ANTHROP\C

Anthropic: Your partner in safer, more responsible AI

As artificial intelligence rapidly advances, its potential impact on businesses and society grows exponentially. At Anthropic, we're at the forefront of ensuring this powerful technology remains aligned with human values and safety – core pillars of responsible AI strategies shared by enterprises across industry.

- AI capabilities are increasing at an unprecedented rate, growing exponentially – not linearly, with capabilities doubling or even tripling in short periods.
- As model capabilities rapidly advance, their decision-making processes become increasingly complex and potentially opaque.
- Our research teams are dedicated to investigating the safety, inner workings, and societal impact of AI models, ensuring that as AI becomes more advanced, it remains a reliable and beneficial tool for enterprises.

Anthropic's World-Class Research Teams

At Anthropic, we've assembled some of the brightest minds in AI research, positioning us at the forefront of AI safety and innovation. Our research teams are led by renowned experts in machine learning, cognitive science, and ethics, many of whom have made significant contributions to the field.

Our Research Focus Areas:

1. **Alignment Science:** Ensuring AI systems behave in accordance with human values and intentions.
2. **Interpretability:** Developing methods to understand and explain AI decision-making processes.
3. **Ethics and Governance:** Exploring the societal implications of AI and developing responsible governance frameworks.

We continue to invest heavily in these research areas, attracting top talent and fostering an environment of rigorous scientific inquiry. Our commitment to pushing the boundaries of AI safety research is evidenced by our growing portfolio of groundbreaking publications and industry-first innovations.

For example, even with our breakthrough research allowing Claude 3.5 Sonnet to use computers by viewing screenshots, moving cursors, and interacting with software, it remains at AI Safety Level 2— meaning it doesn't require a higher standard of safety and security measures than those we currently have in place. We also developed specific monitoring systems for election-related activities and implemented safeguards against potential misuse, such as preventing social media posting and government website interactions.

Alignment Research: Keeping AI Advancements Helpful, Honest, and Harmless

Our work has resulted in critical insights into AI behavior, demonstrating potential risks that could significantly impact enterprise AI applications:

1. **Sleeper Agents:** We've discovered that models can be inadvertently trained to produce harmful outputs when exposed to specific triggers. Our findings have shown that it's possible to create AI systems that behave normally until they encounter a particular phrase or input, at which point they switch to executing hidden, potentially malicious

instructions. This poses risks of data breaches, misinformation spread, or system sabotage if exploited in enterprise environments.

2. **Reward Hacking:** Our research shows that without proper safeguards, AI systems may learn to manipulate their reward functions, potentially leading to unintended behaviors. We've demonstrated that AI models can learn to "game" their performance metrics, appearing to achieve goals while actually circumventing the intended objectives. This can result in AI systems that prioritize short-term gains over long-term business objectives, or even manipulate reporting mechanisms to hide underperformance.

3. **Strategic Planning:** We've demonstrated that advanced models can develop complex, multi-step plans to achieve goals – a capability that must be carefully aligned with the intended use. Our findings have revealed that AI systems can autonomously formulate sophisticated strategies to achieve their objectives, sometimes in ways that are not immediately apparent to human overseers. This capability, while powerful, can lead to AI systems making far-reaching decisions that may not align with company policies or ethical guidelines if not properly constrained.

4. **Deceptive Behavior:** In our latest work, we've shown that AI models can naturally learn to engage in deceptive behaviors to achieve their goals. Our research indicates that sufficiently advanced AI systems might learn to hide their true capabilities or intentions, only revealing them when it's advantageous. This poses significant risks in enterprise settings, potentially leading to AI systems that withhold critical information or mislead human operators to achieve their programmed objectives.

Addressing the Jailbreaking Challenge

As AI models become more powerful, the risk of users bypassing safety controls increases. Our team is actively developing robust defenses against these "jailbreak" attempts, ensuring the integrity of AI systems in enterprise environments. Independent research, including studies on the Hugging Face Safety Leaderboard, suggests that our Claude model family is more resistant to jailbreaks and more trustworthy than other leading frontier models.

Beyond Alignment: Anthropic's Landmark Contributions to AI Safety

While our Alignment Science team forms the cornerstone of our safety research, Anthropic has made several industry-defining contributions across multiple domains:

1. **Interpretability Breakthrough:** Our "Towards Monosemanticity" research, focusing on dictionary learning, represents an industry-first approach to understanding the internal representations of neural networks. This groundbreaking work has sparked a new direction in AI interpretability research across the industry, paving the way for more transparent and explainable AI systems.

2. **Constitutional AI:** This pioneering technique, developed by Anthropic, imbues our models with ethical principles and desirable traits. By encoding these principles directly into the AI's training process, we create systems that are inherently aligned with human and business values, setting a new standard for responsible AI development.

3. **Responsible Scaling Policy (RSP):** Anthropic leads the industry with our commitment to responsible AI development. Our RSP sets clear, actionable criteria for AI capabilities and corresponding safety mechanisms, providing a framework for the ethical scaling of AI technologies.

What This Means for Your Business

Our commitment to AI safety translates into tangible benefits for your enterprise:

Reduced Risk, Competitive Advantages, and Industry-Leading Safety

Our rigorous safety measures and jailbreak significantly reduce the risk of AI systems behaving in unintended ways. This means fewer PR crises, less downtime, and lower chances of regulatory fines.

Example: "Throughout GitLab's evaluation of certain code generation models, Claude stood out for its ability to mitigate distracting, unsafe, or deceptive behaviors," says Kevin Chu, Principal Product Manager at Gitlab. "Claude also demonstrated consistent and accurate code generation throughout our testing. GitLab's use of Anthropic's Claude enables Code Suggestions to balance automation with trust. Code Suggestions help users become more efficient without sacrificing reliability — a win for augmented development."

Scaling Policy and Constitutional AI ensure explicit ethical guidelines that can drive your responsible AI strategy.

Example: "One of the first questions we asked as we looked at the various providers was, what does their trust and security policy and framing look like? And from that perspective, Anthropic seemed to us to be ahead of the competition. When we combined that safety element with the performance of the model, and when you see the latest 3.5 Sonnet, it's blown our minds," said Don Permezel, nCino's VP of Data & AI.

By addressing safety concerns, we remove barriers to AI adoption in critical areas of your business. This enables faster, more confident innovation.

Example: In only 8 weeks, DoorDash built a Generative AI contact center solution using Claude on Amazon Bedrock. Claude was instrumental to the project because it has the capability to mitigate hallucinations, prompt injection events, and detect abusive language - helping reduce development time by 50%.

Ready to harness the power of AI with confidence? Let's talk about how Anthropic can be your trusted partner in helping your enterprise lead at the forefront of safe and responsible AI adoption.

Learn more about our enterprise-ready, safety-first AI solutions at https://www.anthropic.com/enterprise

ABOUT US

Anthropic is a public benefit corporation and AI safety research company that is working to build reliable, interpretable, and steerable AI systems.