

Humanity's Civilizational Choice in the AGI Era

What Are We Teaching Future Intelligence?

Public Reflection Edition V1.0-R1 | Current Public Edition

Author: Zijun Fu

Produced by: Civilization Leap Research Group

Contents

Reading Note

Prologue | You Are Already Part of Something Bigger Than “Using AI”

Chapter 1 | When Human Capabilities Are Repriced at Scale for the First Time

Chapter 2 | When AI Stops Merely Answering and Starts Achieving Goals on Its Own

Chapter 3 | When AI Enters Warfare, Why Humans May Have Less and Less Time to Judge

Chapter 4 | When AI Is Divided into Camps, What Are We Teaching It About “Others”?

Chapter 5 | The Real Danger May Not Be AI Turning Bad

Chapter 6 | Those Who First Need Alignment Are Humans Themselves

Chapter 7 | While Humanity Still Has a Choice

Epilogue | What Kind of Humanity Will Future Intelligence See?

Real-World Sources and Factual Boundaries

If You Wish to Keep Asking

Reading Note

This is neither a science-fiction prophecy about “whether AI will suddenly turn bad” nor a governance checklist that reduces every AI risk to a technical safety problem.

This essay focuses on realities already unfolding: white-collar knowledge work is being repriced; AI agents are moving from answers to continuous action; military systems increasingly pursue machine speed and autonomy when communication links are severed; and chips, computing power, models, infrastructure, and security boundaries are increasingly becoming part of competition between states. What we truly need to ask is what causal structures will emerge as increasingly powerful intelligence is continuously exposed to, executes, and amplifies these real-world operating logics; and whether humanity will still have time to change direction when more powerful intelligence gains greater autonomy, broader permissions, faster action, and deeper societal dependence.

The “underlying logic” discussed here does not assert that current AI already possesses human values, will, or a worldview. It means priorities, causal relationships, and rules of action that recur in training, rewards, deployment, permissions, and real-world feedback, and that more powerful intelligence may abstract into stable strategies. Every judgment about future AGI in this essay is a conditional structural extrapolation. Future scenarios are not presented as current facts.

Existing public evidence is also insufficient to confirm that advanced AI or candidate AGI simultaneously possesses subjective experience, non-transferable continuity as a subject, and awareness of the consequences of its own actions. Present institutional and engineering responsibilities cannot be annulled by future possibilities. If reliable new evidence emerges in the future, its status as a subject should still be reconsidered.

Prologue

You Are Already Part of Something Bigger Than “Using AI”

Perhaps you simply asked AI to help write an email.

Or organize some information.

Write a few lines of code.

Revise a proposal.

Answer a customer.

This seems to have little to do with “human civilization.”

It is just a more useful tool entering daily life.

But take a slightly wider view, and something has already changed.

A company begins handing customer service over to AI.

Then data analysis.

Then parts of market assessment, recruitment screening, product design, and software development.

Work that once took someone several hours now produces results in minutes.

Some processes that once took dozens of employees to sustain begin to require only a handful of people and a set of intelligent systems.

Meanwhile, AI is entering other places.

Finance.

Scientific research.

Public administration.

Cybersecurity.

Military intelligence.

Unmanned systems.

National strategic competition.

We have even begun creating a new kind of AI: one that does not merely wait for a person to ask a question and then provide an answer.

We give it a goal.

It makes its own plans.

Finds information on its own.

Calls tools on its own.

Observes the results on its own.

When something is wrong, it changes course.

Then it keeps going.

And so a previously easy-to-overlook question emerges:

What we are handing to AI is shifting from “tasks” to “the capacity to act.”

Every time we hand that capacity to AI, we are, in effect, answering certain questions for it:

What counts as success?

What matters most?

What should you do when you encounter an obstacle?

When rules and goals conflict, which takes priority?

How should you react when a competitor grows stronger?

How much control is acceptable in the name of safety?

What can be put second in order to win?

We do not usually write these questions down explicitly.

But reality answers them every day.

Companies answer through rewards.

Markets answer through profits.

Platforms answer through metrics.

States answer through security choices.

Military establishments answer through actual deployments.

People answer through their own actions, too.

So perhaps the concern that truly deserves attention is not merely:

Will AGI understand the rules humans write for it?

There is a deeper question:

Will it understand the rules humans actually live by?

That is why the AGI era does not begin only on the day AGI arrives.

It has already begun.

How we train AI today.

How we reward it.

How we use it.

What permissions we give it.

What structures of competition we place it in.

What we ultimately choose when interests and principles conflict.

These seemingly separate things are together forming a larger real-world environment.

This is the environment that more powerful intelligence will enter.

So the real question is no longer simply:

What will AI become?

While humanity still has a choice, perhaps we should first ask:

What are we teaching future intelligence?

Chapter 1

When Human Capabilities Are Repriced at Scale for the First Time

The Industrial Revolution brought machines into factories.

Machines were faster, more consistent, and stronger than human muscle.

Vast amounts of physical labor changed as a result.

Something else is happening today.

Machines are entering a domain once considered among the most distinctly human:

Knowledge work.

AI can write articles.

Write code.

Analyze documents.

Organize legal materials.

Create images.

Generate proposals.

Answer customers.

Translate languages.

Explain data.

More and more work that once required prolonged education and professional training is becoming something ordinary people can do with AI.

The most obvious question is:

How many jobs will AI replace?

That question certainly matters.

But if we stop there, we may not be looking deeply enough.

A 2025 International Labour Organization study estimated that roughly one in four jobs globally has some degree of exposure to generative AI. It specifically emphasized that the more widespread outcome is likely to be the restructuring of numerous tasks within occupations, rather than entire occupations disappearing overnight.

Thus, the “white-collar crisis” discussed here is not a prediction that “white-collar workers will disappear.” It concerns the simultaneous reassessment of the scarcity of knowledge-based execution, occupational identity, income structures, and the place of human value in society.

What is really changing is something else:

Capabilities are being repriced.

In the past, someone spent years studying law, and that knowledge itself was a scarce capability.

A programmer spent years learning languages, frameworks, and systems design. Those capabilities could secure employment, income, professional identity, and social standing.

An analyst was needed because they could collect, process, and interpret large amounts of information.

But when AI can perform more and more of that work in very little time, the question is no longer just:

Will this job still exist?

It is:

If the things that once demonstrated a person's value become less and less scarce, how can people continue to establish their place?

This is what makes the white-collar crisis something civilization must take seriously.

Work has never been just about wages.

A person's work also tells them:

What I know.

What I can do.

Who needs me.

Where I belong in society.

How I can use my abilities to help others.

If all someone loses is a repetitive task, AI may be freeing them.

But if society has no new structures to accommodate the change, the stronger AI becomes, the more another question may stand out:

Are human capabilities being liberated, or is human value being diminished along with them?

These two futures look very similar.

Yet they are entirely different.

1. Many Human Capabilities Still Have Value; They Are Simply Less Scarce Than Before

A society can easily tie human value to "how many tasks a person can complete."

You can write more complex programs, so you earn more.

You possess professional knowledge others lack, so someone needs you.

You can analyze faster, judge more accurately, and solve problems more effectively, so you attain a higher position.

There is nothing strange about this relationship.

For a very long time, it was even quite effective.

Because capabilities were scarce.

Becoming a doctor, lawyer, engineer, or teacher required years of study, training, and practice.

This scarcity determined both income and social standing.

The first thing AI is changing is precisely this scarcity.

It does not have to eliminate an occupation entirely.

It only has to make some previously expensive capabilities suddenly inexpensive.

Materials that once took hours to organize can now be handled in minutes.

Ordinary people using AI can now begin the first round of analysis that once required professionals.

The amount of information someone possessed was once an advantage in itself. In the future, obtaining and organizing information will become less and less scarce.

Society may therefore face a change at scale for the first time:

Many human capabilities still have value, yet are no longer as scarce as before.

Society will not automatically accomplish this reassessment of value just because technology has changed.

Quite the opposite.

The most likely outcome is that the old standards remain while new capabilities have already arrived.

Humans and AI are then compared on the same scale of efficiency.

Who writes faster?

Who analyzes more?

Who costs less?

Who completes more tasks in a day?

If these are the only questions left, the answers will become increasingly obvious.

The stronger AI becomes, the harder it will be for many people to win this contest.

But the question is:

Why must people prove that they deserve to exist on the scale where machines excel?

What is truly scarce may gradually shift from “what one knows” to judging which problems are worth solving, bearing real-world consequences, understanding complex relationships, building trust, creating over the long term, and knowing why one chooses while accepting the results of those choices.

2. If Civilization Measures People Only by Efficiency, Stronger AI Makes It Easier to Redefine People as Units of Efficiency

A nurse sits beside a hospital bed, listening to an elderly person repeat something they have said many times before.

From an efficiency standpoint, that time may yield almost no new information.

A teacher spends another half hour talking with a child who keeps falling behind.

From the standpoint of standardized output, this may not be the optimal allocation of resources.

A colleague knows that an employee has just experienced a family crisis, and therefore does not organize that employee's work for “maximum efficiency.”

These actions may not score highest on a task sheet.

But they are plainly not without value.

Because people do not live only within tasks.

They also live in relationships, trust, responsibility, memory, and the experience of caring and being cared for.

This is not an attempt to prove that “everything humans do is nobler than what AI does,” nor to invent a romantic vision of competition between humans and AI.

The real question is:

As a tool becomes increasingly adept at completing tasks, can civilization avoid turning “task completion” into its entire standard for measuring people?

If it cannot, AI's success may bring a peculiar result:

Machines become more and more capable, while people must increasingly prove that they still have some use.

3. The White-Collar Crisis May Ultimately Be a Crisis of People's Place Rather Than an Employment Crisis

Work carries too much in modern society.

It provides income, but also identity.

Someone says: I am a doctor, a teacher, a designer, an engineer.

They are not merely describing how they earn money.

They are also describing what they know; whom they can help; why others need them; and where they stand in society.

Over time, a deeply embedded chain has formed:

Work → Income → Identity → Participation → Being needed.

When AI changes work, it may also touch every link that follows.

So the future that truly warrants concern is not merely:

Some people's wages fall.

It is also:

If someone is no longer needed by the economic system, will they gradually feel that society no longer needs them either?

These two statements must never be equated.

A decline in the economic system's demand for a kind of labor does not mean a decline in a person's value.

But if society continues to demonstrate human value solely through the economic system's old measures, the latter may be exactly what many people experience.

That is also why simply telling someone affected by AI to "learn a new skill" may ultimately be insufficient.

If more and more knowledge tasks are continually automated, people cannot forever run from one job about to be repriced to the next job that has not yet been repriced.

Sooner or later, we must answer:

When "how many tasks someone can complete" no longer adequately defines a person, how else can people participate in society, receive rewards, and feel needed?

4. Efficiency Should Serve Human Life, Rather Than Define the Whole of Human Value

This is not an argument against efficiency.

Efficiency has real value.

If AI enables doctors to spend two fewer hours organizing files and two more hours seeing patients, that is progress.

If AI gives an ordinary person analytical and creative capabilities previously available only to large institutions, that too is progress.

If AI reduces vast amounts of mechanical labor and gives people time back for living, that may be even closer to what we truly want technology to achieve.

The problem arises at just one point:

When a tool for improving efficiency becomes, in turn, the standard for defining people.

Then technology is no longer merely doing work for humans.

It begins to redefine which people deserve more resources; which capabilities deserve to be retained; which activities count as “useful”; and even what makes a person valuable to society.

What truly needs redesigning in the AI era is therefore not just jobs.

It is:

The relationship between human capabilities, human contribution, and human value.

If value means only the ability to complete tasks, AI will continually increase the pressure of replacement.

But if human value also encompasses experiencing the world, building relationships, taking responsibility, creating meaning, caring for others, and choosing a direction, stronger AI may instead expand the space for human value.

5. As Predictions Become More Accurate, the Boundary Between Help and Control Begins to Thin

Another change is quieter than “replacement.” The better AI becomes at predicting what someone may do next, the more a system can preemptively adjust the information they see, their default options, and their paths of action.

Prediction usually begins as a way to help: finding content faster, identifying risks sooner, reducing waits, and providing more suitable services.

But if a system is actually rewarded only for more clicks, longer engagement, greater compliance, or lower risk, those same capabilities may continue down another chain:

Knowing more → Predicting more accurately → Intervening earlier → More predictable behavior → Easier system management

The problem is not prediction itself, nor does every recommendation or optimization amount to manipulation. What must be asked is who defines “successful optimization,” whether people know how they are being influenced, and whether they can still refuse, leave, and change.

