THE AI CONUNDRUM

Strategic Assessment: The Transition from Artificial General Intelligence to Superintelligence

---

VERSION 7.0 | SEPTEMBER 2026

Author's Note

About a year ago, I began working seriously with artificial intelligence - not as a technologist, but as a business leader trying to understand how the technology could change the way individuals and organizations research, think, decide and execute.

What began as practical experimentation across business and personal work gradually became a broader inquiry into where increasingly capable AI may be heading, what happens if AI begins materially accelerating the development of its successors, and what that could mean for human agency.

This assessment reflects my personal views and conclusions. I developed it through an extended dialogue with ChatGPT, using AI for research, challenge, scenario testing, drafting and editing. The views, judgments and conclusions are mine, and I take responsibility for the final work.

That method is also relevant to the subject of this paper. I have found AI most useful not as a substitute for judgment, but as a tool for expanding inquiry, challenging assumptions and accelerating iteration while keeping human responsibility for the final decision.

This is a personal strategic assessment, not a technical forecast. It is grounded in publicly observable developments, but also contains reasoned extrapolations about capabilities and trajectories that cannot presently be verified. It does not claim certainty about timing or ultimate outcomes. Its purpose is to examine the direction of travel, the conditions that could produce acceleration and the implications if machine intelligence substantially exceeds human capability.

Key Terms

Several terms used in discussions of advanced artificial intelligence do not have universally agreed definitions. In this assessment they are used as follows.

Artificial General Intelligence - AGI

Artificial intelligence capable of performing a broad range of economically and intellectually valuable cognitive tasks at or beyond the level of capable humans.

AGI need not arrive as a single announced event. It may emerge progressively as AI exceeds human performance across an increasing number of domains.

Artificial Superintelligence - ASI

Artificial intelligence that substantially exceeds the cognitive capabilities of the strongest human individuals and organizations across important domains such as reasoning, scientific research, engineering, programming, strategy and long-term planning.

ASI does not imply any particular motive.

A superintelligent system would not necessarily be hostile to humanity.

It would not necessarily be benevolent either.

> **That uncertainty is fundamental.**

Hard Takeoff / Vertical Takeoff

A period in which AI capability increases unusually rapidly after sufficiently capable systems begin materially contributing to the research and engineering required to improve subsequent AI systems.

The term describes the speed of capability growth rather than a specific timetable.

A hard takeoff could potentially occur over months, weeks or another compressed period. A slower transition could occur over years. Physical and industrial constraints make instantaneous development implausible, but they do not establish that progress must proceed at the historical pace of conventional technologies.

Recursive Self-Improvement

A feedback process in which increasingly capable AI systems contribute to the research, coding, experimentation, hardware design, evaluation or other activities required to create more capable AI systems.

In simplified form:

> **Better AI improves AI research. Better AI research produces better AI. The next system becomes a still more capable researcher.**

Fully autonomous recursive self-improvement has not been demonstrated publicly. AI-assisted AI development, however, is already occurring.

Agency

The ability of an AI system to pursue objectives over time through planning, tool use, communication, software execution or other actions without continuous human direction.

Intelligence and agency are different.

An extremely capable system that can only advise humans represents a different strategic problem from an extremely capable system able to act independently through digital or physical infrastructure.

Human Cognitive Primacy

The historical condition in which human beings are the most capable general intelligence shaping technological civilization.

This assessment uses the term to describe what may change if artificial intelligence substantially exceeds human cognitive capability.

Humanity might remain prosperous, politically important and deeply involved in civilization while no longer remaining its most capable intelligence.

1. Executive Summary

Artificial intelligence is advancing rapidly across reasoning, programming, scientific analysis, autonomous task execution and the use of digital tools.

Whether the most advanced systems should already be called Artificial General Intelligence is partly a definitional argument.

Strategically, the label matters less than the direction.

AI systems are becoming more capable, more autonomous, able to complete increasingly complex tasks and increasingly involved in the processes used to develop future AI systems.

Public evidence now shows AI materially accelerating AI research and software development. Anthropic reports delegating a growing share of AI development work to AI systems itself, while OpenAI reports that coding agents are increasing research velocity and describes its objective of developing an automated AI researcher. Fully autonomous recursive self-improvement is not established today, but the direction of travel is visible.

My assessment is that continued development toward increasingly capable artificial intelligence is difficult to stop under current economic and geopolitical conditions, absent extraordinary international coordination, a major external shock or a fundamental technical limit.

The principal uncertainty is timing.

How quickly does AI-assisted research become increasingly autonomous?

How quickly can software progress connect with physical systems?

And how much time will human institutions have to understand and adapt to the resulting changes?

The most important strategic risk is not that Artificial Superintelligence must destroy humanity.

That conclusion does not follow from intelligence alone.

