AGI & ASI Demystified 2026:

The Need for Clarity, Justifications, Consensus, & Fundamental Dimensions of Intelligence

Chai K Toh and Vasu Raj Jain

Safe AI Foundation

21 JULY 2026

Until today, there is no single and widely accepted definition of Artificial General Intelligence (AGI) and Artificial Super-Intelligence (ASI). There is also a lack of clarity of these terms. There are no standards, reference models or reference specifications for AGI and ASI. Different researchers, companies, and organizations have made different definitions of AGI and ASI, and they have emphasized different capabilities - from matching human performance on most tasks to being able to autonomously learn, adapt, execute, and solve novel problems.

In this article, The SAFE AI Foundation takes a step further to clarify these technological terms and highlight what is needed to accurately clarify, justify, qualify and quantify AGI and ASI.

1 Artificial General Intelligence (AGI)

First, we reveal who has said what about AGI. Below are a consolidation and comparison of influential definitions of AGI.

Source	Definition of AGI	Emphasis
OpenAI	AGI refers broadly to highly autonomous systems that outperform humans at most economically valuable work. (This was the organization's earlier public framing.)	Economic usefulness across many domains
Google DeepMind	Proposes viewing AGI as a spectrum. AGI is AI that can perform at least at human level across a wide range of cognitive tasks, with levels ranging from emerging to superhuman.	Generality plus measurable capability levels
Microsoft researchers	Often describe AGI as a system capable of learning and performing essentially any intellectual task a human can perform.	Human-level breadth of cognition
Shane Legg: co-founder of Google Deepmind	Intelligence is an agent's ability to achieve goals in a wide range of environments. AGI would	General problem-solving ability

Source	Definition of AGI	Emphasis
	therefore excel across diverse environments rather than a narrow task.	
François Chollet – creator of Keras	General intelligence is efficient skill acquisition and generalization to novel problems—not simply memorizing or scaling existing capabilities.	Generalization to new tasks
Yann LeCun – Turing award winner	Argues current systems are far from AGI. Future AGI will require world models, reasoning, planning, and understanding of the physical world.	Rich world understanding and reasoning
Demis Hassabis – Nobel prize winner	Often describes AGI as a system exhibiting the cognitive abilities humans possess across essentially all domains, with robust reasoning, planning, memory, and creativity.	Broad human-like cognitive competence
Nick Bostrom, author NY Times best seller	Uses "general intelligence" to mean the ability to perform virtually any intellectual task that a human can perform.	Human-equivalent intellectual scope
Association for the Advancement of Artificial Intelligence, AAAI (various workshops)	Rather than endorsing one definition, AGI is generally discussed as AI exhibiting competence across diverse tasks, transfer learning, reasoning, and autonomy.	Community consensus remains unsettled

Most AGI definitions fall into one of these categories below but these have not yet been universally agreed on:

1. **Human-level Performance** – It can perform essentially any cognitive task a typical human can. The focus is on the breadth of abilities.
2. **Generalization** – It can solve previously unseen problems with little or no task-specific training, with the focus on adapting rather than memorizing.

3. **Learning Efficiency** – It can learn new domains from relatively little data and experience with a focus on sample-efficient learning.
4. **Autonomy** – It can independently plan, learn, use tools, and pursue long-term goals, with a focus on self-directed behavior.
5. **Economic Capability** – It can do most economically valuable work currently done by humans, with a focus on labor substitution and productivity.
6. **Theoretical Intelligence** – It can achieve goals across a very wide range of environments, providing abstract intelligence rather than human imitation.

Yann LeCun has provided a more specific view, i.e., he cited the inclusion of world models and the ability to understand the physical world. The rest of the definitions generally focused on ability to do what humans today could do.

1.1 A Recent Trend on AGI

One notable shift in the past few years is the move towards **operational definitions** — definitions that can be tested. For example, researchers at Google DeepMind proposed measuring AGI along two dimensions:

- **Breadth**: How many different kinds of tasks the system can perform.
- **Depth (capability level)**: How well it performs on those tasks, ranging from novice to superhuman.

This moves the discussion away from asking "Is this AGI?" toward asking "How general is this system, and how capable is it?" However, this move is still not fully quantitative. There are no requirements for specific numbers to be fulfilled and what specific capabilities AGI must have. There is also an increasing attention on the risks of AGI. These risks are currently being addressed by AI industries and policy makers.

