MIRI's Proposed International Agreement to Prevent the Premature Creation of Artificial Superintelligence ("Plan S")

MIRI

BOTTOM LINE

AI companies are racing to build AI far smarter than all humans. Many experts think such "artificial superintelligence" (ASI) poses an extinction risk to humanity and should not be built.[1] MIRI proposes an agreement to prevent the creation of superintelligence, globally, for as long as is desired. It allows the use of existing AIs while banning the development of smarter AIs. It is backed by verification methods so the U.S. and the PRC can coordinate without trusting each other.

This memo assumes the reader is concerned about catastrophic risks from superintelligence. If not, we recommend other memos first.[2]

The full agreement text and commentary are available at intelligence.org/agreement.

The Window for Action May Close Soon

- AI companies are rapidly automating the process of AI development itself. If they succeed, and those AIs are widespread, it will be much harder to slow down.
- Nobody can predict when it will be "too late." Urgent, preemptive action is needed: leaders should immediately halt the training of smarter AI models.

The Solution: A Global Ban on the Development of Superintelligence

- No AI developer in the U.S. nor the PRC knows how to keep superintelligence on a leash. If anyone builds it, everyone dies.
- Therefore, America needs to prevent superintelligence development globally.

Key Components of the Agreement

- **The U.S. and PRC lead an international effort to enact, verify, and enforce an agreement.**
- Trust is not required: The U.S. and PRC use national intelligence and standardized verification procedures to ensure everyone is following the rules. Monitoring focuses on AI inputs.
- **Key Idea: No training smarter AIs. Existing AIs can keep running.**[3]
- AI chips are consolidated into declared, monitored data centers where they can run existing AIs but not train new ones. Intelligence agencies search for undeclared sites.
- No new large-scale AI training; exceptions can be made for safe, beneficial uses of AI such as medical research. Medium-scale training must be overseen.
- No researching frontier AI methods or techniques to circumvent monitoring.
- World leaders who recognize ASI as an extinction threat will seek to prevent it globally. This agreement gives them well-defined, checkable rules to do so. Countries that cheat or pursue ASI outside the agreement face clearly stated escalation: export controls, economic sanctions, counterproliferation, and sabotage of their AI development.

Governance Structure

- Aim for global participation
- U.S.-China Executive Council, can expand to include others
- Restrictions and verification coordinated by technical body

Chip Controls

- Training compute thresholds
- Locate and consolidate existing AI chips, track new production
- Verify chip use in monitored facilities

Research Restrictions

- Restrict ASI-relevant research
- Restrict research that endangers verification
- Verification via interviews, whistleblowers, etc.

Nonproliferation and Enforcement

- Export controls on AI chips and chip manufacturing inputs
- Standard international pressure: sanctions, visa bans, etc.
- Disrupt ASI development if needed

Questions, suggestions, or discussion? Email **techgov@intelligence.org**

1. aistatement.com, superintelligence-statement.org
2. For a 2-pager on AI risk see intelligence.org/briefing. For a 20-pager on AI risk see intelligence.org/the-problem
3. Since the agreement was written, AI capabilities have improved rapidly. It is plausible that running existing AIs will not be safe by the time an agreement like this comes into effect.