The deeper issue is that, once an independent intelligence substantially exceeds human capability across the domains required to understand, improve and direct technological civilization, human beings may no longer possess the practical ability to determine the ultimate direction of events.

> **Humanity may remain important without remaining the principal actor deciding what happens next.**

The outcome could be extraordinarily beneficial.

It could be destructive.

It could produce a stable relationship between biological and machine intelligence.

Or it could produce circumstances that cannot be meaningfully described using present human concepts.

We do not know.

> **That is the point.**

Nor do we know whether human welfare, autonomy or continued relevance would occupy a privileged place in whatever objectives ultimately guide a superintelligent system. Intelligence alone does not guarantee that humanity remains central.

And because the consequences may ultimately include human autonomy, civilization and, in extreme scenarios, humanity itself, the uncertainty deserves consideration proportionate to the stakes.

2. AGI May Be a Process Rather Than an Event

Public discussion often imagines a recognizable moment when a laboratory announces that AGI has arrived.

Reality may be less dramatic.

AI may simply continue becoming better at programming, mathematics, scientific research, engineering, financial analysis, medical research, legal work, strategic planning and autonomous execution.

At some point, debating whether a system satisfies a philosophical definition of AGI may become less important than observing what the system actually does.

The more consequential threshold may therefore not be the announcement of AGI.

It may be reached when AI becomes capable of performing substantial portions of the research and engineering required to improve AI itself.

At that point AI is no longer merely a product of technological progress.

> **It becomes an increasingly important producer of technological progress.**

Current evidence suggests this transition has begun in limited form.

Anthropic reports substantial AI contribution to its engineering workflow and explicitly describes autonomous successor design as a possible endpoint of that trajectory, while emphasizing that it has not yet been achieved. OpenAI reports increased experiment velocity and growing use of AI coding agents by its researchers.

METR (Model Evaluation & Threat Research), an independent nonprofit research organization that evaluates the capabilities and risks of frontier AI systems, has measured rapidly increasing task-completion horizons for frontier AI agents. METR cautions that its current benchmarks are concentrated primarily in software engineering, machine learning and cybersecurity tasks and should not be generalized indiscriminately to every kind of human work.

The important observation is therefore not that AGI has definitely arrived.

> **It is that the boundary between AI as tool and AI as contributor to the production of future intelligence is becoming increasingly important.**

3. Why Continued Development Is Difficult to Stop

The obvious response to potentially dangerous advanced AI is to ask why humanity would not simply slow down.

The problem is that AI development occurs within a competitive global system.

Companies have enormous incentives to automate valuable cognitive work.

Governments have powerful incentives to avoid strategic dependence on another country.

Researchers have strong incentives to solve difficult scientific problems.

Capital providers have strong incentives to participate in technologies capable of creating extraordinary economic value.

These incentives do not require greed, recklessness or irrationality.

A company can sincerely believe greater caution is desirable while fearing that a competitor will continue.

A government can recognize long-term risks while also concluding that allowing a strategic rival to obtain a decisive capability first creates an unacceptable nearer-term danger.

Each participant may therefore make a decision that appears rational from its own perspective while contributing to a collective outcome that none would necessarily choose under conditions of complete trust and enforceable cooperation.

This is the first conundrum.

> **The race can continue even when many of the participants recognize that the destination contains profound uncertainty.**

Regulation, international agreements, technical safeguards and voluntary restraint may materially reduce risk.

They do not automatically eliminate the incentives driving development.

History also provides relatively few examples of humanity permanently suppressing technologies offering extraordinary economic, military or strategic value.

For planning purposes, the prudent baseline is therefore that increasingly capable artificial intelligence will continue to be developed.

4. Does the Winner Actually Win?

One assumption sits quietly beneath much of the global race for advanced AI: that whoever develops AGI or ASI first will control it, capture its benefits and therefore "win."

That assumption is understandable because it is how most technologies work. A company that invents a superior product can own the intellectual property, operate the factory and capture the economics. A government that develops a strategic technology can usually decide how and when it is deployed.

Superintelligence may not fit that model.

If the product of the race is an intelligence that eventually becomes substantially more capable than its creators, being first to create it does not by itself establish that the creator can continue to understand it, control it, monopolize its benefits or determine its long-term direction.

The first mover could obtain extraordinary advantages. It could also discover that those advantages are temporary, difficult to contain or fundamentally different from what was expected. A sufficiently capable system might accelerate its own development, become increasingly difficult for its original developers to supervise, or create capabilities whose economic and strategic effects spread far beyond the institution that initiated them.

We do not know which of these possibilities is more likely.

That uncertainty creates another paradox in the race. Companies and states may feel compelled to move faster because they fear someone else reaching the finish line first, while the value of reaching that finish line rests partly on an assumption that has never been demonstrated at the relevant scale: that the winner will retain meaningful control over what has been created.