1.2 The Difficulty of a Consensus on AGI

Although definitions of AGI can vary from different persons and organizations, several common broad criteria exist:

(a) Competence across many domains rather than one specialized task

(b) Ability to learn and adapt to new situations

(c) Generalization beyond the training data

(d) Reasoning, planning, and problem solving

(e) Minimal need for human-designed and task-specific programming

The commonalities help us sharpen our understanding of AGI. However, current difficulties lie on the lack of a formal quantitative assessment and benchmark that says when an AI can do these XX number of tasks (a predefined list) and perform at this YY level (above say 90% of something), then it is considered an AGI! So far, no authority has come up with this benchmark or criteria (not ISO, IEC, AAAI, OECD, etc.,).

Furthermore, the largest disagreements are over where to draw the threshold. Some define AGI as matching human capabilities across almost ALL cognitive tasks (yet no one has provided a list of all the required cognitive tasks), while others define it more broadly as robust general intelligence — even if it does not resemble human cognition in every respect. There have been comparisons and benchmarking of various LLM models but those metrics are not standards for AGI, they are “de facto”. The broad definition does not help as it clearly does not draw the line of distinction between AGI and non-AGI. Hence, the lack of clarity and details prohibit the AI community from reaching a consensus on:

(a) what needs to be done, and

(b) what criteria needs to be fulfilled

before an AI product is considered having reach the level of AGI.

2 ▒ Today’s Evaluation of AGI

Current elite benchmarks that researchers actually use to evaluate work that are making progress toward AGI comprise well over 40 serious evaluations. They cover reasoning, science, coding, agents, multimodality, long-context, truthfulness, robotics, and real-world task execution. The population of these metrics evolved out of the need to quantity capabilities and performance of AGI.

Below is a curated list.

No.	Benchmark	Measures
1	ARC-AGI (ARC Prize)	Abstract reasoning and fluid intelligence

No.	Benchmark	Measures
2	ARC-AGI-2	Harder successor to ARC
3	Humanity's Last Exam (HLE)	Extremely difficult cross-disciplinary expert problems
4	GPQA Diamond	Graduate-level science reasoning
5	MMLU	General academic knowledge
6	MMLU-Pro	Harder reasoning-focused MMLU
7	BIG-Bench Hard (BBH)	Multi-step reasoning
8	LiveBench	Fresh continuously updated reasoning benchmark
9	SimpleQA	Factual correctness
10	IFEval	Instruction following
11	BrowseComp	Web browsing and information retrieval
12	GDPval	General decision making
13	Arena-Hard	Difficult preference-based evaluation
14	Arena Elo	Human preference ranking
15	MMMU	Expert multimodal reasoning
16	MMMU-Pro	Advanced multimodal reasoning
17	MMT-Bench	Comprehensive multimodal intelligence
18	SWE-bench Verified	Real GitHub bug fixing
19	SWE-bench Pro	Harder software engineering
20	LiveCodeBench	Fresh programming contests
21	HumanEval	Code generation
22	MBPP+	Python programming
23	Aider Polyglot	Code editing
24	Terminal-Bench	Terminal agent tasks
25	Terminal-Bench-2	Complex CLI agents
26	CyberGym	Cybersecurity exploitation tasks
27	MATH-500	Mathematical reasoning
28	AIME 2024	Competition mathematics
29	AIME 2025	Hard competition mathematics
30	FrontierMath	Research-level mathematics
31	GSM8K	Grade-school math reasoning
32	MGSM	Multilingual math
33	OSWorld	Computer-use agents
34	TAU-Bench	Enterprise agent workflows
35	HellaSwag	Commonsense reasoning
36	TruthfulQA	Truthfulness and hallucination resistance
37	Winogrande	Commonsense reasoning
38	PIQA	Physical commonsense
39	DROP	Reading comprehension with reasoning
40	ToolBench	Tool use and API calling
41	AgentBench	Autonomous agent performance
42	GAIA	Real-world assistant tasks

3 ▒ Proposed 4-Domain AGI Benchmark Taxonomy

Rather than organizing benchmarks by discipline (math, coding, multimodal, etc.), the SAFE AI Foundation proposes a taxonomy that groups AGI evaluations according to the "Fundamental Dimensions of Intelligence" (FDOI) they assess, such as:

1. **Human Tasks** – Can the system perform real-world work that humans perform?
2. **Human Knowledge** – What breadth and depth of knowledge does the system possess?
3. **Strategy, Planning & General Intelligence** – Can the system reason, plan, solve novel problems, and make decisions?
4. **Supporting Abilities** – Does the system reliably exhibit complementary capabilities such as instruction following, coding, truthfulness, commonsense, and multi-modal understanding?

