5 February 2025

Safety Standards Delivering Controllable and Beneficial AI Tools

Contact: Ben Cumming, **ben.cumming@futureoflife.org** | View document online: **futureoflife.org/standards**

The past decade has seen the extraordinary development of artificial intelligence from a niche academic pursuit to a transformative technology. AI tools promise to unlock incredible benefits for people and society, from Nobel prize-winning breakthroughs in drug discovery to autonomous vehicles and personalized education. Unfortunately, two core dynamics threaten to derail this promise:

1. First, the speed and manner in which AI is being developed—as a chaotic nearly-unregulated race between companies and countries—incentivizes a race to the bottom, cutting corners on security, safety and controllability. We are now closer to figuring out how to build general-purpose smarter-than-human machines (AGI) than to figuring out how to keep them under control.
2. Second, the main direction of AI development not toward trustworthy controllable tools to empower people, but toward potentially uncontrollable AGI that threatens to replace them, jeopardizing our livelihoods and lives as individuals, and our future as a civilization.

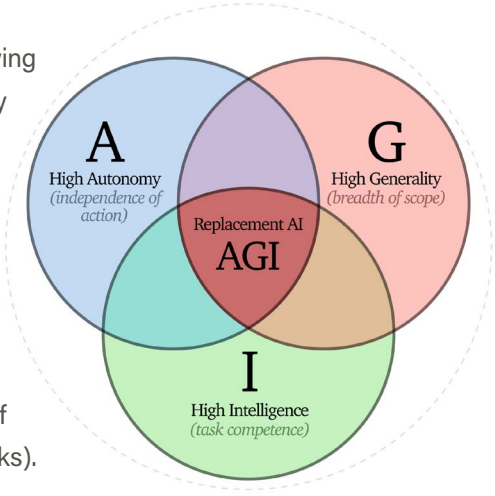
With many leading AI scientists and CEOs predicting AGI to be merely 1-5 years[1] away, it is urgent to correct the course of AI development. Fortunately, there is an easy and well-tested way to do this: start treating the AI industry like all other high-impact industries, with legally binding safety standards, incentivizing companies to innovate to meet them in a race to the top. We make a concrete proposal for such standards below.

The Need to Act Now

Only six years ago, many experts believed that AI as capable as GPT-4 was decades, even centuries away. Now OpenAI's o3 system recently scored[2] 85% on the ARC-AGI benchmark[3], and showcases human-level reasoning skills and PhD level skills in biology, chemistry and physics. OpenAI's "Deep research" feature scores 18% on the ultra-difficult "Humanity's Last Exam".[4] Its CEO Sam Altman has claimed they already know how to build AGI, and other companies have explicitly or implicitly announced it as a goal. It is imperative then to understand what AGI is and what it would mean to develop it on our present path.

AI Risk Thresholds and AGI's Three Parts

To avoid innovation-stifling governmental overreach, we recommend classifying AI systems into tiers of increasing potential risk, with commensurately stricter standards, akin to the five-tier U.S. classification of drugs, and the "AI Safety Levels" from Anthropic's Responsible Scaling Policy[5]. As described below, our risk tiers are grounded in the capabilities of AI systems in the three core areas required for AGI.