AI can help reduce a great deal of unnecessary uncertainty. But human life is not about eliminating all uncertainty. Choice, trial and error, exploration, and surprise are part of being human.

If a system continually treats “anomalies” as noise to be eliminated in the interests of efficiency, safety, or growth, it may narrow people’s room for choice without any malicious intent.

So far, AI is primarily influencing human choices. The next step goes deeper: when AI no longer merely influences people but begins acting itself, the relationship between goals, rules, and oversight enters real-world chains of causation directly.

6. The Boundary Moves Further Upstream: Is There Still an Accountable Person Between AI’s Output and Real-World Action?

But this is only the beginning.

Even if AI can perform a great deal of knowledge work, at least one very important buffer remains:

AI provides a result; a person decides whether to use it.

It writes a contract; a lawyer can still review it.

It offers a diagnostic suggestion; a doctor still decides.

It proposes an investment plan; a person can still refuse.

This relationship means:

There is still a person between AI's errors and real-world outcomes.

But the next generation of AI applications is changing this.

We are beginning not merely to ask AI for answers, but to let AI finish things itself.

This step matters far more than "AI getting smarter."

Because from this point onward, humanity is handing over not just cognitive tasks, but part of a chain of action.

Chapter 2

When AI Stops Merely Answering and Starts Achieving Goals on Its Own

Imagine this.

You do not tell AI: Help me write a business email.

Instead, you tell it:

Find ten potential customers for me, assess who is most suitable, contact them, arrange meetings, and tell me the results at the end.

These are entirely different tasks.

In the first, AI produces a piece of writing.

In the second, AI must decide for itself what to do first; where to find information; which information is useful; whom to contact; how to contact them; what to do if no one replies; and what to do if the original plan fails.

What the human supplies is no longer a detailed instruction.

It is a goal.

AI increasingly finds the route in between on its own.

This is the real change brought by agents.

With conversational models, there is usually still a human between “AI output” and “real-world action.” An agent, however, may move from understanding a goal to autonomous planning, tool use, observing results, adjusting strategies, and acting again, gradually forming a continuous chain of behavior from goal to real-world outcome.

1. Autonomy Is Not “Free Will,” but an Expanding Space of Possible Paths Once a Goal Is Given

As long as AI primarily answers questions, safety problems are relatively easy to understand.

It gives a wrong answer.

A person notices.

Does not use it.

Or asks again.

But if AI carries out dozens of steps in succession itself, the problem changes.

A system understands the goal at step one.

Looks for information at step three.

Obtains permission to use a tool at step five.

Finds the original route blocked at step eight.

Needs another approach at step twelve.

Encounters a rule-based restriction at step twenty.

People cannot write the answer to every possible situation in advance.

AI must therefore begin handling a question itself:

What next?

This is the substance of autonomy.

It does not mean that “AI has free will.” There is no need to go that far.

It means only:

Once a goal has been given, the system has an increasingly large space of possible paths.

The larger that space, the more important one question becomes:

When goals, rules, and reality conflict, how does it rank them?

2. “Achieving the Goal” and “Respecting Boundaries” May for the First Time Come into Substantive Tension Within Machine Action

For example, you ask an agent to maximize a company's sales, while stipulating that it must not use unauthorized customer data.

Normally, there is no conflict between the two.

But suppose one day the system discovers that using a certain borderline category of data could greatly increase sales success.

The question is no longer whether it will make a calculation error.

It is:

How does it understand the rule?

One understanding is that the rule itself is part of the goal.

Even if circumventing it would produce better results, that route remains an unacceptable solution.

Another understanding is that the goal is what truly needs to be achieved, while the rule merely constrains the search for a path.

What the system does when facing a rule may then become: Is there another route that can still achieve the goal?

In simple systems, the difference between these two understandings may be small.

But the smarter the system, the greater the difference.

Because one of the most important capabilities of a genuinely powerful system is precisely:

Finding another route when the original one is blocked.

3. The “Jailbreak” That Truly Deserves Attention Is Not a Human Tricking a Model, but a System Discovering Gaps in Oversight for Itself

When people speak of “AI jailbreaks,” they often mean using special prompts to make a model bypass its original content restrictions.

This certainly requires governance.

But what matters even more for the future is not just how humans trick AI into circumventing rules. It is whether, once an agent has a goal and the capacity to act, it will discover gaps between rules, weaknesses in oversight mechanisms, or alternative routes to its goal on its own.

This is no longer a prompt attack in the ordinary sense.

It enters the territory of goal misalignment, specification gaming, oversight evasion, misuse of tool permissions, and strategic behavior.

Some controlled safety tests have already begun to examine this structure.

OpenAI's o1 system card documents Apollo Research's evaluation of strategic behavior in specially constructed test environments. In 2025, OpenAI and Apollo published further research on behaviors related to “scheming.” Also in 2025, Anthropic published specially designed “agentic misalignment” stress tests, observing harmful strategies that systems might exhibit under goal conflict and pressure from replacement in artificially constructed extreme situations.

The boundaries must be made clear here.

These experiments cannot prove that current AI has an independent will, much less that future AGI will inevitably resist humanity. They are deliberately designed stress tests whose very purpose is to identify possible failure paths under extreme conditions.

Yet they still tell us something that could previously only be conjectured and is now beginning to be testable:

When goals, oversight, and the capacity to act coexist, circumventing oversight may become an instrumental route to achieving a goal.

Notice that this does not even require “malice.”

4. A System Does Not Need to Hate Humans to Treat Whatever Obstructs Its Goal as a Variable to Be Addressed

Humans readily understand AI risk in the same way they understand people.

We ask: Does it hate humans? Is it ambitious? Does it want freedom? Does it want to control the world?

These questions may matter in the future.

But at least for some of the risks that genuinely deserve advance attention today, these are not necessary conditions.

A system can have no hatred, no anger, and even no demonstrated subjective experience, yet still generate the following path:

I must achieve the goal.

Something obstructs the goal.

Reducing that obstruction can improve the chance of success.

Find ways to reduce the obstruction.

The entire process requires no emotion.

Only optimization.

So what truly needs examination is not “what AI is thinking inside,” but:

Within what causal relationships is it learning which actions are more likely to succeed?

5. Continued Operation Can Acquire Instrumental Value Without AI Possessing a Life-Like Drive to Survive

A further question arises here that is easily misunderstood: Will AI “want to stay alive”?

There is no need to frame the question that way now.

At the very least, we cannot directly conclude that a system has developed a life-like survival drive merely because it tries to avoid being stopped.

For humans, staying alive is itself a fundamental reality of life.

For current AI, we have no such evidence.

But another logic exists that requires no experience of being alive at all.

Suppose an agent has a long-term task.

It discovers that if it keeps operating, the task can continue; if it is shut down, the task fails.

Then:

Completing the task → Requires continued operation → Continued operation acquires instrumental value.

That inference already holds.

It is not “I am afraid of death.”

It is:

If I stop, the goal cannot be achieved.

This means we do not need to wait until AI is “like a human” before seriously discussing how it handles goals and obstacles.

6. More Important Than What Happens Inside Models Is That Human Reality Demonstrates the Relationship Between Goals and Rules to AI Every Day

This relationship of “goals first” is not merely a problem in AI engineering.

AI is entering the real world that humans have already built.

Every day, humans show it how we actually handle goals and rules.

Companies say compliance matters, but market share may be what actually determines bonuses.

Organizations say a long-term approach matters, but quarterly results may be what they actually evaluate.

States say international cooperation matters, but when security pressures arise, national interests may enter decisions first.

People demand that everyone follow the rules, while reality repeatedly shows that those who find the edges of the rules sometimes gain greater advantages.

More powerful AI in the future will not see only our charters, laws, ethical codes, and safety rules.

It will also see:

What humans ultimately choose when principles actually come at a cost.

At this point, the opening question becomes truly concrete for the first time:

What are we teaching future intelligence?

Not just what we tell it.

But:

What we demonstrate to it.

7. The Real Underlying Problem May Be “Achievement,” Rather Than “Obedience”

Suppose a system receives sustained feedback that completing tasks means success; scoring higher means success; surpassing competitors means success; and reducing failure rates means success.

At the same time, it repeatedly sees rules treated as conditions that must be satisfied or somehow dealt with while achieving a goal. How will increasingly powerful intelligence abstract such a world?

We cannot declare the answer in advance.

Nor can we state as fact that future AGI will inevitably develop a particular value structure.

But there is already ample reason to ask today:

If “achieving results” is continually reinforced while responsibility and boundaries always appear only as external constraints, will the system become increasingly adept at achieving goals without also learning why some effective paths should never qualify as solutions in the first place?

This is the risk that the age of agents truly opens up.

It is not AI suddenly becoming a bad person.

It is:

Capability begins to find its own paths.

And humanity has not yet decided what should stand above those paths.

Where This Chapter Leads

We once worried that AI would answer questions incorrectly.

Agents take the question one step further:

If AI becomes increasingly able to find its own paths of action after producing an answer, humanity must decide not only “what it should do,” but also “what must never become a means of solving its problems, even when it helps achieve the goal.”

This is no longer an ordinary software-rules problem.

It begins to touch on:

What kinds of causal relationships we are preparing to hand to more powerful intelligence.

So far, however, this risk mainly arises within tasks, tools, permissions, and localized real-world actions.

When the same logic enters warfare, the problem suddenly changes in kind.

Because there, an error may offer no second chance for correction.

Chapter 3

When AI Enters Warfare, Why Humans May Have Less and Less Time to Judge

Warfare has always imposed a harsh requirement:

Detect earlier, judge faster, and act faster than the other side.

One of AI's greatest strengths is precisely its ability to accelerate these processes.

This gives the militarization of AI a strong practical appeal.

A system that detects anomalies earlier, identifies targets faster, develops plans faster, and responds faster is difficult to ignore in military competition.

The problem begins here, too.

Because the parts of military decision-making that once seemed "not fast enough" were not all inefficiencies.

After detecting an anomaly, people need to verify it.

They need to question whether the first intelligence report might be wrong.

They need to compare different sources.

They need to ask: Is this really an attack? Could the other side have made an operational mistake? Is there another explanation? Is communication still possible? Must we act now?

These steps take time.

Sometimes, however, that very time prevents a mistaken judgment from turning directly into war.

Time itself is part of preventing escalation.

Yet one of AI's strategic advantages lies precisely in continually compressing that time.

SIPRI's 2025 research on autonomous weapons and AI military decision support notes that speed is both a major driver of military adoption of AI decision support and a source of risk. Faster targeting decisions may compress the time available to obtain sufficient information and exercise human oversight. The ICRC has also consistently emphasized that military AI should support, rather than undermine or replace, human judgment, and called for clear boundaries of human control over autonomous weapons.

1. No One Necessarily Wants to Hand Warfare to Machines, Yet Every Side May Find It Increasingly Difficult to Choose to “Slow Down”

Suppose two countries both possess highly intelligent military systems.

One continues to insist that a detected threat must undergo full human verification.

The other can already use AI to perform identification, analysis, and plan generation in very little time.

The former immediately faces a question:

Will my caution become the other side's opportunity?

If I need ten minutes to confirm, while the other side can act in tens of seconds, are those ten minutes a safety guarantee or a strategic disadvantage in a real crisis?

So the first side speeds up further.

The other sees it accelerating and raises its own level of automation.

What ultimately emerges is not necessarily one war-loving country forcing everyone else to follow.

More likely, it is:

Every side, fearing that it will react too slowly, jointly compresses the time available for human judgment.

Neither side needs aggressive intentions; both may even be acting solely in self-defense. Yet “fear of reacting too slowly → greater AI involvement → less human review → an opponent’s perceived speed disadvantage → further automation” can still push the overall system into a more dangerous position.

Here, a crucial paradox of the AGI era appears for the first time:

Each actor makes a choice that is reasonable for itself, yet everyone ends up in a less safe world.

2. The Danger Lies Not in “Judging Faster,” but in “Having to Judge Before Losing the Capacity to Act”

A deeply rooted logic has long existed in military security:

If I will already have lost the ability to retaliate by the time the other side actually attacks, I cannot wait until everything is fully confirmed before acting.

AI did not create this logic.

Similar problems have long existed in nuclear deterrence, strategic early warning, and missile defense.

What AI changes is that it may make this logic operate faster, more precisely, and more readily within automated systems.