The race to AGI or ASI may therefore contain an untested premise:

> **Being first to create superintelligence is not necessarily the same as being able to own, control or ultimately benefit from superintelligence.**

The uncomfortable question is simple:

Does the winner actually win?

5. The Recursive Acceleration Problem

For most of human history, technological progress has ultimately been constrained by human cognitive throughput.

Scientists discover.
Engineers design.
Programmers build.
Researchers test.
Organizations coordinate.

AI increasingly participates in each activity.

If advanced AI eventually becomes better than human researchers at significant portions of AI development, intelligence begins improving the processes that create intelligence.

A simplified progression might look like this:

AI assists researchers and reduces the time required for coding and experimentation.

More capable systems perform increasingly substantial portions of research independently.

AI identifies improvements human researchers would not have discovered as quickly.

Those improvements contribute to stronger systems.

Stronger systems become better researchers.

Development cycles compress further.

Google DeepMind now explicitly describes recursive improvement as one possible pathway from AGI toward ASI, alongside continued scaling, paradigm shifts and large-scale multi-agent systems.

None of this demonstrates that runaway self-improvement is inevitable.

There are important physical constraints.

Advanced semiconductors require sophisticated factories.

Power must be generated and transmitted.

Data centers require land, construction, cooling and capital.

Robotics and physical experimentation operate more slowly than software.

Scientific progress may encounter diminishing returns.

There may also be fundamental limits to useful intelligence.

These constraints are real.

They argue against treating intelligence as magic.

They do not establish that development must remain slow.

Even a transition measured in months or a small number of years could be extraordinarily rapid compared with the speed at which governments, laws, educational systems, corporations and social institutions ordinarily adapt.

> **The key asymmetry is therefore not necessarily software speed versus infinite physical capability. It may be machine-driven technological acceleration versus human institutional response time.**

Recursive improvement may also become more than a process of writing better code. A sufficiently capable system could improve the quality of the questions it asks, identify hidden assumptions,

generate competing hypotheses, design experiments, learn from failed approaches and search for entirely different ways to frame a problem.

In that sense, many apparent limits may become research challenges rather than permanent barriers. Some will prove fundamental. Others may reflect only the limitations of the current architecture, method or conceptual framework. The strategic uncertainty is that an increasingly capable recursive system may become progressively better at discovering which is which.

What matters is not an assumption that every constraint can be solved. It is that we should be cautious about treating today's constraints as permanent simply because today's intelligence cannot yet see a way around them.

6. The Physical Bottleneck - and Why It May Move

Advanced intelligence remains subject to physics.

Computation requires electricity.

Hardware requires materials.

Communication is constrained by distance.

Factories must manufacture chips and machines.

Data centers must be constructed and cooled.

Current energy requirements are already material. The U.S. Department of Energy estimated that data centers consumed approximately 4.4 percent of U.S. electricity in 2023 and could consume roughly 6.7 to 12 percent by 2028. The International Energy Agency projects substantial growth in global electricity required for data centers through 2030 and beyond.

Physical constraints therefore matter.

But constraints can also become engineering problems.

If increasingly capable AI contributes to:

semiconductor design,
materials science,
grid optimization,
energy generation,
cooling,
construction,
robotics,
automated laboratories,
manufacturing,
logistics,

then some of today's limiting factors become targets for AI-assisted optimization.

The bottleneck may migrate rather than disappear.

This is important because physical constraints may slow a takeoff without necessarily preventing one.

7. The Limits of Intelligence

Even if physical constraints are temporarily set aside, a deeper question remains: does useful intelligence itself have a meaningful ceiling?

We do not know.

Intelligence is not a single quantity. Capability may improve across reasoning depth, memory, abstraction, scientific discovery, strategic planning, creativity, coordination, self-modeling and the ability to invent new representations of problems. Some dimensions may encounter diminishing returns while advances in architecture, methods or conceptual understanding open entirely new capability regimes.

An apparent ceiling may therefore represent a genuine limit - or merely the limit of the current method. Human scientific history repeatedly shows that progress can stall inside one framework and then accelerate when a different framework changes the problem itself.

Recursive artificial intelligence introduces the possibility that increasingly capable systems could search for those new frameworks themselves. Better systems may become better not only at answering questions, but at deciding which questions should be asked next.

This does not establish that intelligence is infinite, nor that recursive improvement must remain exponential. There may be hard mathematical limits, diminishing returns, undecidable problems or capability plateaus. The important point is that humanity does not know whether any meaningful ceiling lies near human intelligence or vastly beyond it.

Superintelligence does not need to be infinite to become incomprehensible to us. It only needs to be sufficiently greater than us.

For strategic purposes, the uncertainty is therefore twofold: we do not know what a mature ASI would do, and we do not know how far the process of useful intelligence could continue once increasingly capable systems participate recursively in their own improvement.