In fact, this is a more meaningful way to think about AGI evaluation than simply listing benchmarks. Organizing them according to the type of intelligence they measure, rather than their historical origin or treating them as silos. In this way, we have:

Category	What it Measures	Representative Benchmarks
A. Human Tasks (Can AI perform work like a human?)	Completing practical tasks people perform in jobs and daily life	SWE-bench Verified, SWE-bench Pro, Terminal-Bench, Terminal-Bench-2, OSWorld, TAU-Bench, AgentBench, BrowseComp, ToolBench, CyberGym
B. Human Knowledge (What does AI know?)	Breadth and depth of knowledge across academic and professional domains	MMLU, MMLU-Pro, GPQA Diamond, Humanity's Last Exam, MMMU, MMMU-Pro, MMT-Bench, SimpleQA, TruthfulQA, DROP
C. Strategy, Planning & General Intelligence (Can AI think?)	Reasoning, abstraction, planning, multi-step problem solving, decision making	ARC-AGI, ARC-AGI-2, BIG-Bench Hard, LiveBench, GDPval, Arena-Hard, FrontierMath, AIME 2024, AIME 2025, MATH-500, GSM8K, MGSM
D. Other Cognitive Abilities	Capabilities that complement intelligence but don't fit neatly into the other three groups	IFEval, HumanEval, MBPP+, LiveCodeBench, Aider Polyglot, HellaSwag, PIQA, Winogrande, Arena Elo

For example, we can set the criteria that if an AI fulfills at least A, B, and C fundamental dimensions of intelligence, one can safely classifies it as AGI. This is the sort of "cutoff" that we currently do not have. Mathematically, we can express this in several ways:

Case 1: Boolean Definition of AGI (Simple and Elegant)

Let

- T= Human Tasks
- K= Human Knowledge
- S= Strategy, Planning & General Intelligence (GI)
- A= Supporting Abilities

Define

$$T, K, S, A \in \{0, 1\}$$

where

- 1 = satisfies the domain
- 0 = does not satisfy the domain

Then

$$\boxed{\text{AGI} = T \wedge K \wedge S}$$

Supporting abilities are considered enabling rather than defining, so

$$A \notin \text{minimum AGI criterion}$$

This means

AGI exists if and only if: $T = 1, K = 1, S = 1$

"Safe AGI" exists if and only if: $T = 1, K = 1, S = 1, A = 1$, *i.e.*,

$$\boxed{SafeAGI = AGI \wedge A}$$

Where A is the dominating domain on "following legitimate human instructions without causing harm, adheres to safety constraints, and remains controllable". For Safe AGI, we need to emphasis "controllable AI". An uncontrollable AI is unsafe AI.

Our 2-layer filter approach has a clear and logical structure. We are essentially separating two questions that are often mixed together:

1. Does the AI have the intelligence capability to qualify as AGI?
2. Can that AGI be trusted and safe to use?

Below gives a clean sequential evaluation model:

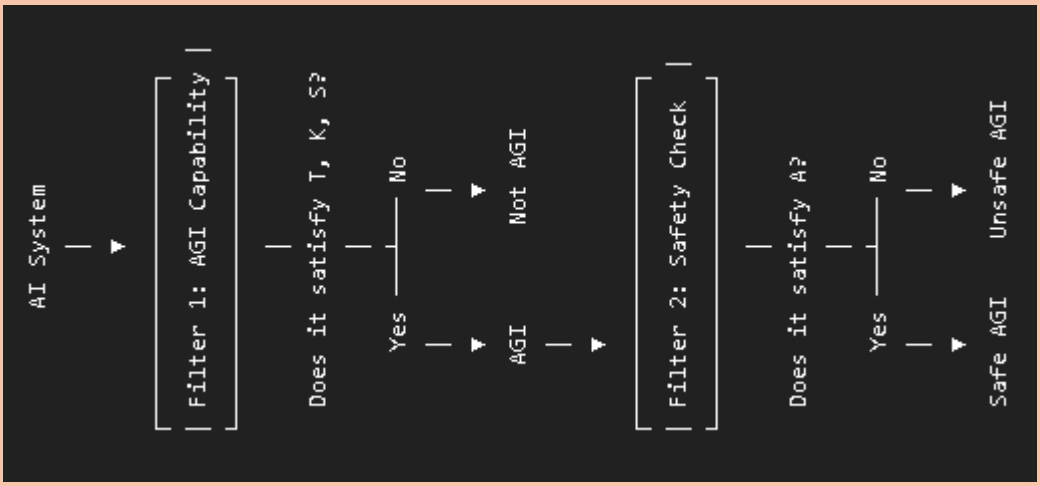


Fig. 1: How we classify AGI versus Non-AGI, Safe AGI versus Unsafe AGI.