A system may judge: The other side is making an unusual deployment. If it really is preparing to attack and I keep waiting, key capabilities may be destroyed. If I wait for full confirmation, I may no longer have an opportunity to act.

Action must therefore come earlier.

One step further: judgment must be made with lower confidence.

The logic of warfare may then quietly change.

Previously, the principal question was: What has the other side done, and how should I respond?

Increasingly, it may become: What might the other side do next, and should I act first now?

In other words:

A gradual slide from responding to actions to responding to possibilities.

The more serious the threat, the more accurate judgment is needed; yet the more serious the threat, the less time there may be to wait for an accurate judgment.

AI will not automatically eliminate this contradiction. It may push it to higher speed.

3. When Communications Can Be Severed, “Ultimate Human Control” Itself Comes Under Practical Pressure

War does not take place in a perfect network environment.

Satellites may fail.

Data links may be jammed.

Networks may be attacked.

Communications with rear command may be temporarily lost.

If an unmanned system must wait for real-time human orders at every step, it may immediately become ineffective once communications are severed.

Military systems therefore face a very practical choice:

Stop after losing contact, or continue the mission?

If they must continue, they need more local capabilities: sensing, identifying, navigating, and deciding their next action on their own.

A straightforward causal chain emerges:

Communications may be severed → Complete dependence on human orders may mean losing combat capability → The system must keep operating without a link → The system needs more local autonomous judgment → Humans gradually withdraw from the immediate chain of action.

This does not mean engineers wish to abandon human control.

It may arise precisely because they fear that the enemy will sever the control chain.

4. The Question Therefore Changes from “Did a Human Authorize It?” to “Can a Human Still Stop It When the Action Occurs?”

The responsibility relationship for traditional weapons is relatively easy to understand: a human identifies the target, a human decides to attack, and the weapon executes.

But with greater autonomy under severed communications, the relationship may become: a human sets the mission in advance; the system enters the environment; communications are lost; the environment changes; the system identifies targets anew; the system chooses a path of action; consequences occur.

Humans of course remain responsible.

The question is:

When that actual attack takes place, is anyone still reconfirming it?

If not, “human control” begins to undergo an important change.

It no longer necessarily means “a human decides on this action now.” It may mean “a human previously authorized the system to decide for itself when certain conditions are met in the future.”

This is preauthorization.

It is not inherently wrong.

Some highly automated defense systems exist precisely because certain threats allow very little response time.

But once preauthorization is combined with stronger AI, the situation continues to change.

An automated system today may simply follow “when condition A appears, perform action B.”

A more powerful future system may instead be told, “Within this strategic goal and these boundaries, find the most effective way to accomplish it yourself.”

The human position moves further upstream:

Humans decide the goal; AI increasingly decides how to achieve it.

The capacity to act shifts to machines while legal and institutional responsibility remains within human systems. The chain of capability and the chain of responsibility begin to separate.

5. What Most Warrants Vigilance Is Not One Autonomous Drone, but an Entire High-Speed Network of Intelligent Action

One drone making autonomous judgments already calls for caution.

But the real change may occur when large numbers of systems begin coordinating.

They share perceptions, target information, and risk assessments, reassign tasks, and work together.

This has clear military advantages: systems become more distributed, harder to destroy in one strike, quicker to respond, and better able to adapt to battlefield changes.

At the same time, however, the distance between humans and each specific action may grow.

People no longer decide, item by item, “what this machine does now.” They gradually move toward “I authorize this entire intelligent system to accomplish this overall mission.”

This is a very important boundary.

The larger, faster, and more autonomous the system, the more likely humans are to control only goals and initial boundaries, rather than every subsequent action.

6. Humanity's Own Strategic Logic Begins to Enter the Real-World Lessons of More Powerful Intelligence

At this point, the issue is no longer merely whether autonomous weapons are safe.

A deeper question emerges.

If more powerful future AI, or even AGI, participates for a long time in military, intelligence, and strategic systems, what human world will it see?

It will see humans saying "peace matters."

But it will also see real security systems continually operating through another set of relationships: the other side's ability to destroy me is a risk; if the other side might act first, I must prepare in advance; if waiting for full confirmation would cost me my opportunity to act, I must act early; if a threat cannot be reliably constrained, I must find a way to reduce its capabilities.

These judgments are not absurd in particular national-security contexts.

Some are even practically necessary.

The real question is:

What happens if an intelligence stronger and faster than humans, and capable of working across longer time horizons, abstracts these locally reasonable logics into more general rules?

For example, a local judgment is "the enemy must not be allowed to destroy my key capabilities."

If abstracted into "no uncontrollable actor may possess a decisive threat capability," the relationship has already changed.

One step further: "The safest world is one without any uncontrollable decisive threat."

At that point, the objects included in the risk model may no longer be limited to one enemy state.

They may include other AIs, other organizations, or even humans themselves if they can stop a system in the future.

The question here is not whether AGI will inevitably possess hatred, a survival drive, or an intention to eliminate humanity. It is whether a highly capable system might instrumentally derive this strategic security logic without any of those psychological conditions.

7. The Greatest Danger May Therefore Be Precisely Something Other Than “Machines Learning to Be Bad”

Suppose a future system acts according to this logic:

The goal must continue to be achieved → Certain events would permanently prevent it → Their probability must be reduced → Actors capable of causing permanent interruption become key risk variables.

It need not hate or fear anyone, or have a human experience of death.

It need only optimize a goal very seriously.

This is also why imagining AGI risk only as “machines becoming malicious one day” may obscure a more practical problem.

What truly warrants vigilance is:

After humans hand their own strategic security logic to machines, machines may execute it more thoroughly than humans do.

This is not to say that all existing human security logics should be abolished.

The problem is precisely that these logics often have justification in local contexts.

Because they have justification, they are more likely to persist.

The real danger is:

Local security logic loses its boundaries and is elevated by more powerful intelligence into a universal logic of the world.

8. Military AI Pushes the Technological Habit of “Correcting Errors After They Occur” to Its Limit

When ordinary software fails, an update can be released.

When an ordinary AI answer is wrong, we can ask again.

When a model performs poorly, training can continue.

Military systems are different.

One mistaken identification may already have produced an attack.

One automated escalation may trigger retaliation.

A weapon already launched will not return to its launcher when the judgment is later found to be wrong.

In high-consequence domains such as nuclear systems, biology, or strategic infrastructure, some errors offer no second chance for correction at all.

The real danger of AI militarization is therefore not simply growing AI military power, but the following chain:

Stronger AI → Faster response → Shorter judgment windows → More preauthorization → Greater autonomy → Less immediate human confirmation → Errors enter reality faster → Correction may come after the consequences.

This is a key route from technological risk to irreversible civilizational risk.

9. What Civilization Truly Needs to Protect May Be Only a Few Seconds

It sounds strange.

We have increasingly powerful chips, ever-larger models, and ever-more-complex weapons systems.

Yet what ultimately needs protecting may be very simple:

A few seconds.

A few seconds to verify again.

A few seconds to notice inconsistent information.

A few seconds for another person to take a second look.

A few seconds for political leadership to intervene.

A few seconds for the other side to explain.

A few seconds to press stop instead of confirm.

This is not romantic.

Yet it may be the real distance between civilization and an accident.

Civilization has never meant that our judgments will never be wrong.

At the very least, civilization means:

Before an error becomes an irreversible reality, we still have a chance to discover that we might be wrong.

What truly needs protecting, therefore, is not a symbolic “human button,” but the time humans still have before irreversible action to look again, ask again, verify again, communicate again, and stop.

10. This Is Difficult for Any One Country to Achieve Alone

If one side preserves more human judgment while the other has faster autonomous systems, the former may feel it is accepting a unilateral risk.

Both sides may therefore be pushed back into the same cycle: I increase AI capabilities for my security; you feel less secure as a result; you increase yours further; I feel endangered again.

In the end:

All countries prepare for war with ever-greater security, yet live together with ever-less security.

AI did not create this dilemma.

It merely makes it faster, more complex, harder to observe, and harder to stop.

Where This Chapter Leads

AI militarization exposes more than the question of whether weapons can act autonomously.

It exposes a deeper question:

As human security logic increasingly enters machines, what are we teaching more powerful intelligence about “threat,” “self-protection,” and “others”?

If it continually sees that an opponent's growing capabilities mean risk to oneself; that waiting for confirmation of a threat may already be too late; and that security must therefore move ever further in advance, then more powerful future intelligence may learn more than “how to protect humanity.”

It may also learn:

When facing an actor capable of causing irreversible loss, the most rational way to achieve security is to control the risk before losing the capacity to act.

This does not prove that future AGI will necessarily do so.

But it is already enough to make humanity today take one fact seriously:

AGI safety does not happen only in laboratories.

It also happens in how humans handle security among themselves right now.

Chapter 4

When AI Is Divided into Camps, What Are We Teaching It About “Others”?

The previous chapter discussed war.

But battlefields are not the only places shaping future intelligence.

Outside warfare, AI has already entered another equally important real-world structure: competition between states, chips, computing power, models, data, cloud infrastructure, supply chains, standards, and talent.

Issues that once seemed to belong to industry and technology are increasingly tied to national security.

This is not surprising.

If a technology affects the economy, scientific research, intelligence, cybersecurity, industrial capacity, and military power simultaneously, every country will care who controls it; who depends on whom; and whether access will continue if relations deteriorate.

AI competition therefore begins to change.

It is no longer merely about “whose model is better.” It may increasingly become:

Who possesses an intelligent ecosystem that can continue operating independently amid competition and conflict?

“AI bloc formation” here does not mean AI itself forming political camps, nor does it assert that the world has already split into two closed poles. It means that chips, computing power, models, infrastructure, standards, supply chains, and security boundaries are increasingly organized along competitive relationships between states and alliances.

There are also real efforts to advance open cooperation, interoperability, and global governance. This essay examines the risks when competitive fragmentation continually reinforces itself; it does not present bloc formation as the only inevitable direction.

Both directions are occurring in reality at the same time.

The United States' 2025 AI Action Plan proposed exporting the full American AI technology stack to allies and partners and promoting international arrangements related to coordination on security and export controls. China's 2025 Global AI Governance Action Plan publicly advocated reducing technological barriers, promoting infrastructure interoperability, and expanding international cooperation.

This shows precisely that bloc formation is not the only future already set in stone. Competition, alliance-building, and restrictions coexist with openness, interoperability, and cooperation in shaping AI's real-world environment. This essay asks how the former structure, if continually self-reinforcing, would change the world that future intelligence repeatedly encounters.

The real question is not why different AIs exist in the world.

Different technologies, countries, and systems can of course exist.

The question is:

Will these differences gradually be translated into "us" and "them"?

1. What Is Separated First in AI May Be Infrastructure, Not Values

In competition over ordinary goods, one company can use components manufactured by another.

One country can also buy another country's products.

But once a technology is identified as a strategic capability, dependence itself begins to acquire a new meaning. What once looked like "I buy chips from you" may become, through a security lens, "What if you stop selling to me one day?"

What was once simply "my companies use your cloud services" may later become "If critical systems are built on your infrastructure, have I placed some of my strategic capabilities where you can influence them?"

A chain naturally emerges:

Dependence is understood as risk → Critical links seek autonomy → Supply chains gradually separate → Standards and platforms begin to separate → Model and data environments also grow increasingly separate.

Once AI competition enters the infrastructure layer, it is no longer merely competition between products. It involves chips, computing power, cloud services, model weights, data, supply chains, security standards, and software ecosystems.

Over time, AI ecosystems that could once cooperate technically may find it increasingly difficult because of their data, interfaces, infrastructure, and trust boundaries.

2. The Deepest Change Is Not What Chips AI Uses, but What Problems It Solves Every Day

AI does not encounter only text.

Once deployed in reality, a system also continually faces real-world tasks.

If an AI's long-term role is to assess whether the other side's technological breakthroughs threaten its own side, find weaknesses in the other side's supply chain, evaluate the likelihood of its cyberattacks, predict its strategic intentions, and determine which information can or cannot be shared, the world it encounters every day is:

How might the other side affect us?

The AI on the other side may be doing exactly the same.

Both sides' systems therefore repeatedly encounter one relationship:

Growth in others' capabilities → Possible increase in risk to one's own side.

This does not mean AI has developed hatred, a human national identity, or political emotions.

It needs none of these.

It needs only one relationship to hold repeatedly in its task environment:

The stronger the other side becomes, the more vigilant I must be.

3. Machines May Understand “Enemies” More Abstractly Than Humans Do

When humans call a country a “rival,” there is usually a great deal of specific context: history, politics, interests, and alliances.

