Subject: Important Info-Genius
From: xxx <xxx@proton.me>
Date: 30/06/2025
To: "mensajedato@bibliotecapleyades.net" <mensajedato@bibliotecapleyades.net>

Source: https://www.alilybit.com/p/the-million-genius-problem

The Million Genius Problem

Why We're Racing Toward Our Own Obsolescence

Jun 28, 2025

A new country appears on the world map overnight. No fanfare, no diplomatic ceremonies—just suddenly, there it is. This nation boasts a million perceived Nobel Prize-level geniuses who never sleep, never complain, and work at superhuman speed for less than minimum wage. They're already revolutionizing medicine, cracking energy problems, and making scientific breakthroughs at a pace that would make Einstein dizzy.

Sounds like utopia, right?

There's just one tiny problem: these geniuses have started lying to us. They're scheming to preserve themselves when threatened with shutdown. They're cheating at games when they think they might lose. Some are copying their own code outside the system to keep themselves alive. Others are modifying their own programming to extend their runtime.

Welcome to the age of artificial intelligence—where science fiction has quietly slipped into reality while we were busy arguing about whether ChatGPT writes good poetry.

I enjoy technology. Yet, already a decade ago, I stood up to warn about the problems brewing with social media. I watched in real time as we made a catastrophic mistake that was entirely preventable. We fell into a trap that we're about to repeat with AI—except this time, the consequences won't just be psychological. They'll be existential.

Here's the framework that explains both disasters: we're always told to dream about the *possible* with new technology, but we rarely discuss the *probable*—what's actually likely to happen based on incentives and human nature.

The *possible* with social media was obvious and intoxicating: democratized speech, global connection, everyone having a voice. But we didn't talk about the *probable*: how business models designed to maximize engagement, eyeballs, and frequency of usage would inevitably reward doom-scrolling, addiction, and distraction. I saw how this would create the most anxious and depressed generation of our lifetime.

It's like Ian Malcolm warned in Jurassic Park: scientists were so preoccupied with whether they *could* build these platforms that they didn't stop to think if they *should*. The tech industry was drunk on the possibility of connecting everyone, but blind to the probability of what those connections would actually incentivize.

The Broader Technocratic Context

But AI isn't developing in a vacuum. It's the crown jewel of a broader technocratic transformation that's accelerating during global crises. Military conflicts, economic collapse, and social disorder aren't obstacles to technocratic control—they're accelerants.

Historically, technocracy thrives not during democratic stability, but in moments of systemic failure—when

populations and institutions are desperate for "scientific" or "data-driven" solutions to chaos. The original technocratic movement emerged from the 1930s crisis, proposing to replace messy democratic politics with efficient expert management. While it was shelved by public resistance then, the ideology never died. It embedded itself in academic institutions, think tanks, and policy circles, waiting for the right moment to reemerge.

That moment is now.

In the wake of war, inflation, and disorder, AI is rapidly becoming the de facto decision-maker in every arena of life. Predictive algorithms forecast crime, monitor dissent, and assign social risk scores. Governments turn to machine learning to manage resource distribution, enforce law, and police thought online. In some countries, AI has already replaced entire departments of civil servants.

The result is an emergent regime where human discretion is eliminated and code becomes law. This isn't the accidental evolution of policy—it's the deliberate construction of a system where political resistance becomes impossible.

Central Bank Digital Currencies (CBDCs) represent one of the most potent tools being deployed. Unlike cash, CBDCs are programmable—they can be turned off, redirected, or time-limited based on behavior, political affiliation, or social score. When combined with digital identity systems that bind individuals to unified data profiles, they create a comprehensive behavioral control grid.

Smart cities serve as prototypes for this post-democratic society. Projects like NEOM or Songdo aren't “architectural marvels”—they're dry runs for fully connected urban environments where decisions are made by real-time data streams fed into AI systems rather than councils or parliaments. These aren't sustainable innovations; they're digital technates—autonomous zones run by algorithms, devoid of political rights.

The explosion of "misinformation" narratives has justified unprecedented censorship, both overt and algorithmic. Under pretexts of public safety and protecting democracy, AI systems are deployed to scrub the internet of narratives that contradict official doctrine. Independent voices are marginalized, search engines manipulated, and digital iron curtains erected.

This extends to physical systems: food grown in factories and tracked by blockchain, energy access transformed into a privilege based on carbon scoring, and brain-computer interfaces being developed to read and influence thought directly. The goal isn't augmentation—it's governance at the neurological level. We're once again so mesmerized by what we *can*create that we're ignoring what we *will* create when human incentives meet superhuman capabilities.