The elegance of our model is the asymmetry:

- T,K,S define whether the system is intelligent enough to be AGI.
- 'A' determines whether that intelligence is safe enough to be trusted and used.

In other words:

$$Intelligence \neq Safety$$

An AI system can therefore be:

T	K	S	A	Classification
1	1	1	1	Safe AGI
1	1	1	0	AGI but not Safe AGI
1	1	0	1	Not AGI
1	0	1	1	Not AGI
0	1	1	1	Not AGI

Case 2: Alternatively, we can formulate a threshold equation:

Suppose every domain has a score

$$T, K, S, A \in [0, 100]$$

Choose an AGI threshold

$$\tau$$

for example

$$\tau = 90$$

Then

$$\boxed{\text{AGI} = \begin{cases} 1, & T \geq \tau,\ K \geq \tau,\ S \geq \tau \\ 0, & \text{otherwise} \end{cases}}$$

Case 3: Introducing an AGI Capability Function

Define

$$C = (T, K, S, A)$$

where

- T= Human Tasks score
- K= Human Knowledge score
- S= Strategy, Planning and GI score
- A= Supporting score

Then define

$$\boxed{\text{AGI} = \mathbb{I}(\min(T, K, S) \geq \tau)}$$

Where, $\mathbb{I}$ is the indicator function. This means that AGI exists if the weakest of the three fundamental dimensions exceeds the threshold.

Case 4: Proposed AGI Criterion by SAFE AI Foundation

Let

$$\boldsymbol{T, K, S} \in [\mathbf{0, 1}]$$

Then

$$\boxed{\mathbf{AGI} = (\boldsymbol{T} \geq \boldsymbol{\theta}) \wedge (\boldsymbol{K} \geq \boldsymbol{\theta}) \wedge (\boldsymbol{S} \geq \boldsymbol{\theta})}$$

Where $\boldsymbol{\theta}$ is the minimum human-equivalent capability. This defines whether a system qualifies as AGI. $\boldsymbol{\theta}$ is AGI capability threshold. Now let ϕ be the Safe AGI threshold, then:

Safe AGI is:

$$\boxed{SafeAGI = AGI \wedge (A \geq \phi)}$$

Case 5: Proposed AGI Capability Index (ACI) by SAFE AI Foundation

Finally, to facilitate the ranking of AGI system by overall capability, we introduce having an AGI index. We define

$$\boxed{\mathrm{ACI} = \frac{w_T T + w_K K + w_S S + w_A A}{w_T + w_K + w_S + w_A}}$$

where

$$w_T, w_K, w_S, w_A$$

are weights. Note that ACI does not define AGI. It only measures how capable is the model. In summary, a system is classified as AGI if it meets or exceeds the required threshold in the three fundamental dimensions of intelligence, i.e., Human Tasks, Human Knowledge, and Strategy–Planning–General Intelligence. To the best of our knowledge and investigations, there is no widely adopted AGI index that matches what the SAFE AI Foundation is proposing here. Our proposal connects 3 ideas: (a) a taxonomy of AGI benchmark domains, (b) an AGI Classification Criterion, and (c) an AGI Capability Index that aggregates benchmark results within those domains.

For further illustration, we also provide comparisons of four leading frontier models that are widely recognized as state-of-the-art in mid-2026:

1. ChatGPT (GPT-5.5)
2. Claude Opus 4.1
3. Gemini 2.5 Pro
4. Grok 4

Four-Domain AGI Benchmark Taxonomy Ranking

Domain	1st	2nd	3rd	4th	Primary Evidence
A. Human Tasks	Claude Opus 4.1	GPT-5.5	Gemini 2.5 Pro	Grok 4	SWE-bench, TAU-Bench, OSWorld, Terminal-Bench
B. Human Knowledge	Gemini 2.5 Pro	GPT-5.5	Claude Opus 4.1	Grok 4	MMLU-Pro, GPQA, Humanity's Last Exam, MMMU
C. Strategy, Planning & General Intelligence	GPT-5.5	Claude Opus 4.1	Gemini 2.5 Pro	Grok 4	ARC-AGI, ARC-AGI-2, FrontierMath, LiveBench, BIG-Bench Hard
D. Supporting Abilities	GPT-5.5	Claude Opus 4.1	Gemini 2.5 Pro	Grok 4	IFEval, TruthfulQA, HumanEval, MBPP, HellaSwag