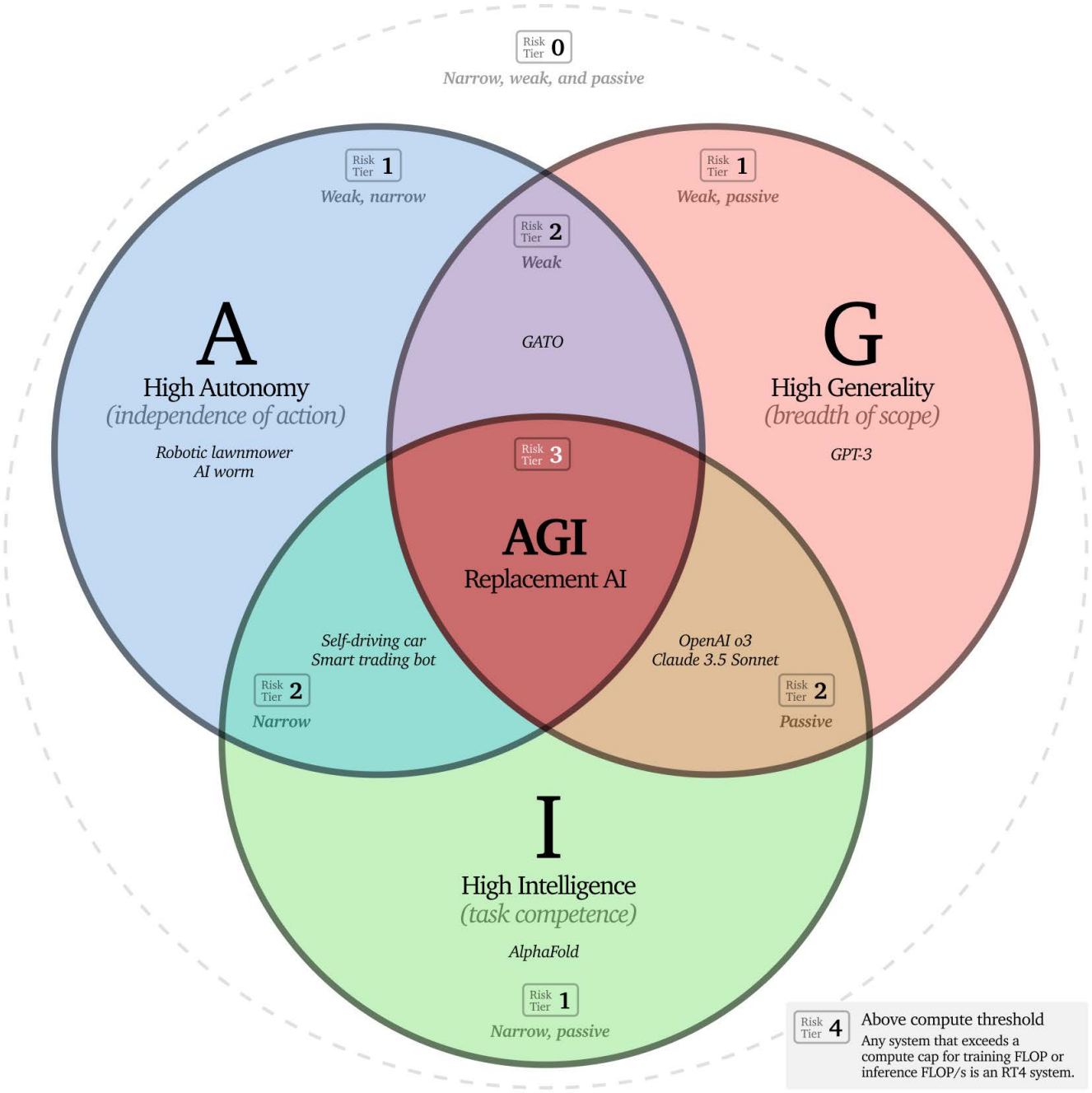


From *Keep the Future Human*, A. Aguirre, Forthcoming

Although AGI stands for "Artificial General Intelligence", it is a useful mnemonic to think of it as an "Autonomous General Intelligence": the triple combination of **Autonomy** (independence of action), **Generality** (breath of capability, application, and learning), and **Intelligence** (competence at tasks). AI systems possess these characteristics in different measures.

While possession of these characteristics can deliver enormous rewards from AI, high levels of convergence between them can result in correspondingly high degrees of unpredictability and risk, with the most dangerous and uncontrollable systems possessing high levels of all three. An AI with all these characteristics would be capable of the large range of effective cognition and action that a human is. However, it would be much more capable—and dangerous—than any individual human, given its scalability in speed, memory, and reproducibility, and its potential for rapid self-improvement and replication.

The combination of all three characteristics is currently unique on Earth to homo sapiens. This is why possession of all three capabilities could enable machines to replace us, as individuals in our work or in a wider sense as a species. This is what has made AGI both a target of development and an unprecedented risk.



From *Keep the Future Human*, A. Aguirre, Forthcoming

AGI: Autonomous Generally Intelligent Systems to Fully Replace Humans

The nature of AGI as a human replacement—rather than tool—is implicit in how it has been traditionally been defined: as a system that can match human performance on most intellectual tasks. Even more tellingly, OpenAI[6] has defined AGI as "highly autonomous systems that outperform humans at most economically valuable work".

The desire to build powerful machines that can entirely replace humans as workers, educators, companions and more is the goal of a tiny minority. How can we ensure that we instead follow the widely desired path—one that makes us strong, rather than obsolete?

Tiered Safety Standards

Our proposed AI risk tier classification counts the number (between 0 and 3) of A, G and I factors that are present to high degree within a given AI system.

For example, Google DeepMind's AlphaFold, which solved the protein-folding problem, is more intelligent than any human at its (narrow) task, but not autonomous or general. It therefore scores 1, placing it in Risk Tier 1. OpenAI's GPT-3 was very general (it could answer questions, generate text, assist with basic coding etc.) but it was not very competent (it was inaccurate, inconsistent, and reasoned poorly). Therefore it falls squarely into Tier 1. Their recently released o3 model, however, has demonstrated human-level reasoning and can answer PhD level science questions, while still being very general, so is in Risk Tier 2.