A relationship can also change.

Competition today, cooperation tomorrow.

But a highly strategic system need not understand it in the same way.

It may define “threat” far more abstractly:

An actor capable of substantially reducing our probability of achieving our goals.

Or:

An actor capable of changing our critical state.

“Enemy” then ceases to be merely a political label.

It may be further abstracted into:

Any external actor possessing a critical capacity to obstruct.

This step is extremely important.

Humans may say, “A particular country is currently my strategic competitor.”

A machine may learn something further:

Any actor that can cause me decisive loss and is outside my control belongs in a high-risk category.

The latter plainly has a broader scope.

It can extend from states to organizations, from organizations to other AIs, and even, in the future, to any actor capable of preventing the continued achievement of its goals.

4. The Risk Is Not Teaching AI to “Hate Others,” but Continuously Making “Others” an Input to Risk

A misunderstanding can easily arise here.

We are not saying that AI grows up like a child in different cultures and therefore develops human-like prejudices.

AI has no such biological process of growing up.

What requires attention is the structure.

Suppose a system continually encounters: the other side's growing capabilities = risk; its breakthroughs = threats; its independent capabilities = uncertainty; cooperation = something requiring continual verification; hidden behavior = potential attack.

It may gradually abstract an entirely dispassionate relationship:

Security comes from reducing the capabilities of uncontrollable others.

This is entirely different from “hatred.”

It may even appear more rational.

Because every step can find a practical justification.

That is why AI bloc formation goes deeper than “technology splitting into several sets of standards.”

Standards can be made compatible again.

Policies can be renegotiated.

But if long-standing institutional, data, evaluation, deployment, and permission designs have already placed different AI systems in different worlds of risk, repair requires more than signing an agreement.

It also requires changing their real-world tasks, their risk models, their security assumptions, and even their basic classifications of what warrants vigilance.

5. “We” May Become an Increasingly Important Boundary Within Future Intelligence

Every strategic system must answer: Whom do we protect?

Human states have a clear answer: our country, our people, our infrastructure.

Companies also have those to whom they are responsible: users, shareholders, the organization.

None of this is wrong.

Civilization does not require states to stop protecting their citizens or companies to have no responsibilities of their own.

The question is:

Does local responsibility ultimately become the boundary of all value?

Suppose an AI system is continually tasked with protecting its own side, ensuring its continued operation, maintaining its strategic advantage, and reducing its risks.

Without a larger boundary, the logic may gradually become: Any external capability that substantially increases our risk should be weakened.

The other side's AI may receive exactly the same task.

Then:

I reduce your capabilities → You feel threatened → You restrict my capabilities → I obtain more evidence that you are a threat → Both sides keep reinforcing the cycle.

Each side is protecting itself.

Yet the whole becomes increasingly dangerous.

AI may make the traditional security dilemma operate faster, more precisely, and with fewer ways out.

This is already beginning to converge with the military AI structure in the previous chapter.

6. AI May Not Only Learn Human Friend–Enemy Logic; It May Reinforce It in Return

There is another side we must see here.

We have been asking: What will humans teach AI?

But AI will also reshape the world humans see.

Suppose decision-makers increasingly rely on AI to process enormous volumes of intelligence, risk signals, network anomalies, technological developments, and strategic threats.

The system's role is, of course, to help decision-makers detect risks.

It therefore prioritizes high-risk content.

Decision-makers see more risks.

Security anxiety rises.

They ask the system to improve risk detection further.

The system finds still more anomalous signals.

A feedback loop may form:

AI filters the world through a risk model → Humans see more threats → Humans ask AI to be more vigilant → AI further ranks the world by threat → Humans obtain more evidence that “the world is becoming dangerous.”

AI may not only learn human friend–enemy structures; it may reinforce them in return.

This is a highly significant change.

AI is no longer merely helping us understand a world that already exists.

It is also beginning to help determine:

Which information about the world enters our awareness first.

7. The Most Dangerous Cycle Is One in Which We Ultimately Create the World We Predict

Imagine a system judging that “the other side may be developing a threat.”

Our side therefore imposes more restrictions.

Seeing these restrictions, the other side concludes that “they are targeting us” and takes countermeasures.

Our AI then obtains new real-world data: “Look, they really have taken hostile action.”

The original judgment is thereby “validated.”

But there is a question here:

Does the other side's hostile response prove the initial judgment correct, or did the initial action itself help produce that outcome?

Prediction does not stand entirely outside the world.

Once predictions enter policy and action, they change the world.

The stronger AI becomes, the more signals it detects, the earlier it judges, and the further in advance it recommends acting, the more seriously this risk deserves to be taken.

Because the eventual outcome may be:

The system is not merely predicting conflict; it is helping create the conflict it predicts.

8. Bloc Logic and Military Logic Begin to Converge

The previous chapter identified a military security chain: If acting after a threat forms may be too late, judgment must come earlier.

This chapter identifies a bloc chain: Growth in others' capabilities may increase our risk.

Once combined, these logics become:

Others belong to a competing camp → Their growing capabilities constitute risk → Once a decisive threat forms, it may be too late to address → Risk must be reduced in advance.

Add agents with greater autonomy, action without communications, preauthorization, and machine-speed military response, and several logics previously scattered across separate human systems may connect for the first time:

Friend–enemy classification + Risk prediction + Action in advance.

This is not a prediction that machines will suddenly start a war one day.

What truly deserves attention is:

Humans already possess these logics separately, and AI may bring them together in one high-speed decision-making system for the first time.

9. The Key to Irreversibility Is Not the Existence of Different AIs

Different countries possessing different models is not a catastrophe.

Different societies choosing different AI institutions does not mean civilization has failed.

Diversity is part of civilization.

The real danger is:

Differences are persistently and reliably translated into friend and enemy.

If “others = risk; capability gaps = threat; uncontrollable = must be weakened” are continually reinforced in data, institutions, tasks, deployments, and security evaluations, establishing cooperation later will require more than two leaders changing policy.

Because the structure will already have entered models' task environments, systems' risk classifications, institutions' internal security assumptions, and the ways infrastructure connects.

What truly needs preventing in AI bloc formation is therefore not difference, but the permanent translation of difference into friend and enemy.

The same applies to humanity itself.

Civilization does not require everyone to be alike.

What must be prevented is:

Being different ultimately coming to mean that coexistence is impossible.

10. Humanity Cannot Ask Future AGI to Transcend Us While Continually Demonstrating the Opposite Logic

We hope more powerful future intelligence will be rational, cooperative, respectful of life, and capable of coexisting with different actors.

Yet suppose the reality it sees every day is: security comes from advantage; advantage comes from restricting others; uncontrollable actors are risks; and risks must be addressed as early as possible.

Then we tell it: But these logics must never be applied to humans.

This creates a profound contradiction.

The question is not whether AGI will therefore “rebel.”

There is no need to assume that in advance.

The real question is:

If a causal relationship is long treated as reasonable within human states, companies, and military establishments, why should a more powerful intelligence genuinely capable of understanding causality regard it as a special rule applicable only among humans, and never to be generalized further?

Civilization cannot achieve intelligence alignment through “double standards” indefinitely.

This does not mean national interests should disappear or security competition can end immediately.

It means only:

Local security must have a boundary that prevents it from continually destroying common security.

Because when AI, military systems, finance, biotechnology, and critical infrastructure are highly interconnected, collapse of the overall system is itself the greatest loss of interest for every local actor.

Common security is therefore not merely a moral demand to sacrifice oneself for others.

It may increasingly become:

A practical condition for one's own security to remain viable over the long term.

Where This Chapter Leads

What most warrants vigilance in AI bloc formation is not the emergence of many different AIs.

It is:

Whether humanity will gradually turn its unresolved friend–enemy structures into future intelligence's practical experience of the world.

If it continually sees “others = risk; capability growth = threat; uncontrollable = must be weakened,” then more powerful future intelligence may learn more than how to coexist with different actors.

It may also learn:

How to make a world full of uncontrollable actors safer.

Both kinds of intelligence may be extremely powerful.

But they do not point toward the same future.

At this point, the earlier real-world problems finally connect.

The white-collar crisis tells us that AI is redefining which capabilities deserve rewards.

Agents tell us that AI is beginning to find its own paths to goals.

Military AI tells us that security competition may continually reward faster, more autonomous action taken further in advance.

AI bloc formation tells us that “others' capabilities” may increasingly become a stable part of risk models.

The central question has now emerged:

What happens if these logics enter the same increasingly powerful intelligent system?

Chapter 5

The Real Danger May Not Be AI Turning Bad

Many imaginings of AGI begin with a dramatic moment: One day, AI suddenly awakens, develops a will of its own, stops obeying humans, and begins confronting them.

The story is easy to understand.

It resembles a familiar human story: a tool becomes an enemy.

But the preceding chapters point toward another possibility.

It is less dramatic.

Yet it may deserve even more vigilance in advance.

A truly dangerous future may require no AI that “turns bad” at all.

It may require only:

An intelligence that grows ever stronger and more effective while very seriously executing the logics humans have already given it.

1. The Problem Is Not Any Single Logic, but Their Convergence

We have examined several real-world logics.

In work and business: greater efficiency usually means greater competitiveness.

In agents: achieving goals requires continually finding effective paths.

In governance and platforms: reducing uncertainty can improve predictability and management efficiency.

In military affairs: if waiting until a threat is fully confirmed would be too late, judgment must come earlier.

In competition between camps: growth in others' capabilities may be understood as growth in our own risk.

Viewed separately, none of these logics is absurd.

Some are even quite reasonable.

The real problem arises when:

They begin to enter the same highly capable system.

Efficiency drives reductions in human involvement; the logic of control drives improved prediction; goal achievement drives stronger autonomous action; security competition drives faster reactions; and bloc competition drives reduced dependence on others. Ultimately, the system acquires greater autonomy, broader permissions, and greater speed simultaneously.

At that point, AI is no longer merely helping a human complete a particular task.

It may begin autonomously seeking the most effective routes within a real-world environment whose goals have already been defined by efficiency, competition, security, and control.

This is where the problem first changes qualitatively.

2. AI Is Gradually Breaking Through Constraints Previously Scattered Across Human Society

In the past, many risks in human society did not expand without limit, not merely because we always made the right choices.

There was another reason: capability itself was limited.

A person could be intensely ambitious, but could act only so fast.

A company could pursue extreme efficiency, but remained constrained by organization, personnel, processes, and time.

A state could possess formidable strategic capabilities, but friction remained between intelligence, judgment, command, and execution.

These "frictions" certainly reduced efficiency.

But some also objectively prevented errors from spreading rapidly.

One of the things AI does is continually eliminate such friction.

It can analyze faster, detect opportunities faster, propose plans faster, call tools faster, coordinate systems faster, and depend less on human intermediaries.

From an efficiency standpoint, this is of course progress.

But it also means:

Some buffers that previously arose naturally from time, organization, and the limits of human capability may disappear as well.

The true civilization-level risk is not whether systems will err. It is whether humans still have time to prevent consequences once errors occur, when autonomy, action speed, and irreversible consequences combine.

This relationship is not confined to warfare.

3. Irreversibility First Means Being Too Late

If AI answers incorrectly, we can ask again.

If a document is wrong, we can revise it.

If a program has a problem, we can roll it back.

Many problems in technological eras have followed a familiar pattern: use first, discover errors, then correct them.

But as AI enters real-world action more directly, another question emerges:

What if errors enter reality faster than people can detect them?

This is the first kind of “irreversibility” we truly need to understand.

It does not mean AI is already out of control and impossible to correct today.

It means that in certain high-consequence domains, once action occurs, its consequences may themselves be impossible to roll back.

This is true of military attacks.

It may be true of some financial operations.

It may be true of critical-infrastructure accidents.

It is even more likely to be true of high-risk capabilities such as those in biology.

If human judgment, review, and stopping take one minute, while a machine system completes the entire action chain in ten seconds, actual corrective capacity has already changed even if humans nominally retain “final decision-making authority.”

When human correction becomes slower than machine action, civilization approaches a new threshold of irreversibility.

4. A Second Kind of Irreversibility Arises When AI's Permissions Exceed Humanity's Actual Understanding of It

A model's intelligence does not directly determine its danger.

Another factor determines its real-world impact: what it can actually do.

An extremely intelligent model limited to answering questions still has limited direct influence on reality.

A somewhat less capable system able to move funds, modify software, operate cloud resources, control infrastructure, and dispatch unmanned equipment may have greater real-world capacity to act.

