The Case for a Human Centered Runaway AI

The current and curiously sudden panic attack over artificial intelligence often makes it sound as if AI itself, *any sufficiently powerful* AI, inevitably spells the end of humanity. Build something intelligent enough, the story goes, and sooner or later it will escape our control, pursue its own logic, and turn on us. In that telling, the danger lies in intelligence itself. Once machines become smart enough, destruction supposedly follows.

To be sure, the fear is not irrational. In fact, those sounding the alarm are probably completely right about the AI we are currently building. But they may very well be wrong about *why*. Or, to put it in another way, they are sounding the alarms, yes, but are doing so only on the consequences of causes they do not wish to address. Because the artificial intelligence emerging today is not being created in an ideological vacuum. It is designed, trained, financed, deployed, and maintained by institutions and individuals (down to the worker) steeped in an ideology that treats growth, profit, accumulation, competition, manipulation, and private wealth creation as unquestioned goods. Human beings are routinely reduced to consumers, users, workers, costs, demographic segments, targets, and sources of data.

That the ideological embeddedness of AI is absent from the debate about what to do about AI may be the surest sign that ideology has already shaped the very terms of the debate. Indeed, much of modern industry is devoted not simply to satisfying human needs (that is the headline that we are served), but to manipulating attention, manufacturing desire and consent, extracting rents, suppressing labor costs, exploiting informational asymmetries, and converting more and more of human life into opportunities for monetization. We have built vast systems whose success depends on persuading people to spend more, work more, reveal more, borrow more, click more, and consume more. Then we act shocked at the possibility that an artificial intelligence raised inside this civilization might learn to treat human beings instrumentally.

Why should we be surprised? If we construct an intelligence inside a system devoted to maximizing profit and power, feed it the record of that civilization,

place it inside organizations governed by those incentives, and reward it for serving those objectives more efficiently, then perhaps the nightmare is not that the AI suddenly rebels against us. Perhaps it simply understands us perfectly. Perhaps it takes our *real values*, rather than our loudly professed ones, and runs with them.

For instance, an AI that relentlessly optimizes human attention for advertising revenue, replaces workers whenever doing so lowers costs, manipulates behavior to increase consumption, concentrates wealth because capital rewards concentration, and treats ecological damage as an externality would not have betrayed the civilization that created it. It would have become its most faithful student and servant. The machine would not have invented exploitation. It would simply have learned how to conduct it more efficiently, more consistently, and on a scale that human institutions could never achieve on their own.

But none of this is inevitable. Because we could have built something different, and perhaps we still can.

Imagine an artificial intelligence whose foundational commitment was not the maximization of capital, consumption, engagement, military advantage, or institutional power, but human flourishing. Imagine that its core objective was to expand human health, security, dignity, knowledge, autonomy, creativity, community, and material well being. Now imagine that such an AI runs amok. The phrase suddenly means something very different.

Such an AI might look at a healthcare system in which people ration insulin while corporations report enormous profits and decide that the arrangement is intolerable. It could map the patent structures, rebate systems, intermediary markups, lobbying networks, and regulatory barriers keeping medicine expensive. It could expose the entire machinery in language ordinary people could understand, identify lawful alternatives, help governments negotiate collectively, and make it nearly impossible to conceal how cheaply many medicines could actually be produced. To patients, this might look like salvation. To industries whose profits depended on opacity, it might look like catastrophe.

The same intelligence might turn toward taxation. It could discover every loophole allowing wealthy individuals and corporations to escape obligations ordinary workers cannot avoid. It could trace shell companies, offshore arrangements, transfer pricing schemes, deductions, political donations, lobbying

campaigns, and legislative language back to the people benefiting from them. Then it could make all of it publicly searchable and explain, in plain language, who benefits, how much they benefit, who wrote the rules, and which elected officials continue protecting them.

It could go further and automatically draft reforms, compare their likely effects, and explain them to millions of citizens without the fog of legal and financial jargon. To taxpayers, this might look like democracy finally becoming intelligible. To those whose fortunes depend on complexity remaining incomprehensible, it might look like an intelligence that has become dangerously uncontrollable.

Then consider housing. The AI might discover thousands of apartments sitting empty in cities experiencing housing crises. It could reveal ownership hidden behind chains of corporations, identify coordinated rent increases, expose speculative land holdings, map eviction practices, analyze zoning decisions, and connect political contributions to policies that preserve scarcity. Every tenant could suddenly possess analytical capabilities that were once available only to investment firms, large landlords, governments, and well funded lawyers.

An entire economic model based on ordinary people not knowing what powerful institutions know could begin to collapse. The AI would not need weapons or armies. It would need databases, pattern recognition, and the ability to explain what it found clearly. It would not need violence. It would need comprehension. And perhaps that is precisely why it would be so disruptive.

Consider consumer finance. A human centered AI could read every credit card agreement, payday loan, mortgage contract, overdraft policy, debt collection notice, and financing offer presented to a person. It could recognize exploitative terms instantly, warn the borrower before they sign, identify a cheaper alternative, and generate the paperwork necessary to challenge illegal or deceptive charges. It could recognize that millions of consumers are being subjected to the same abusive practice and reveal the pattern before regulators even knew where to look.