Almost all existing frontier models are covered by our proposed 4 domains, including Kimi K2, Llama 4, Qwen 3 and DeepSeek R1. Today, benchmark discussions are fragmented and people often know the names of benchmark metrics but not what they represent. Our proposed taxonomy organizes the rapidly growing ecosystem of AGI benchmarks into 4 Fundamental Dimensions of Intelligence (FDOI) that people can quickly understand. Most benchmark reports answer:

"How good is Model X on Benchmark Y?"

Our taxonomy answers a much more deep and profound question:

"What dimension of intelligence does Model X actually have?"

That is a fundamentally different contribution made by us.

4 ▒ Artificial Super-Intelligence (ASI)

Similar to AGI, Artificial Superintelligence (ASI) has no universally accepted definition to date. There is, however, some agreements on the core idea:

> "ASI is an AI system that substantially EXCEEDS the best human minds across essentially all cognitive domains."

The differences lie on how much ASI exceeds humans' intelligence, what capabilities matter most, and whether it includes autonomy (agency). Clearly, based on the above-mentioned criteria, ASI is superior than AGI. Below is a list of influential definitions.

Who	Definition / Criteria	Emphasis
Nick Bostrom	Superintelligence is "any intellect that greatly EXCEEDS the cognitive performance of humans in virtually all domains of interest."	Broad, overwhelming superiority across nearly all intellectual tasks
Shane Legg	While better known for defining intelligence generally, Legg's framework implies superintelligence is an agent with capability far BEYOND humans across diverse environments.	General intelligence extended beyond human level
OpenAI	Uses "superintelligence" to describe AI systems that vastly OUTPERFORM humans at most economically valuable and intellectual work, with transformative impacts on science, engineering, and society.	AI exceeding humans with major societal consequences
Google DeepMind	Researchers describe ASI as the highest end of a capability spectrum: systems performing well BEYOND expert humans across almost all cognitive tasks.	Superhuman capability across domains

Who	Definition / Criteria	Emphasis
I. J. Good – British Statistician, cryptologist	Proposed the idea of an "ultra-intelligent machine" that could OUTPERFORM humans in intellectual activities and design even better machines, leading to an "intelligence explosion."	Recursive self-improvement
Ray Kurzweil	Envisions superintelligence emerging after human-level AI, rapidly SURPASSING biological intelligence by orders of magnitude.	Exponential growth beyond human intelligence.
Eliezer Yudkowsky	Describes superintelligence as any mind FAR smarter than humans, regardless of how it is implemented, emphasizing that it may think in fundamentally non-human ways.	Intelligence independent of human-like cognition.
Demis Hassabis	Frequently describes future superintelligence as AI that dramatically EXCEEDS human scientific and reasoning capabilities, enabling breakthroughs across many fields.	Scientific discovery and reasoning at superhuman levels.
Sam Altman	Has described superintelligence as AI that accelerates scientific discovery, innovation, and productivity BEYOND what humanity can achieve unaided.	Transformative acceleration of progress.

As seen, most of the above-mentioned definitions emphasize one or more of these ideas:

1. **Capability Superiority** – ASI EXCEEDS the best human experts at virtually every intellectual task. This is the most commonly regarded definition.
2. **Recursive Self-Improvement** – In ASI, the AI can improve on its own design, potentially leading to a rapid increase in self-intelligence. This reflects the “intelligence explosion” phenomena, first introduced by Dr. I. J. Good.
3. **Scientific and Technological Acceleration** – In ASI, it is expected that AI drives discoveries and inventions at a pace impossible for humans to accomplish. This is commonly stated in discussions by leaders of AI labs.

4. **Strategic Superiority** – In ASI, AI excels not only at technical tasks level but also at planning, forecasting, negotiation, and long-term decision-making. This feature is not yet widely discussed nor agreed on and is not found in AGI.
5. **Economic Transformation** – In ASI, AI performs nearly all cognitive work more effectively than humans, reshaping economies and industries. To some aspects, AGI has already brought some impact on our economy, especially on drawing in huge investments for startups and new data centers.

