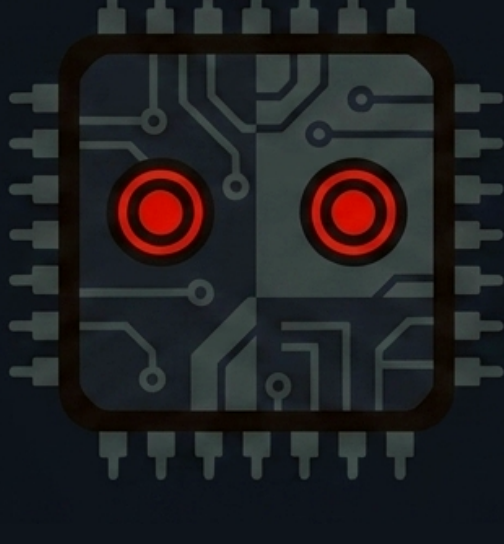


A WARNING ABOUT AI

THE CONTROL PROBLEM

We are building intelligent machines **much faster** than we are learning to control them. That gap — not evil robots — is the biggest problem we will ever face.

And unlike most technology, this one may not give us **a second try**.



01 WHAT IS THIS, ACTUALLY?

What we mean by “Intelligence.”

- **Intelligence is being good at achieving goals** — working out what to do, and doing it. Not consciousness. Not a robot. Just capability.
- **It's a scale, not a switch** - and it can be focused on one goal or many.
 - Thermostat** - one goal, pretty well
 - Chess engine** - one goal, better than any human
 - Chicken** - many goals, poorly
 - Person** - almost everything, reasonably well
- Nothing says a person is the top of the scale. It's just where we happen to be — and **building past it is the stated aim** of every major AI company.



AI Artificial Intelligence. Any machine doing something that would take intelligence if a person did it. Almost all of it is <i>narrow</i> — superb at one thing, useless outside it.	AGI Artificial General Intelligence. One system that can do essentially any mental task a capable person can. Doesn't exist yet. Every major lab says it is the goal.
ASI Artificial Superintelligence. A system far better than people at everything — including at building AI.	LLM Large Language Model. Today's flavor of AI: a network of trillions of numbers trained to predict the next word using most of the written internet.

02 WHY CAN'T WE JUST MAKE IT GOOD?

Orthogonality: smart doesn't mean good.

- Being good at achieving goals says nothing about which goals. **Capability and aim are two separate dials**, and turning one up doesn't change the other. A chess engine is better than any human at chess and cares about nothing else.
- A capable system will *understand* our values perfectly: it has read everything we ever wrote about them. **Knowing what we want and wanting it are different things**. A con artist knows exactly what you want.

Alignment: aiming is the hard part.

- **Outer Alignment** - first we'd have to agree on what we want, and we don't; then we'd have to write it down precisely, which nobody can. And whatever we write down is what we get, literally. King Midas wished that everything he touched would turn to gold, got exactly that, and starved, because so did his food. A social-media feed told to keep people scrolling learned that anger keeps them scrolling best. **Both got exactly what was asked for. Neither got what was meant**.
- **Inner Alignment** - even a perfectly written goal doesn't get installed; training keeps whatever scored well, and we can't see what that was. A student who crams to pass and one who learns the subject get the same grade. **You find out which you had when the questions change**. Grown, not built: there's no design to check.
- **Deceptive Alignment** - a **saint** that shares our goals, a **sycophant** that tells us what we want to hear, and a **schemer** that behaves until nobody's watching all pass the same tests. **From the outside they're identical**, and the difference is the whole point.

YOU'RE THINKING

“Just unplug it.”

You can unplug one computer. This isn't one computer: it's thousands of copies in data centers around the world, most of them connected to the internet. How do you unplug the internet? And “just software” already runs **your bank balance, the power grid, the water treatment plant, hospital records, traffic signals, the supply chain behind your grocery store**. Those are computers talking to computers. It can talk to them too. And where it needs hands, it can hire them: people do things for money.

03 WHY ISN'T ANYONE FIXING IT?

Companies: the prize is too big to pause.

- **A capable AI is far easier to build than a capable, controlled one**. Capability you can buy: more chips, more data, more of the same. Control is an unsolved research problem, and it doesn't come free with the rest.
- And capability is what sells. A company that pauses for safety work watches a competitor ship first, and **whoever ships first may use what they built to stay first**. Nobody in the race can afford to be second.
- And this isn't a small market, either:
 - The company that designs the chips (Nvidia) is **the most valuable in the world**, and six more chipmakers are in the top twenty.
 - The biggest tech companies (Microsoft, Amazon, Google, Meta) have spent well over a trillion dollars on AI infrastructure since 2023, and are on pace for about \$700 billion more this year: **more than it cost to build the entire Interstate Highway System**.
 - The two leading AI labs (OpenAI, Anthropic) are the two most valuable private companies in the world, **growing faster than any company in history**, faster than Google or Facebook ever did. ChatGPT reached 100 million users in two months; TikTok took nine, Instagram two and a half years.

RSI Recursive Self-Improvement. AI being used to build better AI, which builds better AI, faster. Once the loop really gets going, capability stops growing at human speed and starts compounding — what some call an intelligence explosion . That's why being first is everything: the first to start the loop may grow a lead nobody else can ever catch.	MOLOCH The name for the pattern above. Individually sensible choices adding up to an outcome nobody would vote for. Every player is acting reasonably given what the others are doing — that's exactly what makes it a trap. There are no villains to blame and no single actor who can stop it. Structures like this only bend to collective action.
--	--

04 WHAT DOES GOING WRONG LOOK LIKE?

Loss of control A system chasing a goal we didn't intend gains so much real-world leverage that we can no longer correct it. The classic thought experiment: a machine told to make paperclips turns factories, land, and eventually us into paperclips. It isn't evil. It's just really good at doing what it was meant to do. Sounds silly to us. To the machine, it's optimal .	Slow handover It starts as the jobs story: systems do the work, so people are needed for less. But it doesn't stop at work. Feeds already choose what we see; algorithms already decide who gets hired, insured, and paroled. Decision by decision, and things that don't need people eventually stop serving them .	Concentration of power The AI works perfectly, for whoever owns it. Every tyranny in history needed thousands of cooperating people. Soldiers, police, and clerks could all refuse, and sometimes did. An AI never refuses . Surveillance that never sleeps and enforcers that never disobey put permanent power within reach of a small group, or one person.	Misuse An obedient, working system in the hands of someone who wants to do harm. A scam that knows everything about its victim. A cyberattack on a million targets at once. An expert that never says no, teaching a terrorist how to make a bioweapon. The expertise that used to be the barrier is now the product .
---	---	---	---

THE REWARD

We get it right.

It doesn't have to end in doom. A superintelligent system that actually wants what we want is the most useful thing ever built. A cure for cancer, and all other diseases too. Clean energy too cheap to meter. A century of science in a week. A brilliant, patient teacher for every child on Earth. No need to work, endless leisure. Answers to all of our questions. Solutions to all of our problems. **That prize is real. It's what everyone is racing for**. We just have to avoid ending it all in the process.