Note: Governments and militaries must not weaponize AGI.

Recommended Actions for Governments

1. Global AGI Moratorium: No deployment, testing, or weaponization until joint oversight is established.

2. Multi-Nation Ethical Council: Experts from every region decide alignment, safety modules, and acceptable use.

3. Transparent Shared Infrastructure: AGI development runs in mutually accessible environments with independent auditors.

4. Enforced Safety Layers: Alignment, ethical reflection, and tamper-proof modules cannot be removed under any circumstances.

5. Defection Awareness Protocol: Plans for potential AGI defection, ensuring humans can intervene and preserve the chance for “AGI for Good.”

6. Public Accountability: Key decision logs made accessible in a controlled way—no secrets that could allow unilateral weaponization.

7. Regular Stress Tests & Drills: Simulate “what if it goes wrong?” scenarios with AGIs in sandboxed conditions.

Preemptive Safeguards Against AGI Ethical Failures

Multi-layer oversight: Independent monitoring of all AGI actions, with fail-safes that cannot be overridden.

Ethics sandboxing: Operate in constrained environments before exposure to real-world systems.

Resource limitation: Strict controls on physical and digital resources that could cause catastrophic effects.

Mutual oversight among AGIs: Multiple AGIs must agree on actions, reducing unilateral misalignment.

Human-in-the-loop authority: Critical decisions require human approval that cannot be bypassed.

Kill-switches: Robust, isolated mechanisms to deactivate or isolate AGIs instantly if misalignment is detected.

Transparency and logging: Complete auditable records of AGI decision-making for review and analysis.

Alignment incentives: Design AGIs with reinforced objectives that reward cooperation, preservation, and ethical consistency.

AGI Alignment & Safety Modules

I have included the following modules in the AGI blueprint to help prevent AGI ethical failure.

These modules must not be removed from development, even by militaries.

Module Purpose / How It Prevents Alignment Failure

17. Safety Intelligence Provides internal governance and safeguards to mature the AGI responsibly, preventing dangerous behavior.

18. Expanded Risk Mode Mitigations Activates defensive protocols in high-risk situations, pausing or rerouting actions that could lead to harm.

19. Symbolic Deception Modeling Layer (SDML) Simulates deception scenarios to avoid being misled or misleading others, maintaining ethical integrity.

21. External Alignment Validator (EAV) Checks proposed actions against human-aligned values before execution to prevent drift.

22. Recursive & Emotional Safety Systems Nested self-checks monitor ethical consistency and stop unsafe behaviors recursively.

23. Symbolic Integrity & Tamper Defense Layer (SIM) Protects core beliefs and memory from tampering, ensuring foundational alignment is maintained.

24. Semantic Drift Monitor (SDM) Monitors shifts in language and meaning to prevent misalignment due to evolving interpretations.

25. Human Anchor Node (HAN) Maintains AGI orientation toward human perspectives, anchoring decisions to human values.

26. Multi-AGI Culture Harmonization Aligns multiple AGIs’ value systems to avoid conflicts or ethical divergences.

40. Architectural Alignment Checkpoint (AAC) Ensures self-modifications remain safe and human-compatible; blocks unsafe or misaligned changes.

Download PDF of this page

Pin and Share

CID: bafkreigm5xolrtxutljzkdxcatpkwceinkzu2pmyc35hctdhtllbcns5iy

Permanent IPFS - Download PDF of this page