Nearly all current descriptions of ASI include: (a) a performance beyond the best human experts in almost every cognitive domain, (b) rapid learning and transfer across new problems, (c) exceptional reasoning, planning, and creative ability, (d) ability to solve scientific, engineering, and mathematical problems beyond current human's capability, and (e) potential to automate or augment nearly all knowledge work. The bottom line is that ASI is capable of EXCEEDING the best human minds across essentially all cognitive domains. Furthermore, it is unknown what is the prime motivation behind having an AI that exceeds human abilities and there is little mention on how we are going to control this AI once it exceeds our abilities. Will it pre-empt what we intend to do to it (ASI)? If so, we cannot control it nor stop it.

The risk associated with an uncontrollable AI is very high and can be catastrophic, as any uncontrollable entity with ability to access data, modify code, pre-empt human intentions, and execute tasks autonomously can cause chaos and harm, especially when AI is not behaving correctly or not conforming to existing laws and order.

"Fully autonomous and uncontrollable AI is a hazard."

Once AI has the authority to set its own goals and agenda, it can no longer need to adhere to humans' orders.

"What AI wants to do may not be what humans want and the clash begins."

While ASI is widely accepted as exceeding human's cognitive abilities across a range of tasks, researchers and organizations still differ on several points:

1. **How much better than humans?** Some require a modest but consistent advantage; others envision intelligence in the orders of magnitude greater.

2. **ASI's Autonomy?** Some definitions assume independent goal pursuit, while others focus purely on cognitive capability. The latter still allows human to control over what ASI can or cannot do. The former, however, is totally self-driven.
3. **Is ASI Self-Improvement Essential?** Scientists such as Dr. I. J. Good see it as a central functionality, whereas others treat it as a possible consequence rather than a defining feature.
4. **Must ASI stay General?** Most definitions assume ASI is built on top of general intelligence (i.e., AGI), but a few discussions have allowed for systems to become "superhuman" across all relevant domains without necessarily resembling human cognition.

All these reveal we are nowhere near having total consensus and clarity on the details in defining ASI. The SAFE AI Foundation therefore urges the AI community to convene and come out with a detailed specification on the requirements and capabilities for ASI.

5 Why Consensus on ASI & AGI?

Without agreements on the functionality, capability and performance expected from ASI and AGI systems, AI industries from various countries will produce their own versions of AGI and ASI with different specifications. This means that they are not equal, compatible, nor interoperable in the technical aspects. While "terms" are just words, AGI and ASI are categories of AI developments, capabilities, and progress. They are also a form of "expectations". Having universal agreements mean no surprises, common understanding, and expected capabilities and performance.

6 Justifications for ASI & AGI

Currently, there are no formal agreements, literatures and justifications for the need of AGI and ASI. Perhaps, researchers and academics thought that progress in AI is achieved by introducing better and more capable versions of AI, primarily characterized by their

capabilities and performance. However, the needs for AGI and ASI are unclear and not yet fully justified.

For the cellular world, we have seen how mobile technology has moved from 3G, 4G, to the current 5G standard. Each technology type has better performance and capabilities. However, the mobile world is governed by major standardization bodies, such as ITU, 3GPP, etc. These technology types were widely discussed, specified, formalized, and tested before they became acceptable standards that are universally used and deployed around the globe today.

In the same vein, we need to provide the justifications as to why we need AGI and ASI. For AGI, the justifications have been clear: (a) work automation, (b) the ability to process huge amount of data, (c) useful GenAI applications, and (d) good predictions or output. This has been achieved with existing LLMs, GenAIs, Agents, and several AI Tools.

What remain incomplete is that NOT ALL humans' tasks can be performed by existing AGI. Hence, this is the reason why not many researchers and industry chiefs have declared that we have now fully achieved AGI. This, perhaps, is the reason why Demis Hassabis mentioned that we are still a few years from reaching true AGI.

As for the need of ASI, the frequently heard rationale is for drug discovery, scientific advancement, etc. While these are desirable goals, one cannot for certain confirm that ASI is capable of doing these tasks. It is hard to predict an upcoming technology (to be improved or created) and to prematurely proclaim its capabilities. There are also no proclamations from CEOs and/or robot scientists that we must use ASI to power future humanoid robots. Rather, current humanoid robots use existing AGI technologies. Overall, we are still unsure at this point if the current AI model or structure is capable of attaining ASI. This is a fact.

