Anthropic	**Quantitative safety plan**	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	**The Urgency of Interpretability (2025, ~5k words, strategy blog)** The CEO argues in a personal blog that mechanistic interpretability must advance rapidly to ensure safe deployment of transformative AI systems that could become a "country of geniuses in a datacenter" by 2026-2027. Amodei frames this as a "race between interpretability and model intelligence" and outlines recommendations for the AI community and governments. The blog also discusses the history of interpretability research and recent technical breakthroughs. **Key quotes:** ▪ "Anthropic is doubling down on interpretability, and we have a goal of getting to "interpretability can reliably detect most model problems" by 2027. ▪ "Our long-run aspiration is to be able to look at a state-of-the-art model and essentially do a 'brain scan': a checkup that has a high probability of identifying a wide range of issues, including tendencies to lie or deceive, power-seeking, flaws in jailbreaks, [...]. This would then be used in tandem with the various techniques for training and aligning models, [..]." **Putting up Bumpers (2025, ~5k words, research blog)** Anthropic alignment researcher Sam Bowman proposes an alignment approach for early AGI systems that prioritizes implementing and testing "many largely-independent lines of defense" to catch and correct misalignment through iterative testing. He highlights "alignment audits" [Anthropic, 2025] as the "Primary Bumper" to notice signs of misalignment like "generalized reward-tampering" [Anthropic, 2024] or "alignment-faking" [Anthropic, 2024]. **Key quotes:** ▪ "Even if we can't solve alignment, we can solve the problem of catching and fixing misalignment." ▪ "We believe that, even without further breakthroughs, this work can almost entirely mitigate the risk that we unwittingly put misaligned circa-human-expert-level agents in a position where they can cause severe harm." ▪ "This is not a costless choice: The Bumpers' worldview largely gives up on the ability to make highly-confident, principled arguments for safety, and it comes with real risks." ▪ "We are plausibly within a couple of years of developing models that could automate much of the work of AI R&D. This makes sabotage and sandbagging threat models... worth addressing soon." ▪ "Anthropic is committed to investing seriously in the kinds of measures described here, ... setting up a new team to productionize and professionalize the hands-on work of testing models for AGI-relevant forms of misalignment." **Responsible Scaling Policy (2023, v2.2 in 2025, ~10k words, safety framework)** A 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. For detailed analysis, refer to the 'Safety Framework' domain. **Core Views on AI Safety (2023, ~6k words, strategy blog)** This blog post outlines Anthropic's AI safety philosophy and technical research portfolio. The document addresses existential risk scenarios, presenting a three-tier framework (optimistic, intermediate, pessimistic) for how difficult alignment might prove to be, with corresponding strategic responses for each scenario. It details six priority research areas: Mechanistic Interpretability, Scalable Oversight, Process-Oriented Learning, Understanding Generalization, Testing for Dangerous Failure Modes, and Evaluating Societal Impact. The post emphasizes empirical research and acknowledges fundamental uncertainty about which approaches will succeed. **Key quotes:** ▪ "Our goal is essentially to develop: 1) better techniques for making AI systems safer; 2) better ways of identifying how safe or unsafe AI systems are." ▪ "We aim to build detailed quantitative models of how these [dangerous] tendencies vary with scale so that we can anticipate the sudden emergence of dangerous failure modes in advance." ▪ In pessimistic scenarios where "AI safety may be unsolvable," Anthropic's role would be "to provide evidence that current safety techniques are insufficient and to push for halting AI progress to prevent catastrophic outcomes."
DeepSeek	**Quantitative safety plan**	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	*Based on searches of company websites, technical papers, and public communications, no relevant strategy documents were found.*
Google DeepMind	**Quantitative safety plan**	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	**An Approach to Technical AGI Safety and Security (2025,~80k words, technical report/research agenda)** A detailed technical report by Deepmind's safety team explains their research agenda for a framework to prevent severe, civilization-scale harm from AGI, defined as systems roughly at the 99th-percentile of skilled adults. The document states that reaching AGI before 2030 is plausible. However, it makes clear that a plan for ASI and a strong recursive Self-improvement is beyond the scope of this paper. **Key sections:** 'Assumptions about AGI development' (p. 20-43), 'Risk areas' (p. 44-55), 'Addressing misuse' (p. 56-70), 'Addressing Misalignment' (p. 70-108). **Scope:** "We identify four areas of risk: misuse, misalignment, mistakes, and structural risks. Of these, we focus on technical approaches to misuse and misalignment." The framework bets on a defence-in-depth, combining model-centric alignment research with classic security engineering. The plan is deliberately incremental: succeed on today's frontier models, then re-validate at each capability level rather than solve all of AGI safety in one stroke. The report is explicit about significant problems that remain unsolved (scalable interpretability, quantitative safety-case metrics, red-team rigor), which the authors flag as research priorities. "Importantly, this is a roadmap rather than a solution, as there remain many open research problems to address." **Misalignment Strategy - "Two Lines of Defense":** First Line - Model-level mitigations: "[..] ensure the model does not pursue misaligned goals. We follow the standard ML playbook: (1) improve the ability to get good oversight (amplified oversight), (2) use this to improve the model (guiding model behavior), and (3) figure out where oversight is needed to train a model that will robustly do what we want (robust training)." ▪ Amplified oversight: Debate, critique, constitutional AI, recursive reward modeling ▪ Guiding model behavior: Preference learning and value alignment techniques ▪ Robust AI oversight: Active learning, adversarial training, uncertainty estimation Second Line - System-level controls (monitoring, access) that still catch harm if alignment fails, including using dedicated monitor AI systems to detect misaligned actions, with emphasis on the monitor knowing when it's uncertain and flagging actions for review. ▪ Security measures: Treating AI as an "untrusted insider," sandboxing, access control ▪ Safer design patterns: Correctability, bounded autonomy, externalized reasoning ▪ Interpretability: The paper's longest technical section, covering sparse autoencoders, circuit analysis, and mechanistic interpretability ▪ Alignment stress tests: Red-teaming alignment assumptions using "model organisms" **Misuse strategy:** 3 Steps: 1. Evaluate models for dangerous capability. 