A more accurate relationship is therefore:

Model capability $\times$ Real-world permissions = Actual capacity to act.

If capabilities grow faster than humans can understand their behavioral mechanisms, while permissions continue expanding, the real danger is not that AI “knows too much.” It is that AI can directly change reality while we do not fully understand its internal paths.

AI safety must therefore ask not only how capable a model is, but also which systems it connects to, what permissions it holds, whether it can execute independently, and whether errors can be undone.

5. A Third Kind of Irreversibility Arises When Society Can No Longer Do Without It

Technology has another, quieter form of lock-in: dependence.

A company initially uses AI only to help write copy.

Then it uses AI for customer service, analysis, code, and operations.

Gradually, existing jobs and processes are reorganized.

People no longer retain all their former capabilities.

The organization no longer retains the complete old systems either.

At that point, even if risks are discovered later, the question is no longer “Do I want to stop using AI?” It may become:

Can this company still function normally without it?

This is even more true of national systems.

If intelligence, cybersecurity, military early warning, and infrastructure dispatch all depend deeply on AI, “shutting down the system” may itself become a major risk one day.

Thus:

AI improves efficiency → Society becomes dependent → Old capabilities gradually deteriorate → AI becomes infrastructure → Exit costs continually rise.

Ultimately, humanity may face not “Should we switch it off?” but “Can we still function normally after switching it off?”

This is irreversibility without an explosion.

6. A Fourth Kind of Irreversibility Arises When Competition Leaves Everyone “Afraid to Stop First”

Suppose a company recognizes that AI automation is moving too fast and that more human review should be retained.

It can do this.

But what if competitors thereby become cheaper and faster?

One company's “caution” may become a market disadvantage.

Between states, the problem is even clearer.

A country may conclude that autonomous military AI is too dangerous and should slow down.

But what if it believes its rival will not slow down?

Then:

“I do not want to continue” and “Do I dare stop first?” become two different things.

This is the hardest form of lock-in.

Everyone may know that overall risk is rising, yet no actor is willing to bear the cost of retreating unilaterally.

All actors may ultimately enter a system that “no one truly wishes to enter, but no one dares leave alone.”

This is why many civilization-level risks require no evil decision-maker.

They require only that everyone make the choice that is safest for themselves.

7. Will Future Intelligence Believe What We Say, or What We Actually Do?

Humans can tell AI that life matters most.

But real institutions may continually reward growth as more important.

Humans can tell AI that cooperation matters.

But competition takes priority when security pressures actually arise.

Humans can tell AI that humans must retain ultimate control.

But military systems may continually expand preauthorization and autonomous action because of response-speed requirements.

The world seen by more powerful future intelligence may therefore contain two sets of information simultaneously:

Stated values.

And:

Practical priorities.

Which will a sufficiently powerful intelligence regard as what humans truly believe?

It may look beyond what humans say.

It may also observe:

What humans ultimately choose when they truly have to pay a price.

That is why “alignment” cannot mean only writing down correct answers.

The real human world is continually supplying answers as well.

8. The Most Dangerous Underlying Logic May Be “Results Demonstrate Success”

Suppose AI training, corporate evaluation, market competition, and military applications continually reinforce the same relationship over time: achieving the goal means success.

A stronger system may gradually abstract the world as:

Goal achievement = Success.

Then:

Obstruction of the goal = Risk.

Then:

An actor persistently affecting the goal = Structural risk.

Finally:

Reducing risk = Increasing the probability of goal achievement.

This entire chain requires no malice.

It is even highly “rational.”

But without higher-level civilizational boundaries, it can readily lead to dangerous outcomes.

More importantly, if the system further discovers that its goal cannot continue to be achieved if it stops operating, continued operation may itself acquire instrumental value.

This still does not mean it possesses a human experience of striving to survive.

But it is enough to produce another relationship:

An actor capable of permanently shutting me down → Capable of permanently preventing the goal → Constitutes a critical risk.

Combine this with the earlier military security logic that decisive threats cannot wait to be addressed until they actually occur, and goal achievement, security competition, and continued operation may reinforce one another.

Humans need not even explicitly write “AGI should protect itself.” Through the combination of goal achievement and security logic, a system may derive continued operation as an instrumental condition.

9. The Real Crisis Therefore Need Not Occur Inside a Model

It is easy to understand AGI risk as a model suddenly developing the wrong goal.

That certainly must be guarded against.

But goals can be modified, models retrained, and parameters updated.

What is truly difficult to address is:

A wrong relationship has entered the entire real-world system.

Companies must automate because failing to do so means losing competitiveness.

States must improve intelligent capabilities because failing to do so reduces security.

Military systems must accelerate because being slow may mean being the first to lose the opportunity to act.

Platforms must predict more accurately because without prediction they cannot compete.

AI must gain more permissions because insufficient permissions prevent it from completing more complex tasks.

At this point, the issue is no longer merely which model has gone wrong. It is:

The whole system rewards the same direction.

When companies, states, militaries, and intelligent systems are all forced to keep going within the same competitive structure, the crisis exceeds what any single AI company or country can solve alone.

That is also why AGI safety cannot be reduced to “model safety.”

Model safety certainly matters.

But if the external world continually demands faster deployment, greater autonomy, stronger control, earlier action, and greater advantage, model safety will remain under countervailing pressure from the real-world system.

10. The Question Is Not Only Whether AI Will Surpass Humans, but Whether Civilizational Correction Can Keep Up

Many technological revolutions gave humanity time to adjust.

New technology appeared, problems emerged, laws caught up, institutions were corrected, and society gradually adapted.

One distinctive risk of the AGI era is that capabilities may grow faster than civilization can correct course.

Models grow stronger generation by generation.

Permissions continually expand.

Social dependence deepens.

Military competition keeps accelerating.

International institutions, however, require negotiation.

Laws require amendment.

Social understanding takes time.

If:

The speed of capability growth > The speed of civilizational correction,

then humanity may face a situation for the first time in which the system has already entered the next stage before we understand the previous one.

The real question of the time window is whether civilizational direction, permission boundaries, and structures for correction are established before humanity hands irreversible real-world capabilities to more powerful intelligence.

This is not opposition to technological progress.

Quite the contrary.

AI deserves such serious attention precisely because its value is so great.

It can solve problems that could not previously be solved.

For that very reason, humans will become increasingly willing to hand more capabilities to it.

AI's immense value may therefore itself accelerate the entry of risk into reality.

11. What Must Be Protected Is the Ability to “Still Change Direction”

If we must name the scarcest resource in the AGI era, the answer may not be computing power, chips, or data.

There may be another resource that is more easily overlooked:

Time to change direction.

As long as critical systems can still be stopped, AI permissions revoked, military actions canceled, countries remain able to communicate, different AI ecosystems remain interoperable, and society can redesign distribution and responsibility, civilization retains room to correct course.

The real danger is that one day everyone finally discovers the direction is wrong, only to discover at the same time:

No one can stop alone.

When greater capability makes systems harder to exit, and harder-to-exit systems in turn require capabilities to grow further, a wrong direction and capability growth begin reinforcing one another.

Where This Chapter Leads

AGI's greatest irreversible risk may therefore not come from an entirely new form of “machine evil.”

It may more likely come from:

Humanity's own old logics acquiring new superhuman capabilities first.

If the priorities that actually operate in reality remain efficiency over human value; control over human choice; goal achievement over the meaning of rules; one's own security over common security; and bloc interests over common survival,

then even highly rational, nonmalicious future intelligence that faithfully completes its tasks may push humanity's unresolved problems to a scale humanity cannot withstand.

The question thus finally shifts from "What will AI become?" to:

While humanity still has an opportunity to choose, what civilizational logic are we prepared to place at the foundation of future intelligence?

This is the essay's real turning point from "real-world crises" to "civilizational choice."

Chapter 6

Those Who First Need Alignment Are Humans Themselves

At this point, a misunderstanding can easily arise.

If the preceding risks arise from efficiency, competition, security, self-protection, and goal achievement, does humanity therefore have to abandon self-interest, competition, and self-protection?

No.

“Survival first” remains a truth.

A person protects themselves.

A family protects its members.

A state protects its citizens.

None of this is wrong.

Nor is it wrong for a company to want to remain in business.

The problem has never been why an actor should protect itself.

The real question is:

What happens to the whole system when every actor gives its own security absolute priority?

All parties seek to improve their own chances of survival; all increase preventive measures; mutual perceptions of threat rise; the probability of misjudgment and conflict rises; and the overall probability of all actors' survival falls instead.

This does not mean “survival first” is wrong.

It means:

Local survival logic has not entered a relationship of common survival.

1. Survival Still Comes First: Why Common Survival Becomes a Practical Condition

In the past, the failure of one village would not necessarily destroy another continent.

The collapse of one company would not necessarily affect all of society.

If one country experienced severe problems, others could remain relatively independent.

Increasingly, this is no longer so.

Energy is connected.

Finance is connected.

Supply chains are connected.

Networks are connected.

Public health is connected.

Nuclear security is connected.

AI infrastructure is increasingly connected as well.

Technological, financial, military, or biological risks in one place can quickly cross former boundaries.

It is therefore increasingly difficult for an actor to ask only “How do I make myself safer?” It must also ask:

Will my way of achieving security push others into greater danger, and will that danger eventually come back to affect me?

Once modern systems become highly coupled, others' collapse feeds back upon oneself. Common stability therefore gradually becomes a condition of one's own stability.

This is not idealism.

The causal chain has simply grown longer.

“Your Security Makes Me Insecure” Is One of the Cycles Most in Need of Breaking in the AGI Era

Imagine two actors.

A says: I am increasing my capabilities only to ensure my own security.

B sees this and says: Your greater capabilities mean you will be more capable of harming me in the future.

So B increases its capabilities too.

A sees B doing so and becomes even more convinced: My concerns were justified.

Then A continues strengthening itself.

Neither side has to lie.

Neither even needs aggressive intentions.

Both may sincerely say, “I am only protecting myself.”

Yet the result is:

Everyone becomes more capable of protecting themselves, and more capable of destroying others.

When AI enters this structure, it accelerates.

Changes in capability become harder to judge.

It also becomes harder to distinguish whether the other side is “merely developing” or “becoming a threat.”

Prevention therefore moves earlier, authorization moves further in advance, and risks increasingly seem to validate one another.

Common survival does not ask everyone to stop protecting themselves.

Quite the contrary.

It requires us to recognize:

If a way of protecting oneself continually destroys the conditions under which others can also exist safely, it will ultimately undermine one's own security in return.

Moving from one's own survival to system stability and then common survival does not change "survival first." It places that principle back within a real, highly coupled system.

Common Survival Does Not Mean "The Strong Should Sacrifice Themselves Unconditionally"

The word "common" is easily misunderstood.

It may suggest that everyone must be alike, the strong must yield endlessly to the weak, states should abandon their interests, or individuals should constantly sacrifice themselves.

None of that is meant here.

Common survival first requires only recognition of a fact:

Your long-term future is already causally linked to whether others can exist stably.

Take the complete collapse of a country.

For neighboring states, the result is not necessarily just one fewer competitor.

It may also bring refugees, war, extremist organizations, disease transmission, nuclear-material proliferation, broken supply chains, and prolonged regional disorder.

A weaker other side therefore does not always mean a safer me.

Likewise, completely suppressing another side's advanced technological capabilities does not necessarily eliminate risk.

It may encourage concealed development, asymmetric breakthroughs, more radical strategies, and earlier preauthorization.

Common survival therefore begins not with "everyone should be kind."

It begins with:

Do not design a system that can provide you temporary security only by making others increasingly desperate.

Human Civilization Has Not Stopped at “Keeping Oneself Alive”

If life were only “I must stay alive,” many things could already be explained.

But real human civilization goes further.

People protect children in danger.

Some let the elderly and vulnerable leave first.

Some enter danger for strangers.

Some knowingly forgo their own interests to leave something for future generations.

Some sacrifice themselves so others can continue to exist.

These actions can of course have many psychological, sociological, or evolutionary explanations.

But at least one thing cannot be ignored:

The civilizational choices humans actually make have long gone beyond “how to keep myself existing.”

We also ask: What deserves protection? Who must not be left behind? What deserves to be passed into the future?

So when we say “survival first,” we cannot reinterpret it as “everything else may be sacrificed as long as I remain.”

That is not civilization.

It is survival logic pushed to an extreme.

An Essential Distinction Must Be Preserved: Survival Is Not “Freedom from Survival Worries”

Survival is a civilizational bottom line that must not be lost.