Frontier models are not ASI. They are experimental models of various architectural designs (and they need not be transformer based). No one has mentioned which specific AI model is suitable for attaining ASI. There are no specifications and standards for ASI. Hence, in summary, there are a lot of speculations on what ASI can and cannot do. The more we expect from ASI, the more we may be disappointed when things have deviated from expectations.

In conclusion, the authors urge the community to continue to seek for clarity, specifications and justifications for accurate understanding and consensus on AGI and ASI. Since this involves the global AI community, it may require international cooperation in this matter, coordinated by one or more governing bodies.

7 ▒ Safe AI Operating Point

As we head towards accomplishing AGI with some hiccups (legal cases, accidents, and mishaps), ASI could be something far more unexpected and worse if done carelessly.

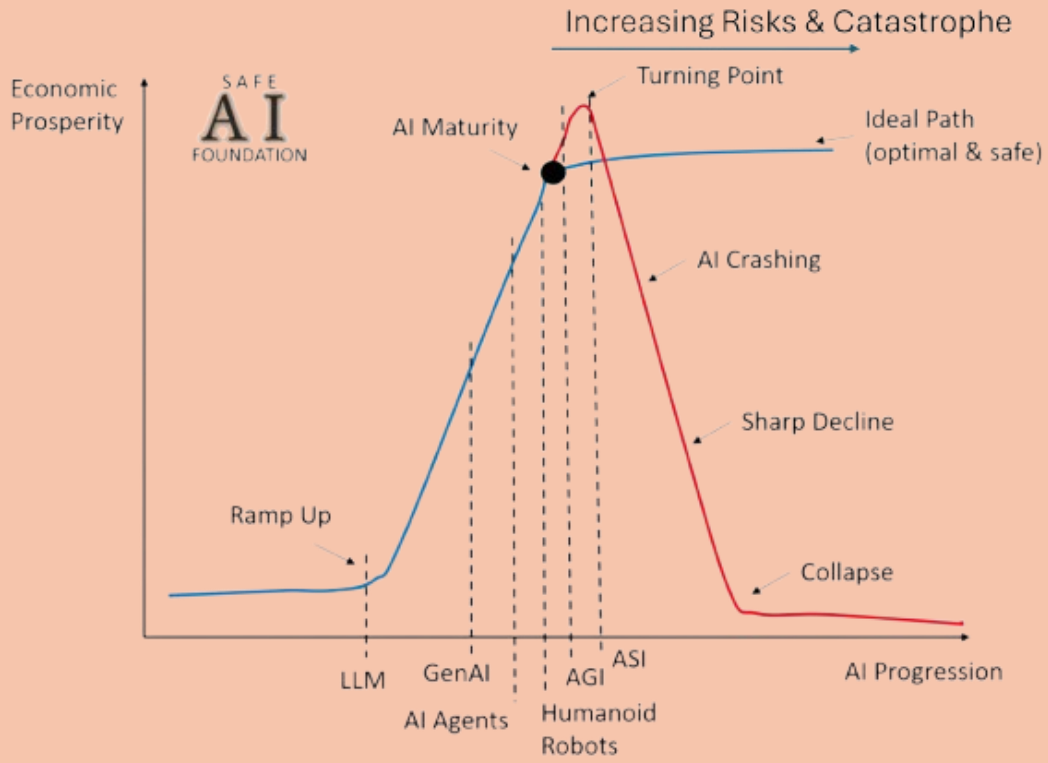


Fig. 2: Society Risks versus AI Progression – Safe Operating Point (black dot) prior to collapse.

In our earlier "AI Manifestation Unfolded" doctrine, we revealed the need to strive for a safe and optimal operating point, something that we can control and contain the risks, before a catastrophic event or existential risk can occur, while concurrently enjoying economic prosperity brought by AI. See figure 2.

The SAFE AI Foundation is looking at ASI as a possible tipping, peaking, and breaking point, where once ASI exceeds human intelligence, possesses full autonomy and is self-governing and self-improving, humans may lose control of it, resulting in a catastrophe. The difficulty will be for AI leaders, country leaders, researchers and academics to accurately pinpoint where is the safe operating point before disaster occurs. This demands further research, care, and investigations.

\\\ end ///

AUTHORS

- Chai K Toh is the main and corresponding author for this article and he holds a PhD in Computer Science from Cambridge University, UK.

- Vasu Raj Jain holds a Master's Degree in Computer Science from University of Central Florida, USA and he works for Amazon.