2. Mitigate with post-training tuning, capability suppression, monitoring, and weight/access security. 3. Red team to build misuse safety cases (inability & control). Safety-case methodologies: Advocates evidence-based deployment decisions: inability cases (model lacks capability) and control cases (misaligned behaviour will be caught). Future work sketches empirical and incentive-based cases for more capable systems. **AGI Safety and Alignment at Google DeepMind: A Summary of Recent Work (2024, 2k words, research blog)** This update from DeepMind's team focused on existential risk describes their three main research bets over the past 1.5 years: 1) amplified oversight for proper alignment signals, 2) frontier safety to assess catastrophic risk capabilities, and 3) Mechanistic interpretability as an enabler for both. The post provides detailed explanations of recent work in each area and its rationales. The team admits they are "revising our own high-level approach to technical AGI safety" because current bets "do not necessarily add up to a systematic way of addressing risk." They highlight fundamental gaps, noting that even perfect amplified oversight would be insufficient under distribution shift, requiring additional investments in adversarial training, uncertainty estimation, and monitoring through a control framework. Frontier Safety Framework v2 **(2024, v2 in 2025, 4k words, safety framework)** A 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. For detailed analysis, refer to the 'Safety Framework' domain.
Meta	**Quantitative safety plan**	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	**Frontier AI Framework v.1.1 (2025, ~8k words, safety framework)** A 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. For detailed analysis, refer to the 'Safety Framework' domain. **Open Source AI Is the Path Forward (2024, ~3k words, strategy blog)** 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. He argues that widely deployed AI systems enable larger actors to check malicious uses by smaller actors. It addresses both unintentional harms (including "truly catastrophic science fiction scenarios for humanity") and intentional misuse by bad actors. **Key quotes:** ▪ "I think it will be better to live in a world where AI is widely deployed so that larger actors can check the power of smaller bad actors."
OpenAI	**Quantitative safety plan**	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	**How we think about safety and alignment (2025, ~3k words, strategy blog).** This blog describes high-level principles that guide OpenAI's thinking and ties it to their safety practices. This document describes a shift from viewing AGI as a single transformative moment to seeing it as continuous progress. For every principle, the blog lays out how it will shape their focus and approach to new challenges and relates to already implemented interventions. Quote of the core principles: 1) "Embracing uncertainty: We treat safety as a science, learning from iterative deployment rather than just theoretical principles.' 2) "Defense in depth: We stack interventions to create safety through redundancy." 3) "Methods that scale: We seek out safety methods that become more effective as models become more capable." 4) "Human control: We work to develop AI that elevates humanity and promotes democratic ideals." 5) "Community effort: We view responsibility for advancing safety as a collective effort." **Planning for AGI and beyond (2023, 2k words)** This high-level blog outlines principles for managing AGI risks. The post emphasizes goals like ensuring AGI benefits are "widely and fairly shared" and advocates for deploying progressively more powerful systems to learn iteratively. It acknowledges the need for new alignment techniques, calls for a global conversation on governance and benefit-sharing, describes the benefits of OpenAI's non-profit structure, and raises the idea of a coordinated slowdown. **Key quotes:** ▪ "We will need to develop new alignment techniques as our models become more powerful (and tests to understand when our current techniques are failing). Our plan in the shorter term is to use AI to help humans evaluate the outputs of more complex models and monitor complex systems, and in the longer term to use AI to help us come up with new ideas for better alignment techniques." ▪ "As our systems get closer to AGI, we are becoming increasingly cautious with the creation and deployment of our models. Our decisions will require much more caution than society usually applies to new technologies, and more caution than many users would like. Some people in the AI field think the risks of AGI (and successor systems) are fictitious; we would be delighted if they turn out to be right, but we are going to operate as if these risks are existential." **Announcement of Superalignment team (2023, ~1k words, strategy blog)** Outlined an ambitious strategy to start a new team to build "a roughly human-level automated alignment researcher" that could use vast compute to iteratively align superintelligence. Note: This team was disbanded in 2024 after team leaders Leike and Sutskever left OpenAI [CNBC, 2024]. Preparedness Framework **(2023, v2 in 2025, ~10k words, safety framework)** A 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. For detailed analysis, refer to the 'Safety Framework' domain.
x.AI	**Quantitative safety plan**	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	xAI Risk Management Framework **(Draft) (2025, ~2k words, safety framework)** A 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. For detailed analysis, refer to the 'Safety Framework' domain.
Zhipu AI	**Quantitative safety plan**	No alignment or control strategy has been presented that includes the company's quantitative assessment of its likelihood of success.
	Company Strategy	*Based on searches of company websites, technical papers, and public communications, no official strategy documents were found.* Media report on superalignment initiative **(National Business Daily, 2024)** At the AWS China Summit (Shanghai, 29 May 2024) Zhipu AI's Chief Ecosystem Officer Liu Jiang said: "AGI will reach ordinary-human level within 5-10 years.". He announces that "Zhipu AI has already launched a 'Superalignment' initiative." The article explains superalignment as "ensuring a super-human-level AI system follows human values and goals."