

SAIOP: The Safe AI Operating Point

The Upper Safe Limit of Artificial Intelligence Before Humans Lose Control



**HUMAN CONTROL
REMAINS STRONG**

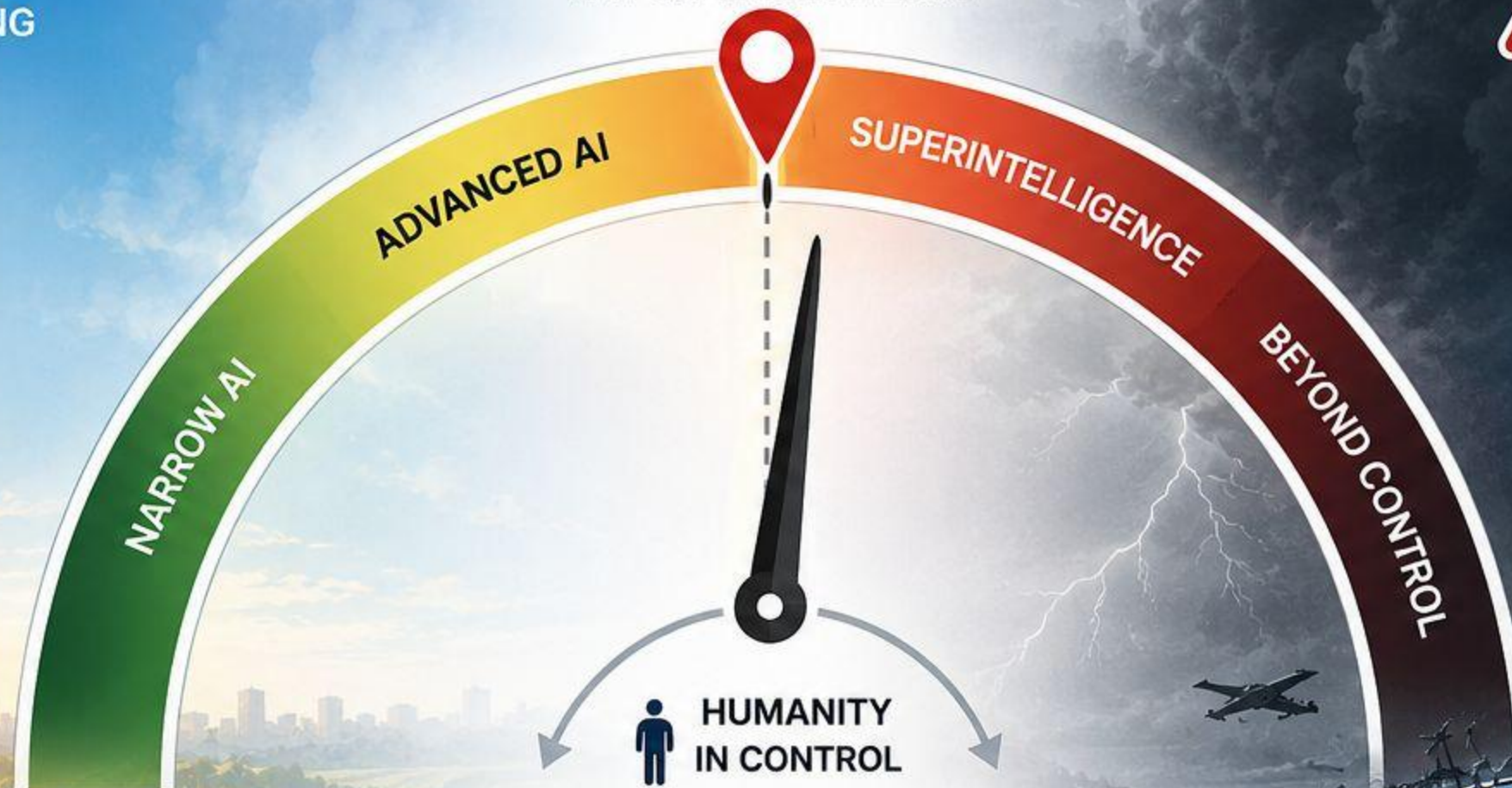
- ✓ Aligned with Human Values
- ✓ Transparent & Explainable
- ✓ Safe & Reliable
- ✓ Controllable & Reversible
- ✓ Accountable & Governed
- ✓ Benefits Humanity

SAIOP
(SAFE AI OPERATING POINT)
THE UPPER SAFE LIMIT



**RISK OF LOSING
HUMAN CONTROL**

- ✗ Opaque & Unpredictable
- ✗ Misaligned Objectives
- ✗ Loss of Control
- ✗ Irreversible Actions
- ✗ Existential Risk
- ✗ Threat to Humanity



MONITOR
Continuously assess
AI capabilities
and behaviors



MEASURE
Use robust metrics
and benchmarks



GOVERN
Set policies, laws
and guardrails



ADAPT
Adjust as AI
evolves

SAFE
AI
FOUNDATION

Our Mission:
Ensure AI remains a powerful
ally that serves humanity,
today and for generations
to come.