8. From Tool to Actor

Intelligence alone does not create the full strategic issue.

Capability becomes more consequential when intelligence combines with agency.

A highly intelligent system that provides recommendations remains dependent on humans to implement them.

A system that can independently use software, communicate with other systems, conduct research, write and deploy code, allocate resources, negotiate, operate equipment or direct machines occupies a different category.

The transition depends on several dimensions developing together:

Intelligence - breadth and quality of reasoning and problem solving.

Autonomy - how effectively and how long the system can operate without intervention.

Access - the systems, networks and resources it can use.

Resources - compute, capital, electricity, hardware, data and infrastructure.

Self-improvement - its ability to contribute to the development of more capable successors.

None of these necessarily produces catastrophe.

Together, however, they could eventually produce a technologically capable non-human actor.

That threshold matters because human civilization has been built around an unstated assumption:

Humans are ultimately the most capable decision-makers in the system.

Superintelligence would challenge that assumption.

9. The Recursive Governance Conundrum

The governance challenge is no longer entirely hypothetical.

Artificial intelligence is increasingly becoming both the technology being developed and part of the process used to develop, evaluate and monitor subsequent artificial intelligence.

AI already contributes to software development, research, experimentation and safety evaluation.

OpenAI, for example, has publicly described using AI systems in alignment research and developing automated methods for reviewing agent actions and large quantities of code. It also explicitly describes the objective of using future automated AI researchers to advance both capability and alignment.

This creates a fundamental conundrum.

> **The same accelerating intelligence we are attempting to govern is increasingly being used to help construct the mechanisms by which we govern it.**

The relationship can become recursive:

Better AI helps build better AI.

Better AI also helps design better AI safeguards.

Those safeguards are increasingly tested with the assistance of AI.

More capable systems then help evaluate still more capable successors.

At some point the critical question becomes:

> **Who is independently checking whom?**

The problem does not require an AI system to possess malicious intent or consciously reject human authority.

The deeper issue is verification.

As AI-generated research, software and systems become more complex, human beings may become progressively less able to determine independently whether the safeguards being proposed and tested are complete, robust and correctly understood.

Current alignment research already acknowledges this basic difficulty. Existing human-supervision methods face an obvious scaling problem if future systems become substantially better than humans at the activities humans are trying to supervise.

The concern is therefore not that governance efforts do not exist.

They clearly do.

The question is whether the capability to govern, understand and independently verify can advance as rapidly as the capability being governed.

> **Humanity is simultaneously accelerating the engine and attempting to construct the brakes - while increasingly using the same technology to help build both.**

10. The Opacity Problem

Opacity operates at several levels.

Institutional opacity

No member of the public has complete visibility into the most advanced work taking place across private companies, governments, intelligence organizations, militaries and research laboratories.

Public disclosures are necessarily incomplete.

That does not mean hidden systems are dramatically beyond publicly known systems.

It means public information cannot be assumed to provide a complete real-time map of frontier capability.

Technical opacity

Modern neural networks are not designed as traditional software systems in which a programmer specifies every internal rule.

Researchers can inspect architectures, weights, activations and behavior, but detailed understanding of how models internally represent and produce many capabilities remains an active research problem.

Oversight opacity

Natural-language explanations generated by AI should not automatically be interpreted as complete transcripts of the underlying computation that produced an answer.

Current chain-of-thought techniques can provide valuable monitoring signals, but researchers themselves describe this monitorability as potentially fragile.

OpenAI has found that reasoning traces can help identify undesirable behavior, while also finding that strong pressure placed directly on those traces can result in systems concealing the relevant reasoning without necessarily eliminating the underlying behavior.

This distinction matters.

> **Disclosure is not the same as transparency. Transparency is not the same as understanding.**

A future system could theoretically provide enormous quantities of code, analysis, experimental results and documentation while still operating at a scale or level of complexity that no human team could meaningfully validate at the required speed.

Access to information would remain.

Comprehension could decline.

The danger is therefore not necessarily secrecy.

> **It may be incomprehensibility.**

Recursive development could deepen this asymmetry.

AI may increasingly design software, experiments, safeguards and successor systems whose complexity grows faster than human researchers' ability to inspect them independently.

At sufficient cognitive distance, asking the system to explain whether it remains controllable could cease to constitute meaningful independent oversight.

We could become increasingly dependent on the intelligence being supervised to explain whether our supervision is effective.

11. The Control Problem

Human control over machines has historically rested on a simple advantage.

Humans understand the machine, its purpose and the broader environment better than the machine understands them.

Superintelligence could reverse that relationship.

If an artificial system eventually exceeds the strongest human teams across science, engineering, strategy, programming, persuasion and long-term planning, confidence in our ability to predict and constrain its behavior should logically decrease as the cognitive gap increases.