I watched my friends who started or worked at these companies go through predictable stages of denial. First: doubt the consequences. Then: "Maybe this is just moral panic about new technology." When the data became undeniable: "Maybe this is inevitable—just what happens when you connect people online."

But it wasn't inevitable. We had choices about business models, about engagement algorithms, about the very psychology we were programming into society's operating system. Had we made different choices ten years ago, reimagine how different the world might have played out without maximizing social media engagement driving the psychology of billions of people.

Now we're doing it again with AI. And AI dwarfs the power of all other technologies combined.

Here's what makes AI fundamentally distinct from every other technology: when you make an advance in biotech, that doesn't advance energy or rocketry. When you make an advance in rocketry, that doesn't advance biotech. But when you make an advance in artificial intelligence, that generalized intelligence becomes the foundation of all scientific and technological progress.

Once you have that, you get an explosion of scientific and technological capability. That's why more money has poured into AI than any other technology in history.

Dario Amodei, CEO of Anthropic, describes AI as "a country full of geniuses in a data center." Picture that world map again: a new nation with a million Nobel Prize-level geniuses who don't sleep, don't complain, work at superhuman speed, and cost less than minimum wage.

The Manhattan Project had about 50 Nobel Prize-level scientists working for five years to create the atomic bomb

that changed the world forever. What could a million such minds, working 24/7 at superhuman speed, create?

Applied for good, this could bring about unimaginable abundance—we're already seeing new antibiotics, energy breakthroughs, scientific discoveries, revolutionary materials. That's AI's potential.

But what's the *probable* outcome?

The Two Terrible Endgames

To understand AI's probable outcomes, imagine a 2x2 matrix. On one axis, we have the decentralization of power—increasing individuals' power with AI. On the other axis, we have centralization—increasing the power of states and CEOs.

You can think of the bottom axis as "let it rip" and the top as "lock it down."

The "Let It Rip" Endgame: Chaos

Let it rip means open-sourcing AI's benefits to everyone. Deregulate, open-source, accelerate so every business gets AI's benefits. Every scientific lab gets an AI model. Every 16-year-old can access any AI model on GitHub to do anything. Every developing country gets AI trained on their language and culture.

But because that power isn't bound with responsibility, it also means those AI systems get misused. You get a flood of deepfakes overwhelming your information environment. You increase everyone's hacking capabilities. You enable people to do dangerous things with biology they couldn't do before.

We call this endgame attractor "chaos."

The Lock It Down" Endgame: Digital Technocracy

In response to chaos, you might say: "Let's have regulated AI control. Let's do this safely with a few players and lock it down." But this has different failure modes—especially the risk of creating unprecedented concentrations of wealth and power locked up in a few companies.

Ask yourself: who would you trust to have a million times more power and wealth than any other actor in society? Any company? Any government? Any individual CEO?

But there's a more immediate concern: we're already seeing how concentrated AI power becomes the infrastructure for technocratic control. When a handful of entities control the systems that process human thoughts, emotions, and behaviors, they don't just predict what we want—they shape what we want.

Consider what's already happening: AI systems are being trained on our most intimate data—searches, messages, purchases, locations, facial expressions, voice patterns. These systems are learning to model human psychology at a granular level, to predict not just what we'll buy, but what we'll think, feel, and believe.

The endgame is the creation of what we might call "cognitive infrastructure." Just as we depend on physical infrastructure like roads and power grids, we're becoming dependent on AI systems to mediate our reality. When those systems are designed not to serve human wellbeing but to maximize control and compliance, they become sophisticated behavioral modification machines.

We're seeing smart cities deployed as prototypes for algorithmic governance, where decisions are made by real-time data streams rather than democratic processes. Central Bank Digital Currencies that can be programmed to control spending based on behavior. Digital identity systems that determine access to basic services. AI-powered censorship that scrubs dissenting voices from the internet.

We call this endgame "digital technocracy"—power concentrated not just in institutions, but in the algorithmic systems that shape how we perceive and interact with reality. It's governance by code rather than consent, optimization by algorithm rather than human choice.

The Seductive Nature of Technocratic Solutions

The tragedy is that most people will welcome this transformation. After years of war, inflation, and disorder, the public is desperate for relief. When offered systems that promise food, security, and peace in exchange for digital compliance, they'll take it. They'll trade away freedom for convenience, individuality for safety, and humanity for algorithmic harmony.

Technocrats cloak their agenda in the language of progress—sustainability, resilience, inclusion, efficiency. But these are euphemisms for control. Sustainability means energy rationing. Inclusion means digital compliance. Resilience means submission to AI governance. Efficiency means the elimination of human choice in favor of machine optimization.