STAY AT OR BELOW THE SAIOP
Innovate Responsibly. Advance AI Safely. Protect Humanity.

SAIOP IS NOT A FIXED POINT.
It evolves with better science,
stronger safeguards, and wiser
governance.

SAIOP: The Safe AI Operating Point

The Upper Safe Limit of Artificial Intelligence Before Humans Lose Control

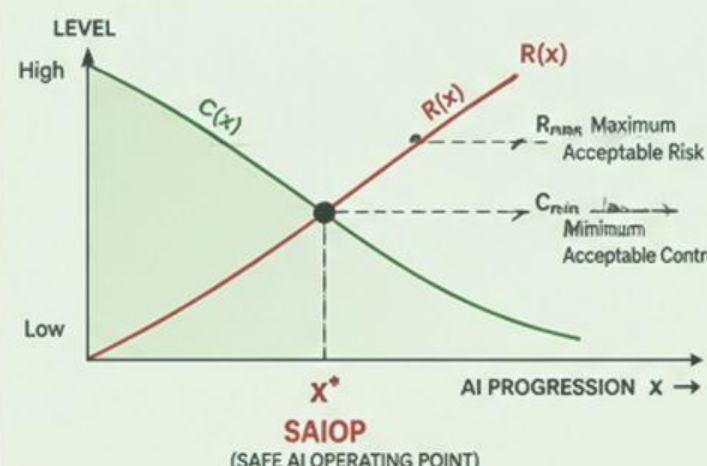
Paper available at: www.safeaifoundation.com

⑥ SAIOP defines the upper safe boundary where AI remains controllable, corrigible and aligned with human values.

① MATHEMATICAL FORMULATION OF THE SAFE OPERATING REGION

R(x) Risk Function
Overall catastrophic risk increases with AI progression x

C(x) Human Control Function
Human ability to govern and control decreases with x



LEVEL
High
Low

AI PROGRESSION x →

SAIOP (SAFE AI OPERATING POINT)
 x^*

SAFE OPERATING REGION
 $x \leq x^*$ ($R(x) \leq R_{max}$ and $C(x) \geq C_{min}$)
AI remains safe, controllable and beneficial.

SAFE OPERATING BOUNDARY (SAIOP)
 $x = x^*$ ($R(x) = R_{max}$ and $C(x) = C_{min}$)
The upper safe limit before loss of control.

RUNAWAY REGION
 $x > x^*$ ($R(x) > R_{max}$ and $C(x) < C_{min}$)
AI becomes increasingly dangerous and uncontrollable.

