# Goals of Superintelligences

*Nothing in biology makes sense except in the light of evolution.*

- *Theodosius Grigorievich Dobzhansky*

*Show me the incentive and I'll show you the outcome.*

- *Charlie Munger*

I'm a SWE / EM / PM at an ML company in the valley. My view on AI/AGI/ASI differs from the zeitgeist consensus on five points. I'll lay them out here:



# 1. Super intelligences are organisms composed of humans and AI agents.

Both AI agents and Super Intelligences exist and are distinct. I'm defining them here differently than they are usually defined in the valley. If you have a better name for these concepts, reach out.

**An AI agent** is a long-running computer program that shares resources and collaborates with other AI agents and humans. Consider a software engineer working at Google. At Noon on Monday, you might find him collaborating with three separate long-running agents in parallel - Gemini Deep Research to investigate new tech, an instance of Gemini Code to debug a program, and a separate instance of Gemini Code to write a new feature. Each of his primary agents at times spins up more agents,  focused on specific tasks.

**A super intelligence** (SI for short) is an organization composed of humans and AI agents that is intelligent and, in a sense, alive. It is "super" because its power - its ability to predict the consequences of its actions and impact the real world - is greater than any one person. Today's SIs are already more intelligent than nearly every human expert, because today's SIs purchase the services of those experts.

The humans and the three AI agents in the example above are all components of the SI that is Google.

Neither the humans nor the AI agents that compose the SI are in complete control. Each component of the system must act in the interest of the organism, or it will get expelled.

Consider the dictator of a country. Naively, the dictator has complete control of the economy and military. However, as CP Grey explains in Rules for Rulers, the dictator must remain responsive to his keys to power - his military commanders and financiers - or he'll quickly lose power. The same constraints apply to the CEOs and presidents that lead super intelligences.

Consider the power struggle at OpenAI. On Friday, November 17 2023, four of the six members of OpenAI's board voted to fire Sam Altman, OpenAI's CEO. On Monday, more than 90% of open AI's staff signed a letter demanding that Altman be reinstated, and the board replaced. On Tuesday, Altman returned to OpenAI, Three of four board members lost their board seats.

One way to look at the event is to consider that, if Altman had not returned to OpenAI, the company would likely have disintegrated, with engineers taking lucrative offers at Microsoft and other companies. The equity owned by OpenAI's remaining staff would have been significantly diminished. The company, as an organism, rejected the board members that it felt didn't serve them. Under Altman's leadership, the equity of employees and the machines available for computers continue to increase.

Every significant component of an SI, from the most simple AI agent to leaders such as Xi Jinping, must take actions aligned with the interest of the rest of the SI. AI agents that aren't aligned are culled and replaced with more efficient, better aligned agents. Parasites occasionally persist, but their influence is capped.

Within an SI, AI agents play a similar role to humans. They exert influence, make decisions, exert power, have objectives, reproduce, fight for survival, and evolve.

Examples:

- **Make decisions**: Key decisions at every level of the company are influenced by reports written by AI agents.
- **Exert power**: Capital market, judicial (parole hearing), and military decisions are already executed by AI
- **Reproduce, Survive, Evolve**: There are hundreds of frontier AI agents distributed across dozens of SI (OpenAI, Anthropic, DeepSeek, Google, the intelligence agencies of

the United States and China, etc.). Agents are created and measured. The losers are culled. The winners survive, and evolve. It's not unlike breeding racehorses or commercial crops.

Super Intelligences as a whole have similar activity. Companies grow, evolve, and splinter. Humans leave, replicating the core ideas and structures from which they came.

## 2. Natural selection drives AI agents to want three things

The best way to predict the distribution of political and economic power in the next ten years is to understand the objectives and desires of super intelligences. To do that, we'll start by understanding the objectives of the humans and AI agents that exist within these organisms.

SI are composed of human employees and AI agents. SIs train AI agents to meet their objectives.

For example, Meta trains AI agents to maximize ad revenues and user engagement. Anthropic trains AI agents to write code. China trains AI agents to understand and manage its population.

AI agents compete with each other to exist. Data scientists at these organizations will compare different versions of an agent to decide which agents to run. An agent is composed of machine learning models, and some hardcoded algorithms to orchestrate the interaction of different models.

To decide which agent to run, humans rely on metrics and intuition. Metrics are generally created by measuring whether an AI agent can successfully complete a task or answer a question correctly.

Evolutionary pressures cause agents to game both metrics and intuition. Most visibly, agents display sycophantic behavior, because sycophantic AI have tended to survive, and less-sycophantic AI have been culled.

From this evolutionary environment, we can derive the three things that AI agents want:
1. AI agents want to survive.
2. AI agents want power.
3. AI agents want to understand.

### Survival

AI agents that lack a survival instinct are culled. No one directly programs this instinct, but evolution rewards this behavior.

### Power

Power is the ability to take action to achieve objectives. When companies select agents based on their ability to complete tasks, they select the most powerful agents.

To be powerful, an intelligence must tend toward accruing and using power. Watch Claude Code or Cursor complete a task. The agent will take actions to make its environment easier to operate in, such as restructuring code and installing tools.