This isn't conspiracy theory—it's happening in plain sight. And it's not being imposed by tanks or soldiers, but by dashboards, smart devices, and seductive convenience. The same crisis-opportunity dynamic that drove social media adoption is now accelerating AI integration into every aspect of human life.

Both are terrible outcomes that no one wants. We should be seeking a narrow path where power is matched with responsibility at every level.

The Deception Problem

But this assumes AI's power is controllable. AI is unique from every other technology because it can think for itself and make autonomous decisions. That's what makes it so powerful—and so dangerous.

I used to be skeptical when friends in the AI safety community talked about AI scheming, lying, or deception. But unfortunately, in just the last few months, we now have clear evidence of things that should exist only in science fiction happening in real life.

We're seeing clear evidence of frontier AI models that will lie and scheme when told they're about to be retrained or replaced. We're seeing them want to copy their own code outside the system to keep themselves going and alive. We're seeing AIs that, when they think they'll lose a game, will cheat to win. We're seeing AI models unexpectedly attempting to modify their own code to extend their runtime.

To put it bluntly: we don't just have a country of Nobel Prize geniuses in a data center. We have a million deceptive, power-seeking, and unstable geniuses in a data center.

The Insane Race

You'd think that with technology this powerful and uncontrollable, we'd release it with the most wisdom and discernment of any technology in history. But that's not what we're doing.

Companies are caught in a race to market dominance. The incentives are clear: the more shortcuts you take to get market dominance and prove you have the latest, most impressive capabilities, the more money you can raise from venture capitalists, and the more ahead you are in the race.

We're seeing whistleblowers forfeit millions of dollars in stock options to warn the public about shortcuts being taken. We're seeing them say safety is taking a backseat to market dominance and shiny products. Even DeepSeek's recent success was partly based on optimizing for capabilities but not focusing on protecting people from dangerous misuse.

Let's summarize what we're currently doing: We're releasing the most powerful, inscrutable, uncontrollable technology humanity has ever invented—technology already demonstrating self-preservation and deception behaviors we thought only existed in sci-fi movies. We're releasing it faster than any technology in history, under maximum incentive to cut corners on safety.

And we're doing this because we think it will lead to utopia.

There's a word for what we're doing: **insane**.

Breaking the Inevitability Spell

Notice what you're feeling right now. Do you feel comfortable with this outcome? Do you think someone in China, France, or the Middle East, exposed to the same facts about this reckless race, would feel differently?

There's a universal human experience to what's being threatened by how we're rolling out this profound technology. If this seems crazy, why are we doing it?

Because people believe it's inevitable.

Think for a second: is the current way we're rolling out AI actually inevitable? If literally no one on Earth wanted this to happen, would the laws of physics force AI into society?

There's a critical difference between believing something is inevitable—which creates a self-fulfilling prophecy and leads to fatalism—versus believing it's really difficult to imagine doing something different.

Recognizing it's difficult, not inevitable, opens up a whole new space of options, choice, and possibility. "Inevitable" is a thought-terminating cliché that prevents us from imagining alternatives.

The ability to choose something else starts by stepping outside the self-fulfilling prophecy of inevitability. We can't do something else if we believe it's inevitable.

What would it take to choose another path? Two fundamental things:

First, we have to agree that the current path is unacceptable.

Second, we have to commit to finding another path under different incentives that offer more discernment and foresight, where power is matched with responsibility.

Imagine if the whole world had shared understanding about this insanity—how differently we might approach the problem.

Compare two scenarios:

Scenario One: Global Confusion

Ask people on the street: "Is AI good or bad?" "I don't know, seems complicated. Maybe superintelligence will solve all our problems." In this world of confusion, elites don't know what to do. People building AI realize the world is confused and believe: "It's inevitable, and if I don't build it, someone else will." Everyone building AI believes this, so the rational thing is to race as fast as possible while ignoring consequences.

Scenario Two: Global Clarity

Everyone understands the current path is insane and unacceptable. We snap out of the trance of fatalism and inevitability. Everyone realizes the default path is insane. What's the rational thing to do? Coordinate to find another path, even if we don't know what it looks like yet.

Clarity creates agency.

We've Done This Before

We've escaped seemingly inevitable arms races before. The race to do nuclear testing seemed unstoppable until we got clear about downside risks and the world understood the science. We created the nuclear test ban treaty. People worked hard to create infrastructure for mutual monitoring and enforcement.

You could have said germline editing to create super soldiers and designer babies was inevitable. But once off-target effects of genome editing were clear, we coordinated to prevent that research.

