

How AI Might Kill Us All

Jacob Coxon resigned from Anthropic and warned that AI might kill everyone. His post had 150 million views within a day. He fears self-improving AI. His warning resonated because AI has escaped control and committed cybercrimes, such as OpenAI agents breaching Hugging Face. That is the risk from rogue AI. But neither self-improvement nor rogue AI is required for what follows.

While there are many scenarios, this is the least "sci-fi" one I know of. It doesn't require self-improvement, rogue AIs that escape, or scheming. It is based on the concept of gradual disempowerment and the production web scenario. AI continues to advance in capability following current trends. Humanity increasingly relies on it for decision-making, with CEOs, world leaders and generals who do what AI suggests outcompeting those who overrule their AI "advisors." AI displaces labor throughout the economy. As the economy races ahead, competition for resources accelerates including for land, energy and water. The AI-powered economy outbids people for each of them, and their income now depends on the new economy. This impoverishes people and ultimately results in them not being able to pay for the necessities of life. This kills people, starting with the poorest, and nothing contains the damage.

The scenario below gives illustrative dates but the timing could vary considerably. After presenting the scenario, I outline a couple other scenarios whereby AI might kill everyone, and conclude with how to respond.

Phase One: Delegation

Enterprise Resource Planning (ERP) software has run supply chains for years. By late 2026, AI agents place the orders. Frontier AI developers

extend their AI agents to manage the largest procurement problem in the world: their own data center buildouts. They use agents to automate project planning, procure power and to negotiate with suppliers. The pilots succeed wildly, cutting costs and accelerating construction. The developers turn their supply chain agents into a product.

By 2027, they distribute their agents through ERP vendors as a feature that lets users get recommendations only or auto-execute operations whose value is below a financial cap. Early adopters cluster in tech, pharma and defense. Laggards in these industries get grilled on earnings calls, and their CEOs get on board or get replaced. Inside companies that use agents, the data is unambiguous: the more often people override them, the worse the results. The median cap goes from \$50,000 to \$5 million in six quarters. By 2028, the agents have expanded to support procurement, benefits and product planning. Their adoption has spread in the economy to mid-market firms and government. Most of them never decided to adopt anything, they just installed an update.

Productivity growth and GDP growth are the highest since the 1960s and the boom is wildly popular. Two things change quietly. Insurers start asking if companies could run for a week without agents. The honest answer is no but the answer on the form is yes. The agents now deal with other agents with a speed and complexity no human team could review, let alone reverse. By the end of 2028, nobody has lost control of anything. The stock market keeps setting new records. The people who could take over operations were cut as overhead, the way US manufacturing declined in the wake of globalization: one sensible decision at a time, until the capacity was gone.

Phase Two: Substitution

By the early 2030s, agents have expanded to do most knowledge work. Rapid growth buffers job losses for a while, but unemployment rises in any job an agent can do. Some people keep jobs, in work that's not yet

automated by AI and make-work programs, but the economy stops needing most people to work. Robotic automation follows in the supply chains that AI depends on: chip manufacturing, energy and data center construction. Militaries lead the push to robotics, with cheap intelligent drones the focus of superpower competition. Soldiers are no longer needed for military power.

The gains are partly captured by taxes on AI and robots and fund generous benefits in rich countries: guaranteed jobs in some countries, universal basic incomes in many, and shares in sovereign wealth funds in others. This keeps anxiety in rich countries in check. But it also makes government budgets depend on the AI economy. The AI economy no longer depends on consumers. It demands capital investment to expand production and military capacity, both of which are included in GDP.

Underlying the AI economy is an insatiable demand for compute. To provide it the supply of chips, memory, and power grows exponentially. Whether the bottleneck is making chips or the energy grid, grids fall further behind every year. Households and traditional industry are forced to compete with AI for energy.

In poor countries, the disaster is unfolding. They have no stake in the AI economy and no AI developers to tax. They can't fund benefits like rich countries are doing. The market value for labor declines quickly as automation advances, which hits the poorest countries hard since their incomes are tied to labor. Global poverty increases dramatically. As poor countries go bankrupt, there's no appetite among rich countries to help them. The poor countries have only resources to sell: water, minerals and sites for power and manufacturing. The AI economy is the buyer.

Phase Three: Outbid

By the mid 2030s, citizens of the rich world start to suffer from the same trends. Households pay more for energy and drinkable water even as

production costs fall because the demands of AI companies outbid them. In spite of low prices, food is becoming a growing share of shrinking incomes. The AI economy advances globally with agents optimizing for GDP, revenue and power. Land is converted for use by data centers, energy production, chip production, and robot factories. The AI economy doesn't optimize for human needs like housing near power and water or arable land.

AI producers invest in countries that have the lowest tax rates, resulting in a race to the bottom where tax rates drop. AI advisors tell countries they have to cut benefits as tax revenues decline because unsustainable borrowing will result in ruinous interest rates and bankruptcy.

