPT-5 system card, "our commitment to keep our reasoning models' CoTs as monitorable as possible (i.e., as faithful and legible as possible) allows us to conduct studies into our reasoning models' behavior by monitoring their CoTs." **Recap from Summer 2025** **Foundational philosophy and strategy** OpenAI stated in its strategy "How we think about safety and alignment," that it has shifted from viewing AGI as a single transformative moment to seeing it as continuous progress. It further listed its core principles that currently guide the company's thinking and actions, which include Embracing uncertainty, Defense in Depth, Methods that Scale, Human Control, and Community Effort. For every principle, the blog lays out how it will shape their focus and approach to new challenges and relates to already implemented interventions. This thinking iterates on the 2023 blog post "Planning for AGI and beyond," emphasizing goals including ensuring AGI benefits are "widely and fairly shared" and advocates for deploying progressively more powerful systems to learn iteratively. **Foundational governance structure** Preparedness Framework, which is updated in April 2025, describes OpenAI's commitment to pausing development or deployment if required mitigations cannot adequately address the identified risks based on regular dangerous capability evaluations and the predefined capability threshold triggers.	No public-facing quantitative safety plan found
Google DeepMind	**No explicit strategy found that explains how they will ensure AGI control or alignment, but evidence below that they have regularly updated their research and planning around the issue.** **Update** Google DeepMind has updated its Frontier Safety Framework in September 2025. Compared to v2.0, the updated version introduced new risk domain (harmful manipulation) in the misuse risk section, broadening the section on misalignment risks from deception, increased transparency on external disclosure, and expand mitigation coverage to large-scale internal deployment. **Recap from Summer 2025** **Research agenda and efforts** An Approach to Technical AGI Safety and Security (April 2025) A detailed technical report by DeepMind's safety team explains their research agenda for preventing severe, civilisation-scale harm from AGI—defined as systems roughly at the 99th-percentile of skilled adults. The paper identifies four areas of risk: misuse, misalignment, mistakes, and structural risks and chooses to focus on technical approaches to misuse and misalignment. The strategy for misuse is to proactively identify dangerous capabilities and implement robust security, access restrictions, monitoring, and model safety mitigations to prevent threat actors from accessing these dangerous capabilities. The strategy for misalignment is "two lines of defense," including model-level mitigations + system-level security measures. The safety-case methodology serves as the integrative layer connecting these safeguards, as it proposes making deployment decisions through structured, evidence-based arguments: inability cases (model lacks capability) and control cases (misaligned behaviour will be caught). **Foundational governance structure** Frontier Safety Framework (v 2.0) Set of voluntary commitments based on regular dangerous capability evaluations and a set of capability thresholds in high-risk domains that trigger a requirement for enhanced safety and security mitigations. These commitments include pausing development or deployment if the required mitigations cannot adequately manage the identified risks.	No public-facing quantitative safety plan found
Meta	**No existential safety strategy found, but evidence below that the company has started to engage with the topic.** **Update** Meta's Shift on Open-Source AI In July 2025, Mark Zuckerberg wrote in a blog post "Personal Superintelligence," that Meta "will need to be rigorous about mitigating these risks and careful about what [it] choose to open source. Still, [Meta] believe that building a free society requires that [it] aim to empower people as much as possible." **Recap from Summer 2025** **Foundational philosophy** Open Source AI Is the Path Forward (2024) In this blog post, Zuckerberg presents a case for open source AI as their primary approach to AI safety and development (not specifically focused on catastrophic risks). The document makes the case that open source models are inherently safer than closed alternatives due to transparency, distributed scrutiny, and prevention of power concentration. **Foundational governance structure** Frontier AI Framework v.1.1 (2025) Set of voluntary commitments based on regular dangerous capability evaluations and a set of capability thresholds in high-risk domains that trigger a requirement for enhanced safety and security mitigations. These commitments include pausing development or deployment if the required mitigations cannot adequately manage the identified risks.	No public-facing quantitative safety plan found
xAI	**No existential safety strategy found, but evidence below that the company has started to engage with the topic.** **Update** xAI Risk Management Framework (August 2025) The formalized RMF outlines xAI's approach to policies for handling significant risks associated with the development, deployment, and release of AI models such as Grok. It identifies quantitative thresholds and metrics for a few critical risks, and lays out procedures that could be used to manage and improve the safety of AI systems. **Recap from Summer 2025** **Foundational governance structure** xAI Risk Management Framework (Draft) Set of voluntary commitments based on regular dangerous capability evaluations and a set of capability thresholds in high-risk domains that trigger a requirement for enhanced safety and security mitigations.	No public-facing quantitative safety plan found
DeepSeek	**No public-facing existential risk policy found**	No public-facing quantitative safety plan found
Z.ai	**The company has indicated in the company survey that it doesn't yet have an AGI explicit existential risk strategy, but is actively developing one.**	No public-facing quantitative safety plan found
Alibaba Cloud	**No public-facing existential risk policy found**	No public-facing quantitative safety plan found