

POLICY MEMORANDUM

RE: Proposal for U.S.-China Technical Cooperation on AI Safety

DATE: October 27, 2024

Overview.

This memorandum proposes a phased framework for technical cooperation on AI safety between the U.S. and China, implemented through the U.S. AI Safety Institute (AISI) with robust security measures. The framework focuses on joint testing of non-sensitive AI systems, confidence-building technical exchanges, and whistleblower protections modeled on California's SB 1047, with oversight from the Department of Justice (DOJ). This effort addresses shared risks from advanced AI systems and aligns with the October 2024 National Security Memorandum (NSM) on AI, promoting global safety standards while safeguarding U.S. strategic interests.

Key Recommendations.

- Establish bilateral working groups through the AISI focused on specific technical challenges, such as AI alignment and safety testing.
- Implement a blue team vs. red team framework, where U.S. and Chinese researchers collaborate to stress-test AI systems and identify safety risks.
- Develop federal whistleblower protections based on California's SB 1047, with DOJ oversight to ensure safe reporting of AI risks.

Background.

As AI technologies advance, both the U.S. and China face growing risks related to AI misalignment, system failures, and unintended behaviors. While U.S. policy has traditionally framed AI development as a competitive domain, areas of mutual interest—such as preventing catastrophic AI failures—create opportunities for limited cooperation. China's participation in initiatives like the Bletchley Declaration and UN resolutions on AI suggests willingness to engage on technical safety issues. Collaborating with China on non-military AI safety challenges offers a path to build trust incrementally and create global safety protocols. With strict security safeguards, the U.S. can pursue these opportunities while protecting sensitive technologies and maintaining its competitive edge.

Situation Evaluation.

Technology Challenges

Advanced AI systems present common risks, such as misalignment, lack of robustness, and transparency issues, that affect all developers regardless of national origin. Current safety testing methodologies are still immature and could benefit from broader collaboration. A blue team vs. red team approach—where U.S. and Chinese researchers

collaborate to identify safety flaws—offers a practical way to improve system reliability. Focusing on non-sensitive applications like healthcare and environmental AI minimizes risks while fostering cooperation.

Domestic Considerations

Ensuring the success of this initiative requires robust domestic safety reporting mechanisms. Whistleblower protections modeled on California’s SB 1047 will allow individuals to report AI risks safely, with DOJ oversight ensuring compliance and accountability. These protections will complement international efforts by encouraging transparency and early detection of potential failures within the U.S.

Political Landscape and Risks

While this cooperation aligns with the NSM’s objectives, it may face political resistance due to broader U.S.-China tensions. Engaging Congress and the public with transparency will be essential to build support. Stringent security protocols—such as sandbox testing environments and pre-approved research scopes—will minimize the risk of technology transfer. If cooperation is misused or becomes counterproductive, off-ramps will allow for immediate disengagement without jeopardizing U.S. interests.

Proposal Evaluation.

This initiative is feasible given existing institutional frameworks and legal models. The AISI provides a secure venue for technical collaboration, while SB 1047 offers a good model for federal whistleblower protections. The project will require \$50 million in initial funding, with an 18-month phased rollout starting with pilot projects in non-sensitive AI applications. The desirability of this initiative lies in its ability to reduce shared AI risks while maintaining U.S. leadership in global AI governance. Establishing communication channels through joint technical efforts will help manage potential crises, and the creation of shared safety standards will strengthen global AI safety protocols. Risks will be mitigated through sandbox environments, limited research scopes, and regular compliance reviews. Transparent engagement with Congress and quarterly updates will ensure the project remains aligned with U.S. interests.

We aim to begin with narrow, non-sensitive pilot projects in areas like healthcare and environmental AI, where collaboration carries minimal risks. Federal whistleblower protections, modeled on SB 1047, will ensure that domestic safety risks are reported promptly and handled transparently. Strict security measures—including sandbox testing environments and limited research scopes—will protect U.S. technologies and national interests. Quarterly reviews will track progress and ensure that cooperation remains on course. If necessary, off-ramps will allow the U.S. to halt participation swiftly if security risks or misuse arise.

Conclusion.

By leveraging selective technical cooperation with China, the US will be able to find common ground on critical AI safety risks and continue to lead in developing global AI safety standards while protecting its interests. The integration of domestic whistleblower protections ensures accountability and early risk detection. If successful, this initiative will strengthen global AI governance, reduce the risk of AI misalignment, and demonstrate that cooperation is possible even amid strategic competition.