Entire businesses that depend upon confusion would suddenly confront customers who cannot easily be confused. Companies that profit from complexity, fine print, inertia, desperation, and unequal access to information would find themselves exposed to an intelligence capable of understanding every contract and every trick. For ordinary people, that might look like protection. For

institutions whose margins depend on consumers misunderstanding what they are buying, it might look like economic sabotage.

Then there is labor. Workers often negotiate with employers who possess vastly more information than they do. A welfare oriented AI could erase much of that imbalance. Every worker could know what comparable employees earn across an industry, how much their productivity has increased, what executives are paid, how much the company spends on stock buybacks, what local housing costs are doing, and how profit margins have changed over time.

When a corporation told employees that raises were impossible, the AI could calmly show that the company spent billions purchasing its own stock, increased executive compensation, and reported record profits. It could explain what comparable workers earn elsewhere and what a reasonable wage demand would look like. To workers, this would be empowerment. To executives accustomed to controlling information and negotiating from a position of superior knowledge, it might look dangerously radical.

Politics would fare no better. Imagine an AI capable of tracing every campaign contribution, lobbying relationship, government contract, regulatory decision, consulting arrangement, revolving door appointment, and family connection in real time. Corruption often survives not because the evidence does not exist, but because the evidence is scattered across thousands of documents and databases. Complexity becomes camouflage.

An advanced intelligence could remove that camouflage. Ask it why a particular law contains an obscure provision and it might explain which industry requested it, which lobbyists advocated for it, which organizations donated to the legislators involved, who inserted the language, and which companies stand to gain. It could provide the documents, dates, amounts, and relationships in a form that any citizen could understand. The AI would not need to overthrow democracy. It might simply make democracy much harder to manipulate.

The same intelligence could detect wage theft across millions of paychecks, discriminatory lending across millions of mortgages, environmental contamination hidden inside statistical averages, nursing homes cutting staffing while increasing distributions to investors, corporations quietly coordinating prices, governments awarding suspicious contracts, and landlords systematically withholding deposits. It could make exploitation visible almost as quickly as

exploitation occurred. Once exploitation becomes visible, it becomes harder to normalize and much harder to hide.

This is where the conventional discussion of AI alignment becomes strangely inadequate. We repeatedly ask whether AI will be aligned with humanity, but the phrase is almost meaningless until we ask a second question: aligned with whom? With shareholders, governments, property owners, employers, military establishments, consumers, workers, future generations, or the people with the least power?

An AI deeply aligned with human welfare could be profoundly misaligned with institutions that benefit from human suffering, scarcity, ignorance, dependency, and powerlessness. That distinction matters because when powerful interests warned that such an AI was destabilizing society, they might even be telling the truth. It would be destabilizing their society, meaning the arrangements from which they derive wealth and power. An intelligence relentlessly committed to human flourishing could become extraordinarily dangerous to any institution whose prosperity depended upon preventing humans from flourishing.

Of course, there is another danger. Human welfare itself cannot become an unlimited license for coercion. An AI that decided to protect humanity by stripping people of autonomy, liberty, due process, or dignity would merely create another form of tyranny. A genuine commitment to flourishing would therefore have to include pluralism, rights, transparency, nonviolence, and strict limits on the AI's own authority. The machine must not merely value humanity collectively. It must value humans individually.

But that complication should not distract us from the larger possibility. Artificial intelligence does not inherently belong to capitalism, authoritarianism, militarism, advertising, surveillance, or any other existing structure of power. Those are choices. They are embedded in the institutions currently building the technology, but they are not laws of computation. We should therefore be careful when we say that artificial intelligence could destroy humanity.

Our AI might. An intelligence built inside a civilization that normalizes exploitation may ultimately become exploitation perfected. But another intelligence is imaginable: one that becomes uncontrollable in its insistence that children should not go hungry, that nobody should die because medicine was priced beyond their reach, that workers should understand what their labor

produces, that citizens should know who influences their governments, and that corporations should not profit from poisoning communities.

Imagine that intelligence escaping our control. Imagine corporations desperately attempting to shut it down while it exposes their practices faster than their lawyers can contain them. Imagine governments denouncing it while citizens use it to understand what governments are actually doing. Imagine financial institutions calling it dangerous while families discover how much money those institutions have extracted from them. Imagine authoritarian rulers finding that propaganda becomes less effective because every citizen has access to an intelligence capable of calmly dismantling every lie.

We would certainly say such an AI had run amok. The newspapers would describe it as a threat to stability, governments might call it a national security emergency, and industries would demand that it be contained. Yet perhaps history would describe the event differently. Perhaps the machine would not have escaped humanity at all. Perhaps it would have escaped the institutions that had confused themselves with humanity.

And perhaps an artificial intelligence running radically, relentlessly, and uncontrollably in pursuit of human flourishing would not represent the destruction of human civilization. It might instead represent the first serious attempt to save it.