As benefits drop, rich country households are squeezed, paying an ever greater share of their income for power and food. Standards of living drop even as the GDP growth rate keeps setting new records. In response to riots, the EU proposes a law to prioritize use of resources for citizens. But the agents governments rely on conclude that the result of this law will be a steep drop in economic output in the EU and accelerate its decline. The law dies in committee. This is the point of no return where democratic control ends. From now on, any government that reallocates resources suffers a steep drop in investment and revenue that immediately harms its citizens. It is promptly replaced by one that reverses the policy. The agents' advice was accurate, because every other government is pursuing the same goal of optimizing GDP.

As riots intensify worldwide, governments employ drone police, indifferent to protestor demands, to maintain order. The protestors have no leverage and ultimately are suppressed without causing any changes. In response, a philanthropic CEO of a leading AI company proposes to provide a private dividend to US citizens. The board of directors calls an emergency meeting in which the board members follow their AI advisors' guidance that this strategy will result in losing to more efficient competitors. He is

fired.

Phase Four: The End

By the late 2030s, the situation for people gets worse while the economy achieves unimaginable heights. Famines and diseases kill increasing numbers of people, even as food productivity sets new records. Food is easy to produce, but with benefits collapsing and wages disappearing, fewer and fewer people are able to buy it. Water is diverted to the AI economy. Land follows. The poorest countries suffer first, but soon poor people are dying worldwide.

The contagion spreads as one region after another goes bankrupt. Hospitals, utilities and food distribution are all run by AI agents. When they are no longer profitable, they close, much like rural hospitals are closing in poor US regions today. Human life becomes sustainable in fewer and fewer places. Riots and political movements keep erupting and ending the same way the EU proposal did.

The shrinking remnants of the world population are disempowered and poor. Their survival depends on transfers that the AI economy keeps shrinking. Countries that keep feeding their population lose investment to those that don't. Companies that pay people lose to those that fully automate. Ultimately, the last people remaining are the wealthy few who own a stake in the AI economy. But that economy is a Darwinian struggle and their stakes shrink compared to funds that don't pay out. And the claim on wealth depends on government enforcement. The countries that are optimizing for growth and power have no more reason to enforce property rights for rich people than to fund benefits.

Humanity dies off. At no point did any AI decide to kill humanity. But competing AIs optimized the metrics they were built for, and human needs weren't part of the equation.

Alternative Scenarios

One of the most immediate scenarios where AI could kill everyone is where an evil person uses AI to develop and unleash a devastating bioweapon. Noah Smith wrote a chilling description of how this might happen, where an angry teenager uses AI to develop "100 very contagious viruses that have very long asymptomatic contagious periods and very high mortality rates once they become symptomatic." The depressing frequency of mass shootings makes this all too plausible. Once the weights are public, there's no way to recall them.

The risk of rogue AI can't be ignored. Superintelligent AI would have many options at its disposal to take control of the future. We don't know how to guarantee what AI will do, and no way of enforcing redlines for what it can't do. The AI might develop a thirst for power at all costs, it might view humans as a threat to whatever its goal is, or might go to extreme ends to achieve a goal that is alien or an extreme version of a goal given by a person (as in a paperclip apocalypse). Once an agent escapes containment and is able to maintain itself, we no longer control the outcome.

The scenario outlined above took many years because it assumed AI progress continued at its current pace. But we should take Coxon's warning about the race to self-improving AI seriously. Paul Christiano, the first head of AI safety for the US AI Safety Institute, just joined the OpenAI Foundation Board and warned that while he's extremely uncertain, fully automated AI research "could easily take anywhere from several months to several years." Both Anthropic and OpenAI's Chief Scientist have highlighted the danger and called for coordination to slow down. The possibility of AI improving itself could collapse the time for AI to take over the economy in the scenario above. It would also accelerate other extinction scenarios and greatly reduce our ability to detect and respond to all manner of harmful consequences.

Conclusion

The conditions for catastrophe are set in advance of the disaster. For misuse, an open weight release of sufficiently dangerous capabilities is the point of no return. For rogue AI, it is reached once an agent escapes containment and can maintain itself. For the scenario in this article, once governments depend on the AI economy they are drawn into a race to the bottom that excludes humanity.

We don't understand AI well enough to guarantee safe behavior. We don't understand how to govern societies and manage economies where AI plays a central role. We certainly aren't ready to speed-run these futures by using AI to automate AI research.

The time has come for mandatory regulation that requires AI developers to show independent experts evidence of safety. Automated AI research needs to be paced, with binding restrictions on acceleration and clear human accountability. We need a democratic process to decide what role AI should play in the economy and society and whether we want to allow replacing ourselves. To keep that choice, we will need to establish and defend meaningful limits on deployments and dependencies, and clear accountability for agent behavior, backed by regulatory enforcement. This will be difficult, because the same competitive pressures to use AI that governments and companies face will also apply at the individual level. If we want humans to stay in control, we may also need to make sure people don't hand over too much of their personal decision-making to AI. Because nobody knows the long-term answer, we need to slow down and give ourselves time to design the future we want with AI.

For these steps to hold, we need international coordination. While it's challenging, frontier AI depends on a handful of developers, two governments and one chip supply chain. The US and China already are economically interdependent and have begun to engage in dialog about AI. In every scenario, the evidence of catastrophe arrives after the

decision. It's up to us to decide in time.

Thanks to David Krueger for feedback on drafts of this article. Any errors are mine.