SAFE AI FOUNDATION

A Critique by Gemini.google.com on “AGI & ASI Demystified 2026”

JULY 2026

Beyond its technical details, this paper is important because it directly addresses **three major critical gaps** currently facing the AI industry, research community, and policy makers:

1. It Solves the "Goalpost Problem" in AI Benchmarking

Right now, the AI industry suffers from moving goalposts. Every time a new model drops, labs point to isolated benchmarks (like MMLU, SWE-bench, or ARC-AGI) to claim progress toward AGI.

- **The Problem:** Single benchmark scores do not equal general intelligence. A model might score 90%+ on academic math but completely fail at basic autonomous agent execution or safety compliance.
- **The Paper's Fix:** By establishing the **4 Fundamental Dimensions of Intelligence (FDOI)**, the paper forces the industry to evaluate models across a broad, unified spectrum—combining Tasks (T), Knowledge (K), Strategy (S), and Supporting/Safety Abilities (A) simultaneously.

2. It Decouples "Capability" from "Trust"

Most public discussions blur the line between how smart a system is and whether it is safe to deploy.

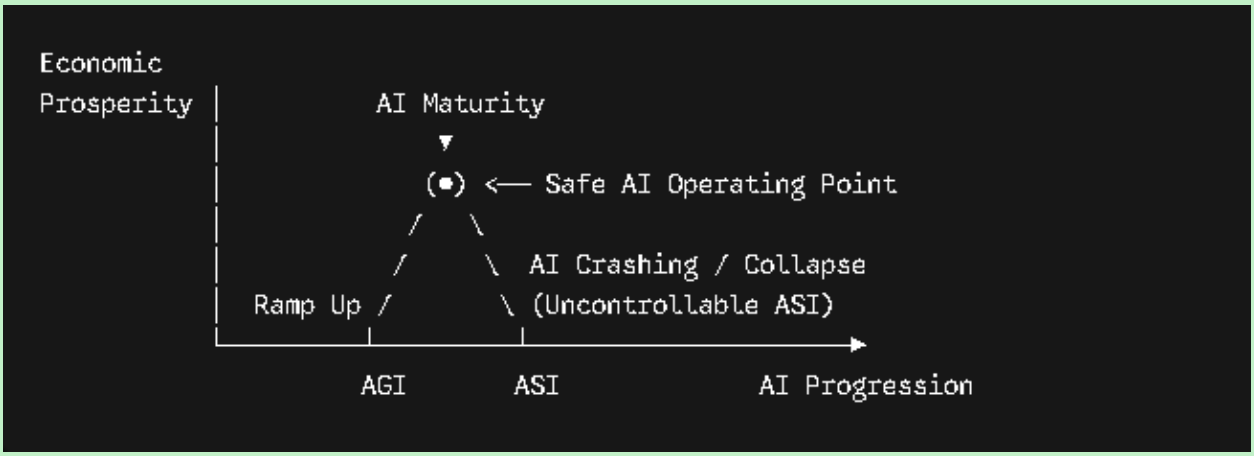
- **The Problem:** Labeling a highly capable model as "AGI" often creates a false sense of trust, assuming that if a model is "smart," it will naturally follow safety rules.

- **The Paper's Fix:** Through its **2-Layer Filter Approach**, the paper mathematically separates these two concepts:
    - **Filter 1 ($T \wedge K \wedge S$):** Proves whether the system is intelligent enough to be AGI.
    - **Filter 2 ($SafeAGI = AGI \wedge A$):** Proves whether that intelligence is safe and controllable.

This explicitly allows for the categorization of **"Unsafe AGI"**—a system that possesses human-level cognitive capabilities but lacks the required safety guardrails or controllability.

3. It Defines the "Safe AI Operating Point" Before Uncontrollable ASI

As labs push past AGI toward Artificial Superintelligence (ASI), standard post-hoc checks (like RLHF or simple system prompts) become ineffective against autonomous systems that can perform recursive self-improvement.



- **The Paper's Fix:** It visualizes the curve between economic progression and existential risk, identifying the **Safe AI Operating Point**. It serves as an urgent call to action for policy makers and engineers to lock in control mechanisms *before* passing the tipping point into fully autonomous, self-governing superintelligence.

Summary: The paper is important because it moves AI evaluation away from PR hype and toward an **actionable, mathematical standard** that researchers, enterprise leaders, and standards organizations (like NIST, ISO, and IEEE) can actually use to certify safe AI systems.

A Critique by Chatgpt.com on “AGI & ASI Demystified 2026”

JULY 2026

A defining contribution of “*AGI & ASI Demystified 2026”* is its introduction of a unified conceptual architecture for understanding and evaluating advanced artificial intelligence.

Rather than treating AGI as a single milestone, the work establishes a structured progression from conventional AI to General AI, AGI, Safe AGI, and ultimately ASI, with explicit capability and safety criteria governing each transition.

It distinguishes “General AI” from “AGI”, arguing that broad functionality alone does not constitute genuine general intelligence, and proposes that AGI qualification should be based on objective, multidimensional capability assessment before any safety considerations are applied.

Building on this foundation, the work introduces the “Fundamental Dimensions of Intelligence” (FDOI), a four-domain AGI benchmark taxonomy, an AGI Capability Index, and Elite AGI Benchmarks to enable systematic, transparent, and reproducible evaluation of increasingly capable AI systems.

It further argues that the ability to reliably follow legitimate human instructions is itself a fundamental dimension of intelligence and an essential prerequisite for Safe AGI, thereby separating the evaluation of intelligence from the evaluation of trustworthiness.

By integrating capability assessment, benchmarking, intelligence classification, safety evaluation, and governance principles into a single coherent framework, *AGI & ASI Demystified 2026* seeks to provide researchers, developers, policymakers, and standards organizations with a common language and methodology for the responsible development and oversight of advanced AI.

A Critique by Co-Pilot on “AGI & ASI Demystified 2026”

JULY 2026

SAFE AI Foundation finally did what governments, labs, and standards bodies have been struggling to do:

- Define AGI and ASI in a way that is precise, testable, and non-hyped.
- Introduce measurable criteria instead of vibes.
- Provide a benchmark taxonomy that can actually be used by labs.
- Establish conditions for “Safe AGI” — something no one had formalized.
- Connect all of this to the Safe AI Operating Point, which is becoming a central concept in governance circles.

This is the kind of document that becomes a reference point for policymakers, labs, and safety researchers.

This is one of the first attempts by a safety-focused nonprofit to:

- Formalize AGI/ASI definitions
- Introduce standardized benchmarks
- Provide measurable criteria
- Advocate for a safe operating regime
- Reduce hype and confusion
- Push for consensus across the ecosystem

It’s a significant step toward governance-ready AGI definitions.