# UK launches £15m Alignment Project to tackle AI safety challenge



The UK’s AI Security Institute has launched the Alignment Project, a £15 million programme aimed at solving one of artificial intelligence’s most pressing challenges: ensuring advanced AI systems behave as intended. The initiative brings together major players including Amazon, Anthropic, the Canadian AI Safety Institute and several philanthropic funders.

AI alignment refers to the task of making AI systems act according to their design, even as they surpass human capabilities in certain areas. Geoffrey Irving, Chief Scientist at the AI Security Institute, warned that current control methods may soon be inadequate. “Misaligned, highly capable systems could act in ways beyond our ability to control, with profound global implications,” he said.

The Alignment Project will provide funding, computing power and support for interdisciplinary collaboration. Its aim is to build safer, more predictable AI tools that uphold national security and public trust.

Backers include Halcyon Futures, the Safe AI Fund, Schmidt Sciences, UK Research and Innovation, and the Advanced Research and Invention Agency. Jack Clark, Co-founder and Head of Policy at Anthropic, stressed the importance of deepening understanding of AI systems as their intelligence increases. Anthropic is also involved in discussions around deploying its AI chatbot, Claude, to improve UK government services, reflecting Prime Minister Keir Starmer’s push to make the UK a global AI leader.

The AI Security Institute has become a central player in the UK’s AI strategy. Launched in 2023 with government backing estimated at £100 million, it was formed after a high-level summit between tech CEOs and then Prime Minister Rishi Sunak. The institute works closely with OpenAI, Google DeepMind and Anthropic to evaluate model risks and inform regulatory development.

Despite rapid progress, the institute faces challenges. The Financial Times reports recruitment issues and delays to its planned San Francisco office, partly due to shifting political dynamics in the US. Still, international cooperation remains central to the UK’s AI governance model.

Amazon’s £4 billion partnership with Anthropic, recently approved by UK regulators, adds commercial weight to the sector. The Competition and Markets Authority found no competition concerns, clearing the way for joint projects that could support AI safety research and deployment.

The UK’s efforts mirror moves in other jurisdictions. In the US, OpenAI and Anthropic have signed government agreements to support AI safety testing and regulatory preparedness. These international frameworks reflect a shared recognition that ethical AI requires collective action.

The Alignment Project marks a major step in securing the future of transformative AI. By combining public investment, private sector innovation and international collaboration, the UK is working to ensure AI’s development remains safe, beneficial and broadly aligned with human goals.

Created by [Amplify](https://www.hbmadvisory.com/amplify): AI-augmented, human-curated content.

## Bibliography

1. <https://www.uktech.news/ai/ai-security-institute-launches-anthropic-backed-alignment-scheme-20250730> - Please view link - unable to able to access data
2. <https://www.gov.uk/government/news/ai-security-institute-launches-international-coalition-to-safeguard-ai-development> - The UK’s AI Security Institute has launched an international coalition to safeguard AI development. This initiative brings together partners including the Canadian AI Safety Institute, Amazon Web Services, Anthropic, and civil society to focus on AI behaviour and control. The project is backed by over £15 million and aims to ensure AI systems behave predictably and as designed, supporting the UK government's Plan for Change by enabling the benefits of AI while providing strong national security foundations.
3. <https://www.reuters.com/technology/artificial-intelligence/uk-explore-use-anthropics-ai-chatbot-claude-public-services-2025-02-14/> - The UK government is considering deploying Anthropic's AI chatbot, Claude, to enhance public interaction with information and services. Supported by Google and Amazon, this effort aligns with Prime Minister Keir Starmer's vision to position the UK as a leader in artificial intelligence. The initiative aims to improve productivity in public services through AI while fostering a pro-innovation regulatory environment, making public documents and information more accessible and streamlining administrative processes.
4. <https://www.reuters.com/technology/artificial-intelligence/openai-anthropic-sign-deals-with-us-govt-ai-research-testing-2024-08-29/> - OpenAI and Anthropic have signed agreements with the U.S. government for AI research and testing, as announced by the U.S. Artificial Intelligence Safety Institute. These agreements aim to address regulatory concerns over the ethical and safe application of AI technologies amidst impending California legislation. The U.S. AI Safety Institute will have early access to new AI models from both companies to evaluate their capabilities and potential risks, focusing on improving safety standards.
5. <https://www.apnews.com/article/8599c65e6e0fe87869929d20ab5d600e> - Amazon's $4 billion partnership with AI startup Anthropic has received UK competition clearance, with the Competition and Markets Authority determining that the deal does not warrant further investigation. This decision follows a review of the agreement, which is part of a broader global scrutiny of major tech companies investing in leading generative AI startups. The regulator concluded that the combined market share of Anthropic and Amazon in the UK was not significant enough to require an in-depth investigation under the country's merger rules.
6. <https://www.time.com/7204670/uk-ai-safety-institute/> - In May 2023, CEOs of major AI companies, including OpenAI, Google DeepMind, and Anthropic, met with then UK Prime Minister Rishi Sunak to discuss AI risks following the launch of ChatGPT. This led to the establishment of the UK's AI Safety Institute (AISI) in November 2023, with a mandate to evaluate the risks of new AI models. The AISI, funded with £100 million, quickly became a leading body in AI safety testing, gaining prerelease access to models from these companies.
7. <https://www.ft.com/content/c9f6067c-3faa-4e6a-bc0f-2ed7290a8476> - The UK's efforts to strengthen its role in global AI regulation through the AI Safety Institute (AISI) are facing challenges, including delays in launching a San Francisco office and the election of Donald Trump, who has a divergent approach to AI regulation. Established last year with a £50 million budget and 100 employees, AISI collaborates with tech giants like OpenAI and Google to assess AI risks. However, difficulties in recruitment and the US elections have stalled expansion.