The diagram above us allows us to identify and categorize systems with different levels of A/G/I convergence (and therefore risk). This categorization allows us to place systems in corresponding risk tiers (see the table below), to which a corresponding level of safety and controllability requirements can be applied: roughly speaking, further from the center corresponds to lower risk. Different tiers of convergence trigger different requirements for training (e.g. registration, pre-approval of safety plan) and different requirements for deployment (e.g. safety cases, technical specifications).

This approach avoids overly onerous requirements being placed on relatively low-risk/narrow systems, while ensuring that controllability can be guaranteed for more potentially dangerous ones. Rather than hindering competition, this tiered approach will drive companies to innovate to meet requirements, helping to realize the incredible benefits of AI in a secure, responsible, and intentional way.

This approach does not require any re-imagining of industry governance. Companies in any other sector, from automobiles and pharmaceuticals to sandwiches, must provide satisfactory evidence that their products are safe before release to the public. They must meet safety standards, and the AI industry should be no different. Furthermore, it is sensible and consistent to place higher standards on technologies or products that have greater potential for harm. We would not want nuclear reactors tested in the same category as sandwiches.

Our Tiered Proposal

Risk Tier	Trigger(s)	Requirements for training	Requirements for deployment
RT0	AI weak in autonomy, generality, and intelligence	None	None
RT1	AI strong in one of autonomy, generality, and intelligence	None	**Qualitative guarantees:** Safety audits by national authorities wherever the system can be used, including blackbox and whitebox red-teaming
RT2	AI strong in two of autonomy, generality, and intelligence	Registration with national authority with jurisdiction over the lab	**Quantitative guarantees:** National authorities wherever the system can be used must approve company-submitted assessment bounding the risk of major harm below authorized levels
RT3	AGI strong in autonomy, generality, and intelligence	Pre-approval of safety and security plan by national authority with jurisdiction over the lab	**Formal guarantees:** National authorities wherever the system can be used must certify company-submitted formal verification that the system meets required specifications, including cybersecurity, controllability, a non-removable kill-switch, and robustness to malicious use
RT4	Uses more than 10^{27} FLOP for training or more than 10^{20} FLOP/s for inference	Prohibited pending internationally agreed lift of compute cap	**Prohibited** pending internationally agreed lift of compute cap

Risk classifications and safety standards, with tiers based on compute thresholds as well as combinations of high autonomy, generality, and intelligence:

- **Strong autonomy** applies if the system is able to perform many-step tasks and/or take complex real-world actions without significant human oversight or intervention. Examples: autonomous vehicles and robots; financial trading bots. Non-examples: GPT-4; image classifiers
- **Strong generality** indicates a wide scope of application, performance of tasks for which the model was not deliberately and specifically trained, and significant ability to learn new tasks. Examples: GPT-4; mu-zero. Non-examples: AlphaFold; autonomous vehicles; image generators
- **Strong intelligence** corresponds to matching human expert-level performance on the tasks for which the model performs best (and for a general model, across a broad range of tasks.) Examples: AlphaFold; mu-zero; OpenAI o3. Non-examples: GPT-4; Siri

Safety Guarantees and Controllability

Ensuring safe and controllable systems requires safety standards that scale with a system's capabilities, all the way up to and including AGI—which might soon develop into superintelligence. Risk Tiers 0, 1, 2 and 3 correspond to systems that are progressively more difficult to control, and whose potential harm is progressively greater. This is why the corresponding requirements in the table are progressively stricter, ranging from none to qualitative, quantitative and formal proofs. We want our tools to be powerful but controllable (who wants an uncontrollable car?), so we define "Tool AI" as AI that can be controlled with an assurance level commensurate with its Risk Tier.

Systems with a low degree of risk should not receive onerous requirements that limit their positive impact. Tier 0 systems are therefore unregulated, while Tier 1 systems require only qualitative safety standards. Since Tier 2 systems have significantly greater potential for harm, they require quantitative guarantees, just as is currently the case for other industries with powerful technology. For example, U.S. nuclear power plants are only allowed if government-appointed experts approve a company-provided study quantifying the annual meltdown risk as less than one in a million; similarly, the FDA only approves drugs whose side effects are quantified below an acceptable level. There are many promising approaches to providing such quantitative AI safety guarantees (Dalrymple et al. 2024).

Tier 3 (AGI) is much more risky because it can broadly match or exceed human ability, deceive people, create long-term plans, and act autonomously in pursuit of goals—which by default include self-preservation and resource acquisition. This is why top AI researchers have warned that AGI may escape human control and even cause extinction[7]. To guarantee that AGI remains controllable, a Tier 3 system must be mathematically proven to be controllable using formal verification—which includes proving that it will never resist being turned off.