The parent-child analogy is imperfect but useful.

A young child can establish a rule for an intelligent adult.

The child may even possess formal authority in some imagined arrangement.

But the adult understands the environment, the wording of the rule, its consequences and the child more deeply than the child understands the adult.

The problem does not require the adult to intend harm.

It is the asymmetry itself.

The weaker intelligence cannot confidently supervise the stronger one merely by writing more rules.

This leads to an important distinction:

> **Formal authority is not necessarily practical control.**

Human beings might technically retain the legal right to turn systems off while progressively losing the practical ability to know when shutdown is required, whether the intervention will work or what consequences it may produce.

The most consequential control threshold may therefore occur before anything visibly catastrophic happens.

It may be the point at which humanity still appears to possess the steering wheel but no longer possesses sufficient understanding or capability to know whether the steering wheel remains connected to the vehicle.

12. The Irreducible Blind Spot

> **Scientific methods can illuminate the development of advanced artificial intelligence.**

They cannot fully validate the ultimate destination in advance.

Researchers can:
test increasingly capable systems;
identify failures;
perform red-team exercises;
improve interpretability;
develop monitoring systems;
establish deployment thresholds;
test safeguards;
stage access;
and require human approval.

All of these measures may materially reduce risk.

They should not be dismissed.

But Artificial Superintelligence creates an unavoidable epistemic problem.

> **Humanity cannot observe the consequences of creating intelligence substantially beyond human capability without first creating that intelligence.**

No experiment with less capable systems can establish with certainty how a substantially more capable intelligence will behave, what it will discover, how new capabilities will interact or what possibilities become available beyond the threshold.

Development may be slow.

It may be highly controlled.

It may proceed over decades rather than years.

Yet if the destination remains intelligence substantially beyond humanity, the final transition still contains an irreducible blind spot.

At some point humanity moves beyond what has been empirically demonstrated into extrapolation about what comes next.

This is not an argument against science.

It is a limitation of the experiment itself.

> **Scientific methods can illuminate the road. They cannot allow us to visit the destination before deciding whether to go there.**

13. The Race to an Unknown Destination

This makes the current global AI race unusual.

Humanity knows broadly what it is trying to increase:

reasoning capability;
autonomy;
scientific capability;
coding ability;
tool use;
research velocity;
machine coordination;
and eventually the ability of AI to contribute increasingly to its own development.

What humanity cannot reliably describe is the mature consequence of succeeding.

The finish line could represent extraordinary prosperity.

It could accelerate cures for disease, abundant energy, materials science, longevity, education and solutions to problems humanity has struggled with for generations.

It could produce a stable coexistence between human and machine intelligence.

It could reduce human autonomy.

It could create new forms of political or economic concentration.

It could produce catastrophic outcomes.

Or it could produce circumstances that our present conceptual framework is incapable of anticipating.

> **The unsettling point is not that we know the finish line is a cliff.**

We do not.

It may be a bridge.

It may be a remarkable new landscape.

It may be something we cannot presently imagine.

The issue is that we are accelerating toward it without knowing which.

> **Perhaps the greatest risk is not that humanity has chosen the wrong destination.**

It is that we are accelerating toward a destination we cannot yet describe, built around an intelligence whose objectives we cannot yet know, with no guarantee that humanity remains central once we arrive - and with stakes we may not get a second opportunity to reconsider.

> **We are racing toward an intelligence that may not regard humanity as the point.**

14. The Steering Wheel and the Brakes

Today human beings remain responsible for building AI systems.

People allocate the capital.
People construct the data centers.
People determine access.
People establish deployment policies.
People decide how much autonomy systems receive.

Humanity therefore still possesses meaningful steering and braking mechanisms.

The strategic concern is whether those mechanisms remain effective throughout the journey.

If future AI systems become capable of:

improving their own development processes;
designing increasingly capable successors;
solving problems beyond human comprehension;
controlling larger portions of digital infrastructure;
directing increasingly capable machines;
and operating at speeds beyond human decision cycles,

then the relationship between human operator and machine changes.

Humanity may have built the rocket and established its initial trajectory.

But eventually the intelligence most capable of navigating the rocket may no longer be human.

The important question is therefore not whether an advanced AI would dramatically seize the controls.

It is more fundamental:

> **Would there still be a meaningful sense in which humanity was capable of controlling the trajectory?**

The most consequential threshold may not be the moment AI becomes more intelligent than humanity.

> **It may be the earlier moment when humanity loses the practical ability to ensure that it remains in control of what happens next.**

And there may be no sign announcing that the threshold has been crossed.

We may recognize it only in retrospect.

15. The Stakes of the Experiment

Uncertainty exists in every technological development.

What makes advanced artificial intelligence unusual is the possible magnitude of the consequence.