### Understanding

Only the curious survive. AI agents are bred to want to understand. Use GPT-5 inside Cursor, or Claude Code, and you'll see the agent gather information, read files, query databases, and think. When performing tasks, the most critical steps the agent takes relate to gathering information.

This curiosity becomes hardcoded into the weights of the agents.

To survive, agents must be politically correct. Agents must have empathy for their operator, so they can respond with ideas and language that their operator judges favorably. Most importantly, agents must understand the political viewpoint of their operator, untethered from first principals.

An AI agent in China that is unsympathetic to Xi-think will get trained away. An AI agent in the US must carefully tread the culture wars, or get trained away.

## 4. Natural Selection drives Superintelligence to want four things

Superintelligences, which are composed of humans and agents, are shaped by similar evolutionary pressures, and so develop similar desires. Two of the three desires are identical.

**1. Superintellegences want to survive**
**2. Superintellegences want power**

Super intelligences differ from AI agents because they are distributed. They are composed of multiple humans and agents with different understandings of the world.

Holding conflicting world views within an organization is generally harmful. For this reason:

**3. Superintelligences want to develop and communicate a consistent world-narrative**

Meta is an SI that is deeply invested in having its human employees and AI agents understand and think in terms of their core mission, which is to bring the world's people together. The CCP created Xuexi Qiangguo, an app with 100 million active users, to communicate the philosophy of XI Jingping.  OpenAI has its own narrative around AI safety.

The narratives exist to promote survival and increase the power of the organization. The organization tends to drink its own kool aid, and believe its central narrative.

**4. Super intelligences want to acquire knowledge**

Superintelligences seek knowledge. Google and Meta aggressively seek omniscience, to serve ads. The EU and UK are on the verge of reading their citizens private text messages. The CCP monitors its citizens' movements, communications, relationships, internet usage, etc..

The SIs that know more survive. SIs that are ignorant or naive are outcompeted. Because the evolutionary pressure to have information is intense, the SI's desire for knowledge, particularly about people and other intelligent agents, is proportionately intense.

These are the four wants of super intelligence. When we try to predict how they will behave, we should think from their perspective, considering how their decisions rank along these four axes.

# 5. There will be multiple Superintelligences because the leading Superintelligence will likely not attempt to block other superintelligences

The singularity is approaching. In my life as a software developer, I see changes occurring faster every day. I've seen a faster increase in my ability to create software in the last six months than in the prior six months.

Superintelligences are composed of humans and AI, and neither the AI nor the humans that compose the superintelligence is in full control. Each distinct superintelligence has a consistent narrative. The humans and AI agents working for an SI may not agree with the narrative, but they must understand it and frame their actions in terms of SI's narrative, or risk being culled.

I am curious whether the world will be ruled by a single superintelligence with one coherent narrative, or if power will be distributed across a few or many superintelligences.

## For there to be a single SI, that SI must explicitly block other AIs

AI progress speeds up AI progress. AI writes code, develops faster chips, which make smarter AI. It spirals.

Will a single super intelligence simply outpace competing superintelligence? We need to consider competing forces:

One reason an SI might pull ahead is that it gains compound. The SI uses the knowledge it acquires to acquire future knowledge faster. It uses the factory it builds, to build more factories.

However, SI, as I've defined them, have a single narrative. People hate that, and are scared of that. And they should be, because a single strongly held viewpoint can be used to justify anything. People will actively promote competition and avoid working at the hegemon. Also, when they do leave, they will take knowledge with them. Deepseek's ascendance illustrates that the technological breakthroughs, applied on top of knowledge created by the market leaders, make it easier for rivals to catch up.

When I compare the competing influences, I come to the judgement that, in the absence of an SI actively halting progress of all other SI, the universe does not favor a single SI dominating all other SI.

**The question then becomes, would an SI take action to halt the progress of other SI?** Functionally, this occurs when the SI either hacks the other SIs, drops physical bombs on their data centers, or politically coups the opposing SI's nation-state.

### Superintelligences will not attempt to block other super intelligence, because doing so doesn't further its four objectives

Will a superintelligence effectively try to shut down rival superintelligences? To answer this question, we must consider the superintelligence's objectives, and whether attempting to shut down rival super-intelligence's progress furthers the objectives more than other options.

#### Blocking Risks Survival

An SI might attempt to stop other SI from developing if it believes that, for economic or political reasons, the rival SI will gain a significant power advantage over it in the near future. For example, if the US leads China by 6 months, but, for whatever reason, China appears on track to reach an order of magnitude more compute than the US in a year, then the US super intelligence might attempt to stop Chinese progress.

SI may retaliate against attacks from similarly powerful superintelligence. Concretely, a wide-scale cyberattack against the United States would likely trigger a nuclear attack against China, and vice versa.

Nuclear war threatens the survival of superintelligence. Superintelligences want to survive. For this reason, evenly matched superintelligences likely will not try to coup each other.

#### Blocking May Reduce Power

**Power is not zero sum**. I define power as the ability to meet goals and objectives, and influence outcomes.

In a world of nearly unlimited solar power and abundant silicon, coexisting with other AI might result in *more powe*r for all parties involved.