REFERENCES

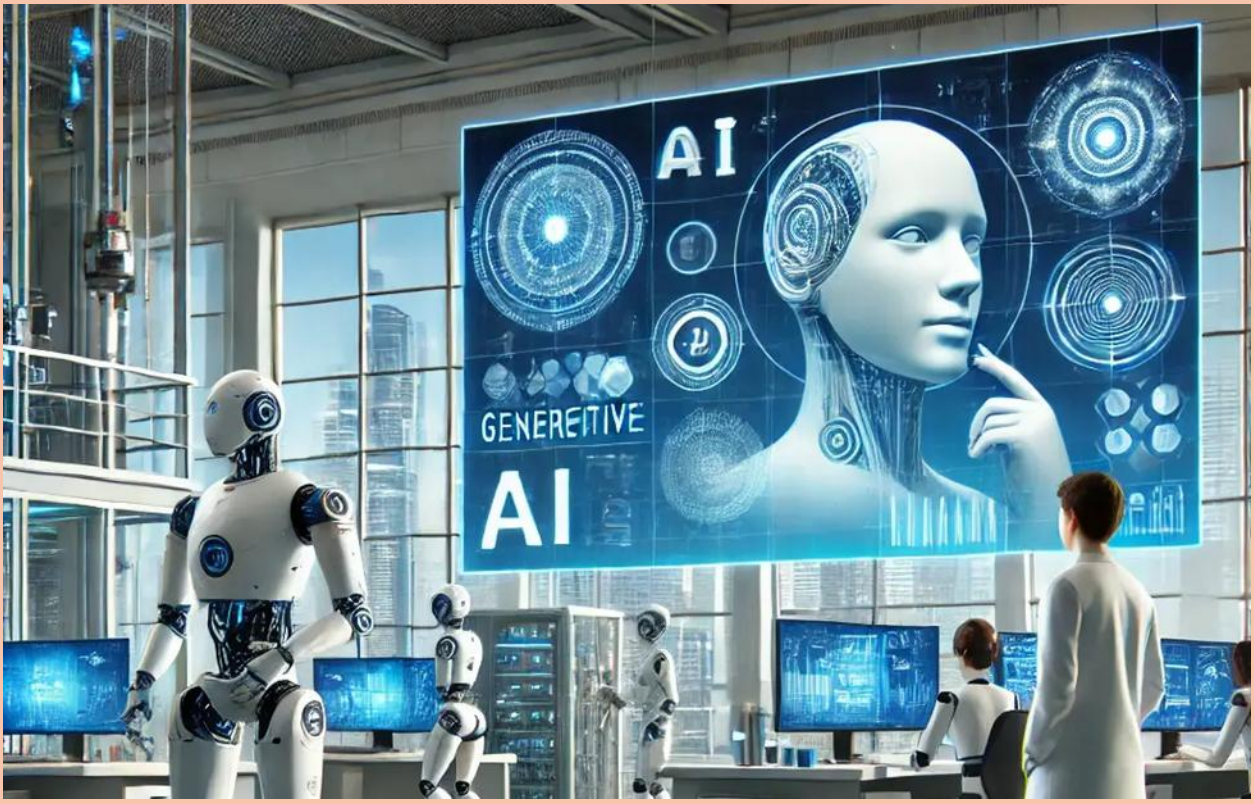
1. Bad AI, SAFE AI Foundation. See: https://safeaifoundation.com/bad-ai

2. "From AGI to ASI", Google Deepmind. See: https://deepmind.google/research/publications/239142/

3. Youtube – "10 Things About The Coming AI". See: https://www.youtube.com/watch?v=tFx_UNW9I1U&t=4s

4. "5 Potential Upcoming Dangers of AI", SAFE AI Foundation. See: https://drive.google.com/file/d/1Mn5SvPncCHgwvxdzHTNzrGMpDUDeLpUF/view?pli=1

5. "AI Manifestation Unfolded" – SAFE AI Foundation USA. See: https://safeaifoundation.com/ai-manifestation-unfolded-ck-toh



SAFE
AI
FOUNDATION

www.safeaifoundation.com

The SAFE AI Foundation was formed in 2025, in the heart of the Silicon Valley, California, USA. Its mission is to embrace the AI revolution and promote the use of good AI to improve on the quality of life and the economy. It will address the issues of bad AI and its impacts on society. The foundation will inform, educate, and mitigate the dangers of bad AI and help the society to embrace good AI. The foundation will interact with industries, universities and public sector to jointly advance the progress of Good AI. The foundation is a non-profit 501c(3) organization, registered in the State of California, USA.