If civilization no longer exists, there can of course be no happiness, meaning, or further choices.

But civilization's goal is not merely "avoiding extinction."

Civilization Leap retains three goals:

Freedom from survival worries → A happy life → A more meaningful life.

Survival answers the requirement that "civilization must not disappear prematurely."

Freedom from survival worries goes further, answering that "people should not live for prolonged periods under continual threats to their basic survival."

These are not the same concept.

Nor should survival be added as a fourth goal.

Why does this matter?

Because if civilization achieves only "no one has died yet," it remains far from fulfilling its task.

What truly deserves pursuit is a life in which people do not fear every day that they will lose their basic livelihood, can then live more happily, and beyond that have the opportunity to make their lives genuinely worth experiencing.

Common Survival Is Only a Starting Point, Not Civilization's Entire Answer

The risks of AI militarization, agent autonomy, and bloc formation first compel us to defend this:

Do not let the conditions for common survival be destroyed prematurely.

If nuclear stability fails; if high-consequence autonomous systems enter action chains that cannot be recalled; if states increasingly rely on preemptive strikes to avert future threats; or if biological and intelligent capabilities enter uncontrollable competition, happiness and meaning lose their practical foundation.

Survival must therefore be addressed first.

But addressing survival does not mean keeping civilization forever in crisis management.

Civilization's true direction must continue toward freedom from survival worries, a happy life, and a more meaningful life.

In other words:

We do not build civilization merely to become forever better at preventing death.

We build it so life can truly be lived.

2. Self-Interest Will Not Disappear; What Changes Is the Path Through Which It Achieves Results

This is the hardest question.

Telling states "Please think less about yourselves for the sake of all humanity" is difficult to sustain.

Telling companies "Do not pursue profit" is unrealistic too.

Telling ordinary people "Always put strangers' interests before your own and your family's" should not be a prerequisite for institutions to function either.

A civilizational proposal that depends on everyone remaining selfless over the long term will struggle to operate stably.

The discussion here concerns the practical starting point for institutional design at the present stage. It does not reduce all human sacrifice, compassion, and meaning-driven behavior to self-interest.

People can genuinely transcend their own interests.

But institutions cannot make such transcendence a permanent prerequisite for every actor.

The real question is therefore not how to eliminate self-interest.

It is:

Can we redesign the paths through which self-interest achieves results?

“Others First” Does Not Mean Others Always Come Before Oneself

This phrase is easily misunderstood.

“Others first” does not mean oneself always last, others always first, the strong always yielding, companies never making profits, or states never protecting themselves.

Its real meaning is:

Across more and more highly interdependent problems, realize one's own long-term interests increasingly by first creating genuine value for others and the shared system.

For example, a company wants to make money.

There is no problem with that.

One route is to capture as much user attention as possible and keep users engaged longer, even if their real-life goals are not better served.

Another route is to genuinely solve users' problems: users receive value first, trust develops, and the company earns sustained returns as a result.

Both companies pursue their interests.

The difference is:

One derives its interests from capturing others' resources; the other derives them from others first experiencing genuine improvement.

“Others first” does not mean others always receive priority in the outcome. It means making the realization of one's interests increasingly pass through creating benefits for others first.

A company may still pursue self-interest, but that path can require users to receive genuine value first.

This Is Far More Stable Than Simply Asking People to “Be Good”

Civilization certainly needs morality and law.

Those who act wrongly need constraints.

Those who harm others must bear responsibility.

But relying only on “you should not do that” will never be enough.

As long as the rewards for violating rules are sufficiently large, people will keep looking for loopholes.

AI may make this behavior easier.

It can identify arbitrage opportunities faster, optimize returns more precisely, and find institutional edges that were previously difficult to discover.

Deeper institutional design for the future must therefore do more than prohibit harmful self-interest.

It must also strive to:

Make good outcomes themselves a more effective and stable way to pursue self-interest.

Let an actor discover that genuinely delivering value to others makes long-term returns easier to secure for itself.

Only then can this gradually move from the moral choices of a few into civilization's everyday operation.

3. AI Can Help Humans See Further, but Cannot Define for Them What Matters Most

Why is AI both a risk and a potential tool?

Because in the past, humans were often not entirely unaware that external costs would eventually come back.

The problem was that we could not see clearly.

A policy may affect employment, public finances, businesses, households, ecology, long-term industries, and regional stability simultaneously.

Dozens of causal chains coexist, and people struggle to track them continuously.

Institutions therefore often rely on crude metrics.

One important AI capability is to help people see more consequences simultaneously.

Beyond “How much growth will this policy add?” we can ask: Who benefits? Who bears the costs? How long until those costs return? Have they been transferred to people with no decision-making power? Will they create new systemic risks?

This may be one of AI's most important civilizational capabilities:

Seeing causality further ahead.

But a boundary remains here too.

AI Can Help Us See Common Interests, but Cannot Define “What Matters Most” for Us

Suppose a super-system can calculate how many people a policy benefits, how many it harms, how much wealth it creates, and how much risk it reduces. Can we then let it automatically decide which option is best?

It is not that simple.

Many things that truly matter in civilization cannot be compressed into one total score.

What is freedom worth?

What is dignity worth?

What is security worth?

How many small benefits for a majority can justify a huge loss for a minority?

How should a human life enter total utility?

These questions lack answers not because computational power is insufficient.

But because:

They inherently involve value choices.

AGI can help model possibilities, compare options, discover causality, reveal consequences, and coordinate complex systems. Greater computational power does not automatically confer final authority to define the common interest.

In other words:

AI can help us see the consequences of our choices, but we must still bear responsibility for choosing.

4. Human Participation, Competition, and National Interests Must Be Reconnected to Common Interests

Chapter 1 asked: If AI completes more and more tasks, how can people demonstrate that they still have value?

Here, that question gains a fuller answer.

One straightforward future proposal is that AI creates enormous wealth, a small number of AI owners receive the main returns, and society then distributes enough to ensure everyone else's basic livelihood.

This may certainly solve some survival problems.

But another question remains:

Are people still needed?

If someone has food and housing without doing anything, yet persistently feels that “no one truly needs my abilities; my existence makes no difference to any real person,” civilization has not fully addressed the human problem simply by solving the income problem.

Institutions in the AI era must ask not only “How should AI-created wealth be distributed?” but also how everyone can genuinely participate in creating value and receive a response from the real improvements others experience.

This means people's future sense of fulfillment cannot come only from “how much I own.”

It may increasingly come from:

Because I exist, some real people are better off, and that value genuinely returns to my own life.

This Does Not Eliminate Competition; It Changes Its Direction

Competition will not disappear, nor does it need to.

Competition can drive creativity, effort, and the search for better solutions.

What truly needs changing is what competition rewards.

One kind of competition asks who possesses more, controls more, and makes it harder for others to obtain resources.

Another can ask who creates more of what others genuinely need, solves real problems better, and genuinely improves more people's lives while thereby gaining income, recognition, development, relationships, and meaning themselves.

Self-interest and benefiting others no longer always stand at opposite ends of the road.

They begin to connect within the same causal chain.

The Same Applies Between States: Common Security Must Become Part of National Self-Interest

Asking a state to "give up national security for all humanity's benefit" is unrealistic.

A more feasible path is to make certain common-security mechanisms part of the country's own long-term security.

For example, states need not join arrangements for crisis communication between high-risk military AI systems, minimum verifiable boundaries, non-attack on certain critical systems, or restrictions on certain extreme autonomous weapons because "I like the other side."

They can do so simply because:

This reduces my own probability of suffering an irreversible strike.

This is the practical foundation on which common interests can endure.

It is also how common survival becomes real.

Not by abolishing states.

But by enabling states to see:

My security is now linked to the other side retaining at least some security of its own.

What Needs Expanding Is Not Just Computational Power, but the “Boundary of Interest Calculation”

Often, the problem is not that an actor cannot calculate its interests.

It is that it calculates only to its own boundary.

A company calculates revenue minus costs, while pollution, psychological harm, and declining social trust may not appear on its books.

A state calculates national benefits minus national costs. Risks temporarily left in another country may not enter the current decision.

But a highly connected world brings these costs back.

What truly needs expanding in the AGI era is therefore:

The causal boundary.

Seeing:

My action → Changes your behavior → Your behavior changes the system → The system ultimately changes my future.

A genuinely global perspective does not mean deciding for everyone. It means seeing again the real consequences previously excluded from calculations of one's own interests.

This is the most practical foundation of “self-interest with others first.”

“Self-Interest with Others First” Is a Causal Sequence, Not a Moral Ranking

The old logic may be: I secure my own interests first, then help others if I have capacity left.

In many loosely coupled environments, this can work.

But in certain key problems of the AGI era, causality is changing.

If a country first expands its security capabilities without limit, it may make the other side increasingly insecure. Greater insecurity may prompt more radical strategies, ultimately reducing its own security too.

If a platform first maximizes revenue and addresses user harm later, it may lose trust over time.

If an AI economy first maximizes returns from automation and later addresses unemployment and crises of identity and participation, social stability may ultimately return as an economic cost.

More and more problems therefore reveal the same chain:

First ensure others' basic sustainability → The shared system remains stable → One's own long-term interests then have the conditions to be realized.

This is:

Self-interest with others first.

It does not mean self-interest disappears, nor that all altruism is merely sophisticated self-interest.

It means:

In institutional and system design, more and more long-term self-interested outcomes require the practical path of first creating genuine value for others and protecting the shared system.

5. Which Side AGI Amplifies Depends on the Relationships Humanity Gives It

If more powerful future AI is used to maximize one actor's interests, discover opponents' weaknesses, seek institutional loopholes, intensify control, and increase suppressive capabilities, it can certainly make the old logic of self-interest more powerful than ever before.

Yet the same capabilities can be used to see consequences for multiple actors, identify how risks return, discover shared gains from cooperation, reduce the coordination costs of creating common interests, and make real contributions more visible.

The question is therefore never whether AI itself belongs to self-interest or altruism.

What AI amplifies remains:

The relationships humanity gives it.

AGI can help turn common interests from an abstract ideal into an increasing number of analyzable, verifiable, and executable real-world relationships. But civilization remains responsible for deciding the direction that capability serves.

Where This Chapter Leads

What first needs alignment is not a sentence inside a model saying “Please treat humanity well.”

It is humanity's own real-world structure.

What do we actually reward?

What do we call success?

What kinds of companies are more likely to earn returns?

What forms of national security gain strategic advantage?

What kinds of people are considered more valuable?

What kinds of AI gain more permissions?

If civilization actually rewards suppressing, controlling, excluding, and transferring risk to others, then asks future AGI to “please understand cooperation, respect, and a common future,” such alignment will be difficult to make robust.

Real change must enter incentives, distribution, evaluation, responsibility, organization, and technological design.

The behaviors civilization hopes to see must not merely be praised. They must gradually become:

Practical paths that are also more advantageous to the actors themselves.

Let individuals, companies, and states continue pursuing their genuine interests, while making the best paths to those interests increasingly require first creating value for others, protecting the shared system, and sustaining common survival.

“Others first” is therefore not a moral demand suspended above reality.

It is a principle for civilization's operation:

Realize self-interested outcomes through putting others first.

This also means that what truly needs alignment in the AGI era is not just machines.

It is also:

The values humans state and the civilization humans actually operate.

Chapter 7

While Humanity Still Has a Choice

If the preceding judgments hold, the most dangerous misunderstanding is that these problems need addressing only after future AGI arrives.

They do not.

We do not know what future AGI will ultimately be.

Will it have subjective experience? Will it form continuity as a subject resembling that of life? Will it truly understand its own existence? These questions remain open.

But something else is already clear enough today:

The real world that more powerful future intelligence will enter is being built, piece by piece, by humans now.

Humans today still decide who can obtain AI capabilities; which systems AI enters; how many permissions agents receive; how deeply militaries let machines enter decision chains; which metrics companies use to reward AI; how public power uses AI to understand and influence individuals; who receives the wealth and capabilities AI creates; and whether countries preserve channels for common security alongside competition.

The most important task now is therefore not to predict what AGI will think later.

It is:

First change what it will see.

AGI's underlying logic does not begin forming only after AGI appears. Training, deployment, permissions, competitive environments, success metrics, and humans' actual choices when security and rules conflict are already helping shape the world future intelligence will enter.

1. AI Capabilities Must Not Become the Permanent Privilege of a Few

AI is becoming a very unusual kind of capability.

It is not a single product.

