PROJECT
MANAGEMENT
GLOBAL
The Place for People in Projects

UK Leads New Global Initiative to Ensure AI Safety and Alignment

August 5, 2025

In a significant move to ensure AI systems are developed responsibly and aligned with human values, the UK government has launched a new global initiative backed by more than £15 million in funding. The “Alignment Project” aims to make AI systems predictable and trustworthy, addressing a critical challenge in the technology’s rapid evolution.

The project, spearheaded by the UK’s AI Security Institute, is a collaborative effort with a coalition of international partners, including the Canadian AI Safety Institute, CIFAR, Schmidt Sciences, Amazon Web Services (AWS), Anthropic, and other key industry and philanthropic organizations. This partnership reflects a growing international consensus that AI alignment is a shared responsibility, vital for national security and public trust.

Science, Innovation, and Technology Secretary Peter Kyle highlighted the importance of the initiative, stating, “As advanced AI systems surpass human performance in some areas, it is crucial we drive forward research to ensure this technology behaves in our interests. This fund will help us make AI more reliable, more trustworthy, and capable of delivering the growth and better public services that are central to our Plan for Change.”

The Alignment Project will provide three levels of support to accelerate research and development:

- **Grant Funding:** Up to £1 million in grants for researchers across various disciplines, from computer to cognitive science.
- **Compute Access:** Up to £5 million in dedicated cloud computing credits from AWS, enabling large-scale technical experiments.
- **Venture Capital:** Investment from private funders to support the commercialization of alignment solutions.

This multi-faceted approach is designed to overcome traditional barriers in alignment research by combining funding, infrastructure, and market incentives. The initiative will focus on key areas, including ensuring AI systems remain responsive to human oversight and continue to follow our goals as they become more capable.

Geoffrey Irving, Chief Scientist at the AI Security Institute, emphasized the urgency of the project, noting that “misaligned, highly capable systems could act in ways beyond our ability to control, with profound global implications. This project tackles this head-on by bringing together a global coalition to close critical gaps in alignment research and increase the chance that transformative AI systems serve humanity reliably and safely.”

The UK is uniquely positioned to lead this effort, building on the work of its AI Security Institute and its world-leading ecosystem of AI companies and research institutions. Governments, philanthropists, and industry partners are encouraged to join the project to help accelerate progress and ensure AI safety keeps pace with the rapid advancement of the technology.