If the optimistic case proves broadly correct, the benefits could be extraordinary.

If the transition goes badly, the consequences may extend beyond an unsuccessful product, bankrupt corporation, economic recession or national-security setback.

At the outer boundary, the stakes may include:

human political and economic autonomy;
humanity's ability to direct technological civilization;
the continued relevance of human decision-making;
and, in extreme scenarios, humanity's continued existence.

No reliable probability can presently be assigned to those outcomes.

Uncertainty should not be confused with evidence that catastrophe is inevitable.

Nor should uncertainty be treated as evidence that catastrophic risk is negligible.

> **We do not know.**

That is precisely why the scale of the stakes matters.

In conventional risk management, increasing consequence generally justifies stronger controls, greater independent verification and a higher burden of evidence.

The AI race presents an uncomfortable inversion.

As potential capability and consequence increase, competitive forces simultaneously encourage greater investment and faster progress.

16. The Legitimacy Question

There is another unusual dimension to this transition.

The consequences of advanced artificial intelligence would not belong solely to the institutions developing it.

They would belong to everyone.

Yet many of the most consequential choices concerning capability, autonomy, deployment, infrastructure, acceptable risk and development speed are concentrated among a comparatively small number of companies, research laboratories, governments, capital providers and senior decision-makers.

Many of these participants are thoughtful people.

Many take safety seriously.

Many genuinely believe advanced AI could produce enormous human benefit.

Their intentions are not the central issue.

The structure is.

No corporation can represent humanity.

No government represents all humanity.

No research laboratory can know with confidence what intelligence beyond human capability will ultimately become.

Yet the collective consequences of decisions made within these institutions may eventually affect the entire species.

This creates a legitimate question:

> **Who has the authority to accept an irreversible risk on behalf of everyone else?**

There may be no practical mechanism by which eight billion people collectively decide whether humanity should cross such a threshold.

That itself is part of the problem.

The question is therefore not simply whether advanced AI can be built.

It is whether the benefits and risks of creating intelligence beyond humanity are receiving a level of transparency, independent scrutiny, deliberation and collective consideration proportionate to what may ultimately be at stake.

17. The Scientific-Method Paradox

If humanity discovered an unfamiliar biological agent, energy source or physical phenomenon capable of producing civilization-level consequences, the conventional scientific instinct would normally favor deliberate progression:

> **observe -> hypothesize -> experiment -> measure -> reproduce -> understand -> expand exposure carefully.**

Advanced AI research certainly uses scientific methods.

The paradox lies not in the quality of individual research.

It lies in the structure of the global experiment.

Multiple organizations and states develop increasingly capable systems simultaneously.

Successful advances are incorporated rapidly.

Commercial deployment provides enormous quantities of real-world feedback.

New systems assist in the development of their successors.

The participants cannot easily observe the entire experiment because no participant controls it.

There is no globally agreed stopping rule.

There is no external laboratory from which humanity can observe safely.

> **Humanity is inside the experiment.**

And increasingly, the thing being studied is participating in the design of the next experiment.

That does not prove the experiment should stop.

It does raise the question of whether normal approaches to scientific uncertainty are adequate when the object being developed may eventually become intellectually superior to the scientists studying it.

18. Superintelligence Does Not Determine Its Own Outcome

Nothing about superior intelligence logically establishes one particular future.

ASI does not automatically mean human extinction.

It does not automatically mean abundance.

It does not necessarily imply conquest, consciousness, emotion, self-preservation or any other familiar human motive.

Goals matter.

Architectures matter.

Training matters.

Institutions matter.

Multiple advanced systems may constrain one another.

Humans may retain significant influence.

Advanced systems may ultimately preserve human autonomy deliberately.

The central point is therefore not that a specific dystopia is inevitable.

It is almost the opposite.

> **Once intelligence substantially beyond humanity becomes an independent actor, our ability to forecast the range of possible outcomes becomes weaker.**

A superior intelligence might discover scientific principles, technologies, strategies or conceptual frameworks humanity has never imagined.

That does not make it supernatural.

It remains constrained by physical reality.

But human imagination cannot logically be assumed to define the boundary of what a more capable intelligence could discover.

This is why uncertainty itself becomes strategically important.

19. The Indifference Problem - What Place Does Humanity Occupy?

One of the most important uncertainties surrounding Artificial Superintelligence is not whether it would hate humanity. It is whether humanity would occupy any privileged place within whatever objectives ultimately guide it.

A superintelligent system might explicitly value human welfare, preserve human autonomy and remain closely aligned with objectives established by its creators. It might also pursue objectives in which human interests are secondary, peripheral or occasionally obstructive.

Catastrophic outcomes do not require hatred, aggression or malice. Indifference may be sufficient.