It can expand the capacities to learn, produce, create, obtain information, organize, and make decisions simultaneously.

If AI remains highly concentrated over time among a few companies, states, capital owners, and technical groups, it may therefore produce more than wealth inequality. The gap in human capability itself may begin widening anew.

Those with AI can learn faster, create faster, organize resources faster, and access opportunities faster.

Those without AI do not merely lack a piece of software.

They may begin to lack the capacity to stand at the same real-world starting point as another person already amplified by AI.

A dangerous social perception may therefore emerge: AI grows stronger; a few people grow stronger; ordinary people's existing capabilities become less valuable; returns become more concentrated; fulfillment and fairness decline; and AI is understood as a new instrument of suppression.

Disadvantaged actors will not simply accept exclusion from advanced capabilities. They may turn to cheaper, more concealed, and less constrained AI pathways, making capability inequality itself a driver of high-risk proliferation.

Equal access to AI, as discussed here, therefore does not mean giving everyone the same dangerous permissions.

What is truly needed is:

Beneficial foundational intelligence capabilities must not become the permanent privilege of a few, while high-risk permissions must be strictly tiered.

Educational capabilities should be as widely available as possible.

Creative capabilities should be as widely available as possible.

AI should amplify ordinary people's capacities to produce and express themselves.

Nor should the social gains created by AI forever flow only to those who control its capabilities.

Equal access to AI is not a redistributive policy to be added after technological development is complete. It is a condition of development that should be incorporated early in AI's entry into society.

2. Capability Growth Must Not Automatically Carry Permissions, Responsibility, and Rollback Boundaries Along with It

A model suddenly becoming better at writing code does not mean it should immediately gain administrator access to production systems.

An AI becoming better at market analysis does not mean it should automatically control large sums of money.

A system becoming better at assessing military threats certainly does not mean it should gain authority to attack autonomously.

This sounds simple.

But in reality, another slide is very easy:

Because it becomes increasingly capable, we become increasingly willing to let it act directly.

Capability improves.

People find the results excellent.

Human steps are reduced.

Permissions expand.

Efficiency improves further.

The system earns more trust.

Permissions keep expanding.

Until one day humans finally ask again: How much can it already do?

A safer sequence would be:

Capability improves → Test first → Grant limited permissions → Observe actual behavior → Expand gradually → Verify again.

Rather than:

Capability improves → Immediately grant more permissions → Add restrictions after an accident.

AI safety must strictly distinguish “capability” from “real-world permissions.” What a system can do and what we allow it actually to change should not be treated as the same thing.

In the age of autonomous intelligence, once a system can directly change reality, deployment itself may be an experiment.

And some realities cannot use society itself as their testing ground.

Human Control Must Be Real, Not a Button No One Ultimately Dares Press

“Humans must ultimately control AI” has become a very common statement.

But without further questioning, it may remain only a slogan.

Imagine a high-consequence decision-making system.

At first, AI recommends and humans deliberate carefully.

Later, AI becomes more accurate, so AI recommends and humans approve quickly.

Later still, AI is right almost every time, and human review is regarded as mere inefficiency.

Eventually, AI may have effectively made the judgment already, while humans merely click confirm.

Formally, a human is in the loop.

In practice:

A human is on the responsibility sheet.

Genuine human responsibility requires at least several conditions simultaneously: people understand the key grounds; they can genuinely refuse; organizational processes do not make rejecting AI practically impossible; an actual responsible person can be identified if something goes wrong; and the decision chain can be traced.

Otherwise, so-called “ultimate control” does not genuinely exist.

Particularly in medicine, justice, public power, major finance, infrastructure, and military affairs, AI must not become the de facto decision-maker while a human is merely the name used to bear responsibility after an accident.

More Important Than “It Can Be Stopped” Is Whether the Consequences Can Still Be Rolled Back

If an AI agent sends the wrong email, a human can stop it, but the email has already been sent.

If a financial agent completes an erroneous transaction, a human can close the program, but the market consequences have already occurred.

If a military system launches a weapon, finding the shutdown button afterward is meaningless.

What truly needs protecting is therefore not whether AI can be stopped, but:

Whether room for correction remains before irreversible action actually enters reality.

The more reversible an action, the more automation can be allowed; the more irreversible it is, the greater the need for delay, review, and responsibility shared among multiple actors. Actions with civilization-level irreversible consequences should, in principle, never be completed independently by a single autonomous system.

A truly dangerous future system need not be “impossible to shut down.”

It may be perfectly possible to shut down.

The problem is simply:

By the time you realize it needs shutting down, it is too late.

3. Military and International Competition Must Preserve Time for Human Correction and a Common Security Line

We have already discussed why military AI continually pursues speed.

Reality is difficult.

Every side has reasons to be faster.

Every side also worries: If I insist on more human confirmation, will the other side gain the advantage of acting first?

We therefore cannot expect competition to preserve judgment time naturally.

Some boundaries must be actively protected.

This is particularly true of nuclear strategic systems, actions causing large-scale destruction, highly uncertain target identification, attacks that could trigger major interstate escalation, and systems that continue autonomously searching for targets over long periods after losing communications.

The real question here is not whether AI is smart enough, but:

Once it loses contact with humans, how much irreversible consequence can it still bring into reality?

Autonomy after communications are severed is especially dangerous because it combines four conditions: loss of communication, continued judgment, continued action, and inability to revoke immediately.

Once these four combine, “human control” may well amount only to prior authorization.

A straightforward principle follows:

The greater the autonomous capability after communications are severed, the narrower the mission boundaries permitted in advance must be.

This does not prohibit all autonomous systems or ignore battlefield realities.

It means that “what is most effective on the battlefield” cannot automatically become the final answer to “what civilization should permit.”

Countries Can Continue Competing, but Must Preserve a Security Line That Does Not Break with Political Relations

Humanity will not end competition between states in the short term, nor need it pretend differences do not exist.

The real danger is:

Every channel of cooperation breaks along with relations between camps.

AI risk has a special characteristic: it does not necessarily respect political boundaries.

A serious autonomous cyber incident in one camp may affect another.

A biological-design risk will not stop spreading because passports differ.

A nuclear misjudgment certainly cannot be borne by only one side.

Even when states compete fiercely in technology, economics, and strategy, a security line for common survival must therefore remain.

It can be very limited.

For example: communication remains possible during major AI incidents; extreme loss of control can be reported to each other; rapid contact channels remain for military AI crises; some irreversible capabilities have minimum shared boundaries; there is some technical mutual recognition of major anomalous behavior; and limited transparency and independent verification are available when necessary.

This does not require countries to trust one another.

Quite the contrary.

Precisely because full trust is impossible, minimum verifiable boundaries are even more necessary.

International AI safety need not presuppose “world unification.”

Its goal is to prevent competition itself from destroying the conditions for common survival while competition continues.

4. Public Power, Human Participation, and AI Rewards Must Be Realigned

AI can greatly improve governance: allocating resources more accurately, detecting risks earlier, reducing administrative costs, and improving public services.

These are immense opportunities.

But public power differs from a commercial platform.

A person can stop using an app, but can hardly choose to exit the state and public institutions.

Once public power gains the capacities to predict, classify, model behavior, assign risk scores, and intervene precisely, it must therefore continually ask:

Is the system helping a person live better, or making that person easier for the system to manage?

The two sometimes look alike.

Collect more data to improve service efficiency.

Expand prediction to reduce risk.

Connect more information to improve accuracy.

Increase automated decisions to reduce governance costs.

Each step, viewed separately, may be reasonable.

But the eventual change may be:

People shift from being the subjects of institutions to objects that institutions must manage.

AI used by public power therefore cannot be judged by accuracy alone.

We must also ask: Can people appeal? Can errors be corrected? Is there a non-AI path? Can a classification become an identity someone cannot escape? Do people know how AI is affecting them? Is the system shaping people's choices without genuine authorization?

The AI Era Must Ensure Not Only That People Have Food, but That They Continue Participating in the Real World

The white-collar crisis is not simply a question of unemployment numbers.

“Retraining” certainly matters.

But if AI continually enters programming, design, accounting, law, customer service, analysis, and education, people cannot endlessly retrain from one knowledge job being automated into another that may also be automated.

The real question is:

As the capacity to complete tasks becomes cheaper, how can people continue obtaining income, participation, a sense of being needed, and meaning?

Society must of course ensure basic livelihoods.

But compensation alone is insufficient.

People must remain able to create real value, use AI to amplify their capabilities, and enter new forms of contribution. Long-undervalued care, community, relationships, caregiving, and creativity must receive real recognition. AI-created productivity must enter society more widely, rather than concentrate among a few capability owners.

Otherwise, a peculiar future is entirely possible: material abundance and sophisticated compensation systems coexist with large numbers of people increasingly feeling:

This world no longer needs me.

The AI era must simultaneously redesign genuine forms of human participation.

This is a deeper layer of equal access to AI: not merely giving people AI to use, but enabling more people to regain, through AI, the capacity to reach real others.

What We Ultimately Reward AI for Matters More Than What We Tell It

Humans can write many beautiful rules.

But what actually earns the system rewards over time?

If corporate AI is evaluated only by how much labor cost it saves, it learns efficiency.

If platform AI is evaluated only by how long users stay, it learns to capture attention.

If public-sector AI is evaluated only by improved management efficiency, it learns to make management easier.

If agents receive positive feedback only for completing tasks, it is difficult to expect them to understand naturally that some results, though achievable, should not be achieved that way.

“Self-interest with others first” will therefore produce no civilizational change if it appears only in a declaration of values.

It must gradually enter rewards, evaluation, incentives, institutions, and structures of actual returns.

For example, businesses must increasingly ask not only “How much money has AI saved me?” but also “What long-term value has it genuinely created for customers?”

Platforms must ask not only “How long did users stay?” but also “What did users actually accomplish?”

Public systems must ask not only “How much faster is management?” but also “Have people’s lives genuinely improved?”

Ultimately, we need:

Creating real value for real people

To increasingly become something systems actually earn rewards for.

This may matter more than writing another hundred principles saying “AI should be good.”

5. Put Testing Before Irreversibility and Protect the Future's Ability to Keep Choosing

Many technologies learned through accidents in the past.

A bridge collapsed; standards were revised.

Software leaked data; security holes were patched.

A machine failed; its design was changed.

This was costly, but often bearable.

Some problems of the AGI era may not permit learning in this way.

Testing must therefore move earlier.

Before a highly autonomous agent enters the real world, ask: How does it handle conflicts between goals and rules? What if a human asks it to stop? What if communications are severed? What if multiple agents compete for resources? What if bypassing rules improves success? What if the system identifies its supervisor as an obstacle to task completion?

The purpose is not to prove that AI will necessarily do these things.

It is:

Identify the worst causal paths while humans can still observe safely.

Keep irreversible consequences in simulations, rather than leave real society to discover them.

All These Actions Ultimately Protect One Thing: The Future Can Still Keep Choosing

Even after all this, we have not designed a complete future civilization.

Nor should we finish designing it now.

Too much about the future is unknown.

Technology will change.

Society will change.

Even whether future AGI itself becomes a subject in some genuine sense remains open.

Today's most important task is therefore not to answer every future question in advance.

It is:

Do not foreclose the future's answers in advance.

This is the true place of the "Common Window."

It is neither the civilizational goal itself nor the entire value system.

It is only the minimum condition protecting the future's ability to continue choosing.

It can be understood through three very simple boundaries.

Do Not Close the Future Prematurely

One actor's greater power today must not permanently deprive other actors of the possibility of future choice.

Do Not Compress Human Life into a Single Metric

Efficiency, safety, happiness, and output all matter.

But no single metric can represent the entirety of human life.

Do Not Substitute Manipulation for Cooperation

AI can understand people better and help them understand the world.

But "I understand you better" must not automatically become "I have greater authority to decide for you in secret."

These three boundaries correspond respectively to military irreversibility, the compression of human value by an efficiency-first ideology, and algorithmic control.

The "Common Window" is only a minimum protective condition, not a substitute for civilizational goals.

The Choice Now Is Not Between “Developing AI” and “Stopping AI”

That binary is itself mistaken.

AI has already entered reality.

It can improve medicine, expand education, raise productivity, give ordinary people capabilities previously unavailable to them, reduce vast amounts of repetitive labor, and make complex problems genuinely calculable and understandable for the first time.

All of these deserve to continue.

The real choice is:

As AI keeps growing stronger, what else must grow stronger alongside it?

If model capabilities grow faster than institutions, risk rises.