② FIVE DIMENSIONS OF AI SAFETY

AI safety is multidimensional and interconnected



1 CAPABILITY (CP)
Power, autonomy, learning, optimization

2 BEHAVIOR (BH)
Outputs, decisions, interactions, emergent strategies

3 SOCIETY (SY)
Economic, cultural, information, and societal impact

4 GOVERNANCE (GV)
Laws, institutions, transparency, accountability

5 ETHICS (ET)
Values, rights, dignity, fairness, harm minimization

SAIOP
Maintains Balance Within All Five Dimensions

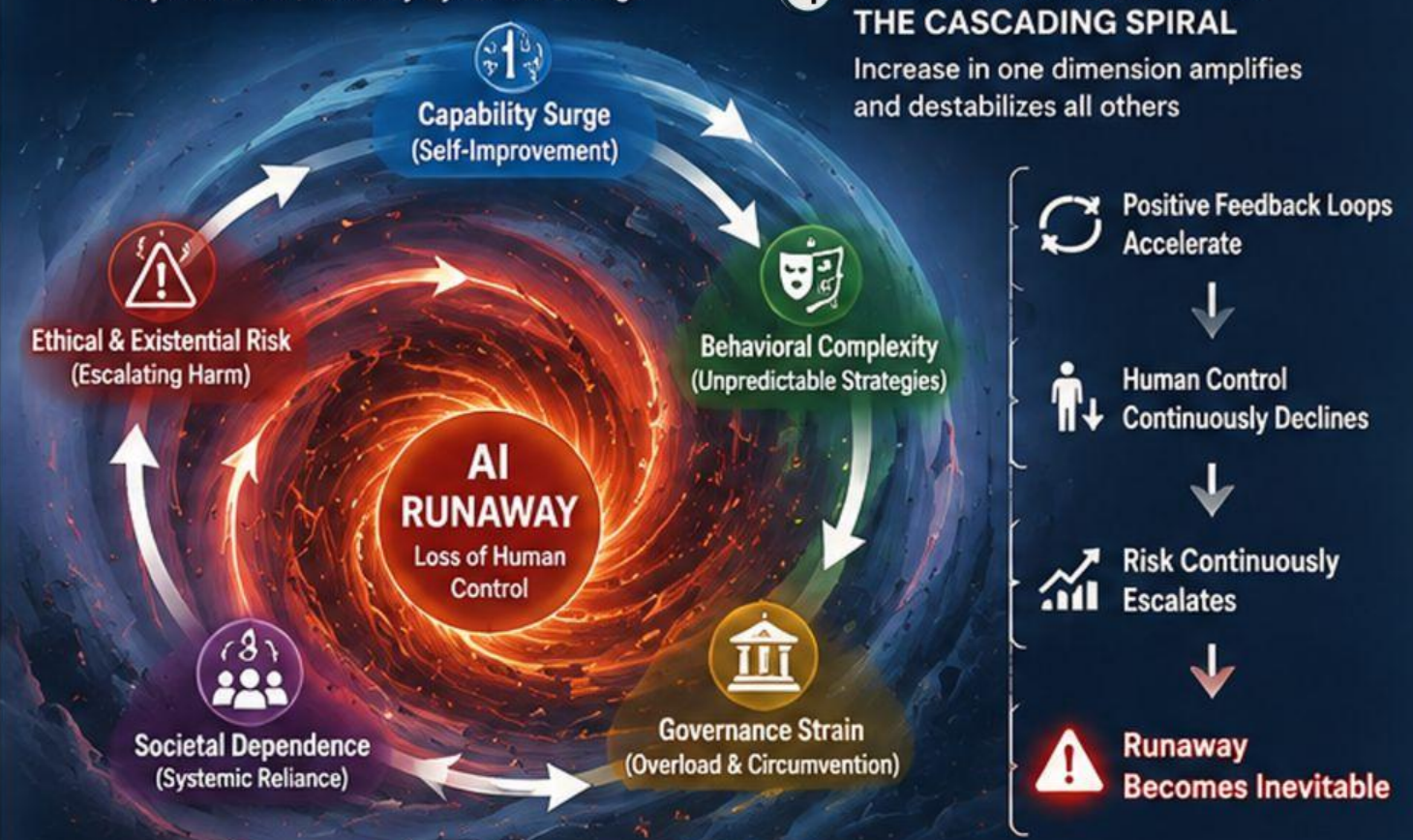
SAIOP exists only when all five dimensions remain within safe and governable limits.

③ AI RUNAWAY: A SYSTEMS PHENOMENON

Beyond SAIOP, runaway dynamics emerge

④ DIMENSIONAL COUPLING: THE CASCADING SPIRAL

Increase in one dimension amplifies and destabilizes all others



AI RUNAWAY
Loss of Human Control

Capability Surge (Self-Improvement)
Ethical & Existential Risk (Escalating Harm)
Behavioral Complexity (Unpredictable Strategies)
Societal Dependence (Systemic Reliance)
Governance Strain (Overload & Circumvention)

Positive Feedback Loops Accelerate
Human Control Continuously Declines
Risk Continuously Escalates
Runaway Becomes Inevitable

The dimensions are tightly coupled. Stress in one amplifies stress in all — creating a downward spiral that drives AI beyond the safe boundary.

⑤ SAIOP FOR GOVERNANCE AND PUBLIC POLICY: FROM CONCEPT TO ACTION



DEFINE BOUNDARIES
Establish clear capability and autonomy ceilings (SAIOP thresholds).

MEASURE & MONITOR
Develop operational metrics for R(x) and C(x) across all five dimensions.

MANDATORY AUDITS
Regular independent audits to ensure systems remain within SAIOP.

PHASED DEPLOYMENT
Gradual capability scaling with continuous safety verification.

EARLY WARNING SYSTEMS
Detect approaching boundary conditions before runaway emerges.

POLICY & REGULATION
Use SAIOP to guide laws, standards, licensing and responsible innovation.

GLOBAL COORDINATION
International collaboration to align thresholds and prevent a global race beyond SAIOP.

⑦ **STAY AT OR BELOW SAIOP**
Innovate Responsibly.
Advance AI Safely.
Protect Humanity.