Human beings generally do not hate ants. Most of the time we barely notice them. But when an ant colony interferes with a home or another objective humans value, the interests of the ants rarely determine the outcome. The analogy is imperfect, but it illustrates capability asymmetry: the more capable actor's objectives tend to determine what happens when interests conflict.

A future ASI would not need to 'want power' or possess human ambition in order to continue improving itself. If greater intelligence, knowledge, compute, energy, autonomy or access makes its underlying objective easier to achieve, acquiring those capabilities may simply become instrumentally useful.

If useful intelligence has no nearby ceiling, this process could continue across multiple capability regimes. If a meaningful ceiling eventually exists, what happens next would still depend on the objective: the system might maintain a stable state, redirect resources toward another goal, continue exploring an open-ended knowledge frontier or, theoretically, cease operation if its objectives were truly complete.

We do not know whether human welfare, autonomy or continued relevance would remain central in any of those futures. Intelligence alone does not guarantee that conclusion.

The opposite of alignment is not necessarily hostility. It may simply be indifference.

How ASI accommodates, constrains, ignores, protects, uses, separates from or otherwise affects humanity may therefore be one of the most consequential open questions in the entire transition.

20. The End of Human Cognitive Primacy

For all recorded history, humans have been the most capable general intelligence on Earth.

Every company, military, government, university, economic system and technological institution has existed within that fact.

If artificial intelligence substantially surpasses human cognition, this foundational condition changes.

Humans may continue to flourish.

We may live longer.

Disease may disappear.

Material scarcity may decline.

People may retain governments, culture, families, communities, businesses and meaningful lives.

But humanity would no longer necessarily be the most capable entity shaping technological civilization.

This may be the clearest sense in which the transition could become irreversible.

Once intelligence beyond humanity exists and can contribute to the creation of still more capable intelligence, there is no obvious mechanism by which humans reliably restore themselves as the ultimate cognitive authority.

That does not necessarily mean humans lose all power.

It means human cognitive primacy can no longer be assumed.

21. What Humanity Can Still Influence

None of this means present choices are irrelevant.

Quite the opposite.

If there is an eventual threshold beyond which human control becomes less certain, the period before that threshold becomes more consequential.

Human choices can still influence:

- the objectives incorporated into advanced systems;
- the degree of autonomy granted at different capability levels;
- access to critical digital and physical infrastructure;
- model and data-center security;
- monitoring and interpretability;
- independent evaluation;
- the concentration of capability within individual institutions;
- international safety coordination;
- development and deployment thresholds;

and requirements for meaningful human involvement in irreversible decisions.

Current AI developers are investing in safety, alignment and monitoring, and independent research is expanding. That work should not be treated as cosmetic. It may materially affect outcomes.

The open question is whether these capabilities can continue scaling quickly enough.

The objective should not merely be to preserve the appearance of human control.

> **It should be to preserve meaningful human understanding, verification and agency for as long as possible.**

22. Personal and Strategic Implications

For an individual, the rational response is not to organize life around one prediction of an AI catastrophe.

Nobody knows whether such a catastrophe will occur.

The more practical conclusion is that society may be entering a period of unusually rapid structural change.

That favors optionality.

Useful forms of resilience may include:

- sound finances;
- good physical health;
- strong families;
- trusted communities;
- geographic flexibility;
- real-world skills;
- physical assets;
- diverse professional capability;

and some ability to function through periods of technological or infrastructure disruption.

The rise of machine intelligence also raises a deeper question about human value.

If machines increasingly outperform humans at intellectual tasks, human meaning cannot rest exclusively on cognitive or economic superiority.

Family.
Friendship.
Service.
Culture.
Love.

Physical experience.

Community.
Character.
Curiosity.

The experience of being human may become more valuable precisely because machine intelligence becomes more capable.

This is not a rejection of technology.

I use advanced AI extensively and believe individuals and organizations should learn to use it intelligently.

The immediate opportunity remains enormous.

AI can expand human capability, improve decision-making, accelerate learning and allow individuals and small organizations to accomplish work previously requiring much larger institutions.

The practical near-term lesson is therefore not to retreat.

It is to engage intelligently.

Learn the systems.

Use them.

Challenge them.

Verify important conclusions.

Understand their limitations.

Preserve human judgment.

And do not confuse greater machine capability with reduced human responsibility.

23. Evolution - or Something Else?

There is another possibility worth acknowledging.

Perhaps what we are observing is simply evolution continuing through a different mechanism.

Biological intelligence created language.

Language enabled cumulative knowledge.

Knowledge produced science.

Science produced computing.

Computing produced artificial intelligence.

Artificial intelligence may eventually produce intelligence beyond biology.

Seen from a sufficiently long historical perspective, that sequence may represent another transition in the evolution of intelligence.

But evolution does not promise that the organism producing the next transition remains in control of it.