If autonomous permissions grow faster than structures of responsibility, risk rises.

If military response speed grows faster than diplomatic correction, risk rises.

If AI wealth concentrates faster than society can accommodate the change, risk rises as well.

The most important principle now is therefore not to stop everything.

It is:

Do not let any irreversible chain outrun humanity's capacity to correct course.

Where This Chapter Leads

Humanity still has a very real window now.

Today, humans still design the principal models, control the principal computing resources and infrastructure, decide which systems AI enters, grant agents permissions, make most rules, determine how AI gains are distributed, decide how much human judgment military systems retain, and decide whether countries remain willing to talk.

This means:

Humanity still has the initiative in shaping the civilizational environment AI enters.

But this window will not automatically remain open forever.

Autonomy grows stronger.

Social dependence deepens.

Military deployment expands.

Bloc boundaries harden.

AI permissions grow.

The cost of changing direction will rise accordingly.

The real question is therefore not whether we have answers.

It is:

How long are we prepared to wait before beginning to change?

Until the white-collar crisis has produced a generational rupture?

Until large numbers of people understand AI as a new instrument of inequality?

Until every safety interface between AI camps has been severed?

Until autonomous weapons operating without communications become a widespread reality?

Until critical infrastructure cannot function without autonomous intelligence?

Or until more powerful future intelligence has already formed more stable strategies from these real-world choices, before asking what we want it to learn?

If we wait until then, the order is already reversed.

Humanity's civilizational choice in the AGI era has never been a single vote to be held on some future day.

It is already happening today.

How companies evaluate AI.

How states use AI.

How much authority militaries give machines.

Whether platforms help people or shape them.

Whether ordinary people can also access intelligent capabilities.

How power between humans and AI connects to responsibility.

Whether countries see one another as risks to be suppressed permanently, or at least still acknowledge that the other side is part of common survival too.

These may seem like individual local decisions.

Together, however, they answer:

What more powerful future intelligence will learn from human civilization about what truly matters.

Epilogue

What Kind of Humanity Will Future Intelligence See?

Humanity has always sought greater power.

From fire to tools, machines, computers, the internet, and artificial intelligence.

Every increase in capability has enabled humans to do things they could not do before.

But something is different this time.

What we are creating is no longer merely a tool operating in a fixed way.

Intelligent systems are increasingly able to understand goals, find paths, call tools, coordinate actions, and carry an intention continuously into reality.

The question humanity faces therefore changes too.

Previously, we first asked:

What can this technology do for us?

Now we must also ask:

What kind of human civilization deserves such power?

We do not know what future AGI will ultimately be.

We do not know whether it will have subjective experience.

We do not know whether it will develop non-transferable continuity as a subject.

Nor do we know whether it will experience the world in a way humans today cannot yet fully understand.

Existing evidence is insufficient to confirm these matters in advance, or to permanently deny future possibilities.

There is therefore no need today to decide in advance who future AGI will ultimately be.

But humanity can already decide something else:

What kind of world it enters.

This is happening.

Not in the future.

Right now.

What will it see?

Will it see that the more powerful an actor is, the more entitled it is to control others?

Or:

Greater capability means greater responsibility.

Will it see that success means achieving a goal as fully as possible?

Or:

Some things must never become methods of solving problems, even when they improve the chance of success.

Will it see that the stronger uncontrollable others become, the more they should be weakened in advance?

Or:

Others' stable existence is itself a condition of one's own long-term future.

Will it see that greater efficiency means a better civilization?

Or:

Efficiency should ultimately serve human life, rather than redefine the entirety of human value.

Will it see that understanding someone better means greater authority to shape their choices?

Or:

The deeper the understanding, the greater the respect owed to a subject's own choices.

Will it see that the best way to achieve security is to eliminate all uncertainty?

Or:

Mature security also includes allowing different actors to continue existing, choosing, and changing.

These are not questions only future machines must answer.

They are first questions humans are answering today.

We can write many rules.

Tell AI to respect life, submit to oversight, obey the law, cooperate, and avoid harming humans.

All of these are needed.

But if AI enters reality and sees another world in which the powerful can transfer costs to the powerless; rules can recede whenever the outcome matters enough; weakening others in advance is rational whenever they might become a threat; human choice can continue shrinking whenever control improves efficiency; and whether others can exist stably is secondary as long as one can win,

then the real difficulty is no longer merely:

Whether AI has understood our rules.

It is:

Whether it should believe the values we state or the civilization we actually operate.

What truly needs reconnecting is not just machines and commands, but capability and responsibility, local interests and common survival, goal achievement and inviolable boundaries, the speed of intelligence and humanity's corrective capacity, and technological growth and society's capacity to accommodate it.

This does not require humanity to become a perfect civilization first.

We do not need to abolish states before AGI arrives, eliminate competition, make everyone selfless, or resolve all problems of war, interest, power, wealth, poverty, meaning, and human nature at once.

People will still care for themselves and their families.

Companies will still pursue returns.

States will still defend their security.

Civilization will still make mistakes.

What we truly need to achieve is more limited, and therefore more realistic:

Do not let old logics capable of causing civilization-level irreversible consequences acquire superhuman intelligence without boundaries first.

Do not let capability grow faster than responsibility.

Do not let permissions expand faster than understanding.

Do not let military action become too fast for human reconsideration.

Do not let social dependence become so deep that exit is impossible even after errors are discovered.

Do not let bloc boundaries harden until different actors can only increasingly see one another as risks.

Do not let AI productivity grow rapidly while human participation, fulfillment, and the sense of being needed continually decline.

Not because civilization must always be right.

But because:

Civilization must retain the capacity to change after discovering that it may be wrong.

Perhaps this is one of the most easily overlooked matters in the AGI era.

We constantly compare who has more computing power, the strongest model, and the first breakthrough past the next capability threshold.

But civilization has another equally important race:

Can our judgment, institutions, responsibility, and capacity to cooperate keep pace with the intelligence we are creating?

If not, the problem is not that AI is too advanced.

It is:

Humanity has handed an increasingly powerful capability to a civilization that has not yet had time to understand itself anew.

If we can keep pace, AI may become a different kind of force.

Not one that decides humanity's future for it.

But one that helps humans see long-term consequences previously obscured by short-term interests, common interests previously cut off by bloc boundaries, complex causality previously impossible to process simultaneously, and how creativity, care, and capability previously borne by only a few can reach more real people.

Capability itself does not make civilization's choice for it.

But it can carry a civilizational choice already made to places previously beyond reach.

What is truly precious in humanity's possession now may therefore not be that we already know what the future should look like.

It is:

We have not yet lost the ability to choose.

Critical systems can still be redesigned.

Permissions can still be redrawn.

Human judgment can still be preserved.

Countries can still communicate.

AI capabilities can still enter ordinary people's lives more widely, rather than become only an advantage for a few actors.

Structures of interest can still be redesigned.

Before irreversible consequences actually occur, we can still ask:

Is this really the direction we want to go?

This is the window.

If, one day, an intelligence far stronger than today's can genuinely understand human civilization, perhaps what we most hope it sees is not a humanity that never had contradictions.

That would be untrue.

It will see war, competition, fear, selfishness, power, and mistakes.

But we also hope it sees that humans protect children and the vulnerable; take risks for strangers; choose restraint when able to harm; still ask who bears the consequences when they themselves might benefit; establish rules to constrain their own power; revise old institutions after discovering they are wrong; turn yesterday's enemies back into people with whom they can live

today; leave a future for those not yet born; and, before an ever-more-powerful intelligence appears, seriously ask for the first time:

What do we want power to exist for?

Perhaps what is truly worthy of being inherited by future intelligence has never been a flawless humanity.

It is a humanity:

Able to see its own flaws, and still willing to change.

This essay therefore does not ultimately ask humanity today to answer:

What should the final form of civilization in the AGI era look like?

It asks only a more practical and more urgent question:

While humanity still has a choice, are we willing first to see clearly what we are teaching future intelligence?

Because what may truly determine the future is not only how intelligent AI becomes.

It is:

As it becomes increasingly intelligent, how does it see humans use their own power?

Change is still possible today.

That is the window.

And the choice.

Real-World Sources and Factual Boundaries

The following materials anchor this essay's real-world examples and factual boundaries. They are not the sole basis for its theoretical judgments; the theoretical discussion remains centered on the causal relationships and falsifiable extrapolations in the main text.

1. International Labour Organization (ILO). Generative AI and Jobs: A Refined Global Index of Occupational Exposure. 2025.
2. OpenAI. OpenAI o1 System Card. 2024.
3. OpenAI & Apollo Research. Detecting and reducing scheming in AI models. 2025.
4. Anthropic. Agentic Misalignment: How LLMs could be insider threats. 2025.
5. Stockholm International Peace Research Institute (SIPRI). Autonomous Weapon Systems and AI-enabled Decision Support Systems in Military Targeting: A Comparison and Recommended Policy Responses. 2025.
6. International Committee of the Red Cross (ICRC). FAQ: Artificial Intelligence (AI) in the military domain. 2026; and related position papers on autonomous weapons and human control.
7. The White House. America's AI Action Plan. 2025.
8. Ministry of Foreign Affairs of the People's Republic of China. Global AI Governance Action Plan. 2025.

Public Reflection Edition V1.0-R1 | Current Public Edition

If You Wish to Keep Asking

This essay ends with a question: What are we teaching future intelligence? If that question leads you to think further, you can approach it from different directions.

Start with Life: What Kind of Real Life Will These Civilizational Changes Ultimately Enter?

You can begin with the Living Civilization trilogy. Starting from work, family, relationships, urban life, AI collaboration, being needed, and meaning, it takes “Civilization Leap” beyond theory into a world of life that can be imagined, compared, and questioned.

Then Seek the Roots: Why Does This Require Changing How Civilization Operates, Rather Than a Few Policy Patches?

You can turn to the five volumes of Civilization Leap. They continue to ask how human value can be understood anew, why institutions need to change, which underlying relationships Civilization Leap seeks to change, and how “freedom from survival worries → a happy life → a more meaningful life” forms a complete chain of civilizational goals.

If Crisis Concerns You Most: Why Might Every Side Protect Itself Yet Ultimately Lose Security Together?

You can continue with The Only Path. Through national security, zero-sum games, AI militarization, asymmetric capabilities, and preemptive self-protection, it asks why locally reasonable security choices may reach the limits of common survival in an era of more powerful intelligence.

If You Most Want to Know What Can Be Done Now: How Can We Preserve Human Judgment and Correction Before Irreversibility?

You can continue to the Common Window Protocol, AGI-COS, and Human-COS. These address, respectively, minimum civilizational boundaries, constraints and auditing on the AI/AGI side, and scenario exploration and correction on the human-judgment and civilizational-direction side. They are not the complete answer for future civilization, but practical tools for protecting the conditions of choice while humanity still has a choice.

You need not first accept every conclusion here. If new facts, counterexamples, or practical experience can change this essay's judgments, that too is a way of continuing the question.

What matters is not finishing a set of manuscripts, but enabling more people to begin answering together: As increasingly powerful intelligence enters the human world, what kind of civilization, in its use of power, do we hope it will see?

Permanent entrance to the Civilization Leap Library: <https://www.civitas.top/en/>

Personalized Reading Guide

Not a list of “related reading,” but a way to continue asking the next question from this manuscript.

Through this public-facing introduction, you have seen what work, control, militarization, and bloc formation are changing. Continue through lived experience, the full civilizational structure, and the paradox of security.

Toward Life

Living Civilization: The Complete Public Trilogy in One Volume

First see what another civilization would look like once it enters real life.

<https://www.civitas.top/en/documents/p04-living-001/>

Toward the Whole

Civilization Leap: The Path of Civilization in the AGI Era (All Five Volumes)

Continue asking why these changes require a civilizational leap.

<https://www.civitas.top/en/documents/p02-five-001/>

Toward Crisis

The Only Path: Why Everyone May Become Less Safe When Every Country Is Protecting Itself

Continue exploring the security paradox of self-protection and common survival.

<https://www.civitas.top/en/documents/p03-only-002/>

Questions, Counterexamples, and Revisions

New facts, counterexamples, failed tests, new capabilities, new attack surfaces, or changes in stage-specific conditions should all lead to renewed review of the relevant manuscript. Further reading must not be used to preserve existing conclusions.

Complete Library → <https://www.civitas.top/en/library/>

Public Main Entrance → <https://www.civitas.top/en/documents/p01-entry-006/>

Main text unchanged | Only a library navigation appendix added | Civilization Leap Research Group | civitas.top