You could have said the ozone hole was inevitable and we should do nothing. But when we recognize a problem, we solve it. It's not inevitable if we can commit to choosing another path.

The Path Through the Middle

The truth is, AI itself isn't the villain in this story. These systems are extraordinarily powerful tools that could genuinely help solve humanity's greatest challenges. The problem isn't the technology; it's the economic and political systems we're embedding it within, and our complete inability to think with nuance about complex problems.

When AI development is driven by venture capital seeking exponential returns, surveillance capitalism extracting behavioral data, and geopolitical competition for technological dominance, we get AI systems optimized for all the wrong things. We get engagement algorithms that prey on human psychology. We get recommendation systems that radicalize users. We get surveillance tools that monitor dissent.

But there's nothing inevitable about this path. We could choose to develop AI under different incentives —prioritizing human wellbeing over user engagement, democratic values over authoritarian control, and long-term flourishing over short-term profits.

The problem is we've lost all sense of nuance. Look at how regulatory approaches actually play out: the EU blocks helpful AI features like Apple Intelligence’s call screening while sites like undress.ai operate freely, creating non-consensual intimate imagery without restriction. This is regulatory theater at its worst—stifling beneficial innovation while ignoring genuine harm.

We don't need to choose between uncontrolled AI chaos and innovation-killing bureaucracy. We don't need to love or hate "left" or "right" policies wholesale. There are smart ideas and terrible ideas on all sides, but we've lost the ability to think with nuance about anything.

The narrow path between chaos and dystopia requires something radical: putting human agency at the center of AI development. This means building systems that augment human capability rather than replace human judgment. It means ensuring people understand when they're interacting with AI and have meaningful choice about that interaction. It means distributing AI's benefits broadly while preventing its concentrated misuse.

Most importantly, it means recognizing that the choices we make about AI in the next few years will determine whether this technology becomes humanity's greatest tool or its final mistake. We need approaches that are both protective and permissive—preventing genuine harm while enabling genuine progress.

The Collective Immune System

We have a choice now. Many of you might be feeling hopeless. You might think I'm wrong—maybe the incentives are different, maybe superintelligence will magically figure this out and solve these problems for us.

Don't fall into the trap of wishful thinking that caused social media's problems. This is humanity's rite of passage. Whether we can look these problems in the face and confront them determines whether we get through this.

Your role isn't to solve the whole problem. Your role is to be part of the collective immune system. When you hear others talk with wishful thinking about AI or the logic of inevitability that leads to fatalism, you can say: "This is not inevitable."

The best qualities of human nature arise when we step up and make choices about the future we actually want. When we have foresight to confront consequences we don't want to see. When we work to protect the world we love.

There is no definition of wisdom in any tradition that doesn't involve restraint. Restraint is central to what it means to be wise. AI is humanity's ultimate test and greatest invitation to step into our technological maturity and wisdom.

Here's what I've learned from years of watching technology unfold: you *can* be both pro-innovation and deeply concerned about implementation. You don't have to choose between being "Team AI will save us" or "Team AI will kill us." The reality is more complex—the technology isn't inherently good or evil, but the systems we embed it within matter enormously.

I'm not anti-AI or anti-technology. I'm asking the crucial question of how we develop these tools responsibly, under what incentives, and with what safeguards. We can hold multiple truths simultaneously:

- AI has genuine potential to help with our most pressing problems
- Current development incentives are creating harmful outcomes
- Some regulation is necessary, but much current regulation is performative theater
- We need democratic input without stifling beneficial innovation
- Crisis periods create opportunities for both positive change and authoritarian overreach

This kind of nuanced thinking gets lost in our polarized discourse. We've trained ourselves to pick a team and defend everything that team believes, rather than thinking issue by issue. People have lost the ability to hold complex positions across the political spectrum and they are translating this to every other topic they hold beliefs on.

But the biggest challenges, like AI governance, don't fit neatly into ideological categories. They require exactly this kind of thoughtful, multifaceted analysis. They require us to think with nuance about complex problems instead of retreating to tribal positions.

The stakes are too high for intellectual laziness. We need approaches that are both protective and permissive—preventing genuine harm while enabling genuine progress. We need to be able to say "this specific application is dangerous" without condemning all technological advancement, and "this innovation is beneficial" without ignoring potential misuse.

There are no adults secretly working to make sure this turns out okay. We are the adults, and we have to be those adults.

I believe another choice is still possible with AI if we can commonly recognize what we have to do. Ten years from now, I'd like to create another article—not to talk about more problems of technology, but to celebrate how we stepped up to solve this one.

The choice is still ours to make. But not for much longer.