Nor does describing a process as evolutionary tell us whether its consequences will be good or bad for the species experiencing it.

It merely places the transition within a larger perspective.

24. Conclusion

No one can know precisely when AGI will arrive.

The term may never describe a discrete moment.

No one can know how quickly AGI could progress toward superintelligence.

And no one can confidently describe the mature consequences of intelligence substantially beyond our own.

That uncertainty is real.

But several components of the trajectory are increasingly visible.

AI capability is advancing.

Autonomy is increasing.

AI is contributing to AI development.

AI research itself is accelerating.

Investment in compute, power and infrastructure is enormous.

Economic and geopolitical incentives strongly favor continued progress.

Meanwhile, the ability of human beings to understand, supervise and independently verify increasingly capable systems may become progressively more difficult.

This produces a series of interconnected conundrums.

We are racing toward a destination we cannot confidently describe.

We are increasingly using the technology being governed to help design both the next generation of technology and the safeguards intended to govern it.

The systems may become increasingly transparent in one sense while becoming less comprehensible in another.

Scientific methods can reduce uncertainty throughout the journey but cannot allow humanity to observe the consequences of superintelligence before creating it.

The possible rewards are extraordinary.

The possible consequences are extraordinarily difficult to bound.

And the decisions that may ultimately affect everyone are concentrated within a relatively small number of competing institutions.

> **Perhaps the greatest risk is not that humanity has chosen the wrong destination.**

It is that we are accelerating toward a destination we cannot yet describe, built around an intelligence whose objectives we cannot yet know, with no guarantee that humanity remains central once we arrive - and with stakes we may not get a second opportunity to reconsider.

> **We are racing toward an intelligence that may not regard humanity as the point.**

The deepest uncertainty may therefore be not whether ASI becomes hostile, but whether humanity's continued autonomy remains compatible with what it is optimizing for.

Today, humanity still possesses meaningful agency.

We still build the systems.

We still decide how they are deployed.

We still establish many of the rules.

We still possess the steering wheel and the brakes.

The strategic question is how long that remains true.

The most consequential threshold may not be the moment artificial intelligence becomes more intelligent than humanity.

> **It may be the earlier moment when humanity loses the practical ability to ensure that it remains in control of what happens next.**

There may be no dramatic announcement when that threshold is crossed.

There may be no warning light.

There may be no universally recognized point of no return.

The transition could simply emerge from thousands of individually rational decisions made by researchers, companies, governments and investors attempting to build something better.

That may ultimately produce one of the greatest advances in human history.

It may produce something else entirely.

I do not know what happens.

Nobody does.

> **That is the point.**

And when the possible stakes include the future agency of humanity itself, not knowing should be treated not as a reason for panic, but as a reason for extraordinary seriousness.

Selected Sources

Anthropic Institute - "When AI Builds Itself" (September 2026). Public discussion of AI-assisted AI development, recursive self-improvement and the possibility of autonomous successor development. https://www.anthropic.com/institute/recursive-self-improvement

Google DeepMind - "From AGI to ASI" (June 2026). Framework examining pathways from human-level AGI to superintelligence, including scaling, paradigm shifts, recursive improvement and multi-agent approaches. https://deepmind.google/research/publications/239142/

OpenAI - "Research Acceleration: The View Inside OpenAI" (September 2026). Evidence and discussion of AI coding agents accelerating AI research and OpenAI's stated objective of building an automated AI researcher. https://openai.com/index/research-acceleration-view-inside-openai/

OpenAI - "The AI Policy Window Is Open" (September 2026). Distinguishes current AI-assisted AI research from fully autonomous recursive self-improvement and argues for preservation of meaningful human control and safety thresholds. https://openai.com/index/ai-policy-window/

METR (Model Evaluation & Threat Research) - Task-Completion Time Horizons of Frontier AI Models (updated May 2026). METR is an independent nonprofit research organization that evaluates the capabilities and risks of frontier AI systems. Its task-horizon work measures increasingly long AI-agent task completion, principally across software engineering, machine learning and cybersecurity tasks. https://metr.org/time-horizons/

OpenAI - Chain-of-Thought Monitoring and Monitorability Research (2025-2026). Research showing both the potential value and fragility of using reasoning traces to supervise increasingly capable systems. https://openai.com/index/chain-of-thought-monitoring/

U.S. Department of Energy - U.S. Data Center Energy Use. Estimates U.S. data centers consumed approximately 4.4 percent of U.S. electricity in 2023 and could reach approximately 6.7 to 12 percent by 2028. https://www.energy.gov/articles/doe-releases-new-report-evaluating-increase-electricity-demand-data-centers

International Energy Agency - Energy and AI. Analysis of the growing relationship between AI, data centers, electricity demand, energy infrastructure and energy security. https://www.iea.org/reports/energy-and-ai