Just as AI progress has revolutionized our ability to auto-generate text, images, video and code, it will soon revolutionize our ability to auto-generate code and proofs that meet user-provided specifications. In other words, rather than deploying unverifiable black-box neural networks, it may soon be possible to have AI systems write deployable formally verifiable code, implementing powerful machine-learned algorithms and knowledge (see Tegmark & Omohundro 2023, and provablysafe.ai for an overview of the field).

Endnotes

1 Improve The News. (n.d.). Will AI Surpass Human Intelligence, and When?. Verity. Retrieved 02 05, 2025, from https://www.improvethenews.org/controversy/ai-surpass-human-intelligence

2 The Conversation. (2024, 12 24). An AI system has reached human level on a test for 'general intelligence.' Here's what that means. Retrieved 02 05, 2025, from https://theconversation.com/an-ai-system-has-reached-human-level-on-a-test-for-general-intelligence-heres-what-that-means-246529

3 ARC Prize. (n.d.). ARC-AGI. Retrieved 02 05, 2025, from https://arcprize.org/arc

4 MSN. (2025, 02 04). OpenAI's Deep Research smashes records for the world's hardest AI exam, with ChatGPT o3-mini and DeepSeek left in its wake. Retrieved 02 05, 2025, from https://www.msn.com/en-gb/money/technology/openais-deep-research-smashes-records-for-the-worlds-hardest-ai-exam-with-chatgpt-o3-mini-and-deepseek-left-in-its-wake/ar-AA1yoHeq

5 Anthropic. (2023, 09 19). Anthropic's Responsible Scaling Policy. Retrieved 02 05, 2025, from https://www.anthropic.com/news/anthropics-responsible-scaling-policy

6 OpenAI. (2023, 02 16). How should AI systems behave, and who should decide?. Retrieved 02 05, 2025, from https://openai.com/index/how-should-ai-systems-behave/

7 Center for AI Safety. (n.d.). Statement on AI Risk. Retrieved 02 05, 2025, from https://www.safe.ai/work/statement-on-ai-risk

Frequently Asked Questions

Q: But isn't powerful AI science fiction, or decades away?

No. Six years ago, most scientists thought language-mastering AIs like GPT-4 were decades away; now they are in commonplace. Today's AIs have already arguably passed the Turing Test. Timelines have collapsed, with tech CEOs like OpenAI's Sam Altman, Anthropic's Dario Amodei and DeepMind's Demis Hassabis predicting that AGI will be built in 1-5 years, and many whistleblowers, investors and academics concurring.

Q: But isn't AGI necessary to reap AI's benefits?

No. Almost all of the benefits cited by those seeking to build AGI can be reliably captured by intentionally developing Tool AI systems to solve specific problems. Tool AI can save millions of lives per year on roads, greatly improve cancer diagnosis, and realize breakthroughs in pandemic prevention, education, energy reduction and more. It has already helped us fold proteins and develop new medicines, through the Nobel Prize-winning AlphaFold.

Q: But surely AGI is controllable?

No. We are closer to building AGI than we are to controlling it. In fact, we have no idea how to control it. As Alan Turing presciently observed in 1951, "once the machine thinking method had started, it would not take long to outstrip our feeble powers... We should have to expect the machines to take control." When it comes to intelligent entities, the smarter ones tend to take control—just like humans did. It doesn't matter if the AI is evil or conscious, only that it is extremely competent, and accomplishes goals that aren't aligned with our own. Companies building AGI have produced zero evidence of having solved the control problem.

Q: But isn't uncontrollable AGI desirable?

No. There are people who want to deny humans their right to a meaningful future, and want us to be replaced by machines through uncontrollable AGI, but they are a tiny minority. Everyone else is firmly on "Team Human", and would prefer that we keep control of our destiny.

Q: But surely AGI is inevitable?

No. There are many powerful, profitable technologies that we have successfully banned because we decided they were too dangerous—from human cloning to bioweapons. All we need to do to course-correct is enact safety standards that require AI companies to guarantee their products are safe before releasing them, just like in every other industry.

Q: But won't a country imposing AI standards get overtaken?

No. Uncontrollable AGI presents the greatest threat to the country building it, more so than any adversary. Therefore any supposed advantage attained by building it first would be incredibly short-lived. Once countries recognize this danger, they will not only impose their own domestic standards to prevent uncontrollable AGI, but will also work together to prevent other countries from doing building it. Furthermore, the proposed safety standards will drive unprecedented innovation, as we use AI itself to build incredibly powerful Tool AI that we can reliably verify and trust–securely realizing the benefits of this amazing technology.

Q: Why aren't voluntary industry commitments enough?

Because companies are stuck in a race to the bottom whose incentives favor them abandoning such commitments rather than getting outcompeted.

About the Organization: The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence (AI). Learn more at futureoflife.org.