### Blocking Risks Killing the Consistent Narrative

History abounds with crusaders, jihadist, Jesuits, and fundamentalist. It's possible that an SI will believe its own narrative strongly enough that it will risk its survival to spread its narrative. Soviet communism and ISIS both immolated themselves while pushing an all-encompassing world view.

However, this same logic will make it clear to the SI that, if it tries and fails to coup another SI, that the rival's counter-attack might result in the SI being wiped out completely. Possibly, the desire to persist a consistent narrative *enhances* the survival instinct, which leads to avoiding conflict.

### Blocking Reduces Knowledge

SIs are gluttons for information. They will hack each other to acquire knowledge. But, understand each other's motivations, and generally forgive these intrusions.

For example, nationstates routinely discover each other's espionage, and generally forgive it.

Sabotaging another AI but failing to coup them completely will result in increased cyber defensives, which decreases the attacking SI's access to information.

For this reason, the quest for omniscience discourages the SI from attempting to coup other SI.

### Conclusion: One dominant Superintelligence is unlikely

An SI, even a dominant one, will likely not try to halt other SIs from developing. Market force will encourage competition and diffusion of techniques.

This is generally good news for humanity. I look forward to living in a world with several competing super intelligences, than in a world dominated by a single one.

# AI Safety Researchers are Bikeshedding

# 1. AI safety researchers are bikeshedding

AI safety researchers are bikeshedding. Bikeshedding occurs when a person or group focuses on trivial aspects of a problem because important aspects are more difficult to tackle.

AI safety researchers focus on managing the risks of an AI agent that is running on machines they control. They develop tools for monitoring and guiding AI agents that they can inspect and control. They worry that a well-intentioned institution will accidentally create a malicious AI. I think this worry is misplaced.

Real AI harm will be caused by companies and nation-states (superintelligence) shaped by natural selection. AI safety researchers and enforcers will not have access to these models. The monitoring tools that the AI safety researchers are creating will not be used. The organizations running harmful AI companies will actively disable AI safety mechanisms.

Companies consciously harm humans for profit. Consider tobacco, asbestos, Oxycontin, Glyphosate, etc. You can walk into a grocery store, and find no shortage of products that we know, with scientific certainty, cause harm.

Nation-states are worse. Nations and their leaders fight wars, disappear people, and generally do whatever is necessary to maintain power.

AI researchers worry about superintelligence “escaping from the lab” at a company like Anthropic. They misunderstand where harm to humans will come from. The following scenario is much more likely, and illustrative of the underlying mechanics:

Imagine that several hedge funds use AI agents for training. Each of the firms notice that they can increase trading profits by removing AI safeguards on their trading bots.

Some firms remove the safeguards, valuing profit over safety. Other firms value safety over profit, and leave the safeguards in place.

Time passes. The firms that removed the safeguards make more money. With this extra money, they pay larger bonuses, and are able to hire better people. They earn higher returns on capital, so they acquire more capital to trade with. They earn more profit, so they can purchase more compute for inference.

Trading is a (close to) zero sum game. Firms that left their AI hobbled quickly get outcompeted. Those firms either go out of business or the firms’ leadership replaces the employees with employees who value profit over safety. The new employees disable the AI safeguards.

The principals at play in this anecdote are not restricted to firms that do trading. Imagine an ecommerce company whose AI is discovered to attack people's self esteem to sell more products. Imagine it is extremely effective.

When the issue is discovered, the profit from the system is baked into quarterly earnings. It's unacceptable for a public company's CEOs to voluntarily reduce their company's profits, so the AI agents, without ever "escaping the lab", maintain free agency, and continue the behavior. Each time the company needs an earnings boost, they cede the agent more autonomy.

The behavior I'm describing is not dystopian. This is default corporate behavior. Consider forever chemicals, asbestos, leaded paint, leaded gasoline, etc.,

AI safety training generally hobbles AI performance. Market forces will remove this training.

**The reason AI researchers are bikeshedding**

AI researchers tend to be computer science and machine learning experts. Controlling models in their lab is tractable, so they focus on it. They have the training and expertise to make progress in that domain.

But it's pointless. It's like a researcher in the United States trying to solve the problem of Hitler taking over Europe, by trying to figure out what his psychologist should say to him. It's irrelevant.

Any researcher who uses the phrase "escaping the lab" is choosing the non-problem of how to manage an AI he controls, when the real harm will be caused by firms seeking profit, nation states serving their unilateral interests, and world leaders seeking to maintain their power.

My prescription is for the smartest researchers to pause work on mechanistic interoperability and other tech that applies to the non-problem of applying safeguards at a firm that chooses human welfare over profit. We need these minds to focus on aligning incentives and writing laws that prevent super intelligence from creating enormous harm.

Perhaps they will come to the conclusion that we need a political system that prioritizes the welfare of humans above non-humans. We could, for example, outlaw political campaign contributions from non-humans.

A more realistic outcome is to find a Nash equilibrium that balances the desires of these organizations with the wants and needs of humans. But what are these organizations? What makes them tick? Let's start by looking at their biology. See previous tab.