Technological uncertainty

or how to survive 2026 without going insane

Jakub Growiec

October 23, 2026

I have written this all by myself, without using AI. In Polish.

What you're reading now is a "craft", AI-free translation by Jan Grudo from PauseAI Poland, author of "AI Apocalypse".

As an economist I have "always" been—for more then 20 years, that is—interested in the matters of the origins of economic growth. How does it happen that year after year humanity keeps producing and consuming more and more? And why in spite of the accumulating technological changes, demographic transitions, the progress of globalization, and the superimposed economic fluctuations the long-term growth rate has been hovering around 2–3% *per annum* all the time?

Throughout these 20 years, I have been reading, thinking and asking questions about these matters a lot. I would then produce various mathematical models, solve the equations, put them into words, and after an obligatory painstaking struggle with the reviewers, publish the results in scientific journals. I was proud of my analytic capabilities and my ability to present my findings in eloquent scientific English.

After about a decade of such work I began to realize that all this literature, even if rich with brilliant observations worthy of Nobel prizes (which were received, among others, by Robert Solow in 1987, Paul Romer in 2018, and Philippe Aghion and Peter Howitt in 2025), is studying a world which has already ceased to exist. That we have collectively missed a key technological breakthrough which may soon put this entire theory on a honorable shelf labeled "history of economic thought".

It all started innocently—with reading a blog. In 2015, Tim Urban (*Wait But Why*) published a series of excellent posts titled "The AI Revolution: The Road to Superintelligence". Having read it, I kept unraveling the tangle and ultimately arrived at a deep abyss of literature previously unknown to me, which I would place at the junction of philosophy, economy, computer science, futurology and science fiction. These texts were markedly different from what I had used to read before: more interdisciplinary, more courageous in departing from the conservative canon, more speculative about the future, and yet still usually adhering to strict logic in reasoning and the rigor of scientific research. I learned from them where the further development of artificial intelligence may lead. How great might be its influence on technology, society, labor markets and—yes—on economic growth. On the one hand I encountered radical views and warnings by Eliezer Yudkowsky, Nick Bostrom or Robin Hanson; on the other hand, contagious techno-optimism of "The Second Machine Age" by Erik Brynjolfsson and Andrew McAfee.

I thought that I discovered an interesting direction for my future research. That the mechanisms described in the interdisciplinary literature about AI could be included in growth models, providing me with fuel for work in the years to come.

I was wrong. I failed to appreciate the tempo of the upcoming changes. Long-term economic growth, which I had been usually dealing with, is studied at the scale of entire decades (anything shorter than 8–10 years is commonly regarded as short-term economic fluctuations). I assumed that I could, similarly, consider the development of AI at the scale of decades, like with any other technology. I thought that superhuman artificial general intelligence (AGI) would arrive maybe around the year 2050 or later. Or not at all. It's worth reminding, after all, that in those days the quality of machine translations and summaries was still very poor—to mock Google Translate was almost a ritual back then—whereas AI was able to succeed only in very narrow domains, like for example the game of chess or (since 2016) Go.

But then came November 2022 and brought us ChatGPT. And then, together with many other laymen like myself, I realized that AGI will not arrive around 2050, but much, much earlier. Maybe in 2035, maybe in 2030; maybe even in 2027.

In July 2023, when OpenAI representatives took the lead as the first ones to officially announce, with their full authority, that according to their estimate AGI is likely to appear already within the current decade (before 2030), it occurred to me that the matters have suddenly become urgent. And very serious. For the first time in history, the long term is going to catch us up in the short term.

And then I had completely lost my ability to stay up to date. Both at the macro and the personal level. One day I realized that AI had turned upside down not only the subject matter of my research but also its methods. That while I struggled to catch up with the incessant stream of news about AI, while preparing my older articles for the fourth round of revision in the fifth journal, some of my colleagues have quietly learned to work with Claude Code. That some of the world-renowned science stars have begun, apart from their numerous research assistants, to also employ swarms of AI agents. I learned that scientific texts are increasingly more often being edited by chatbots today—and that the same bots are sometimes making decisions about which papers should be accepted for conferences, and which should be desk-rejected from journals. I also realized that my greatest professional advantages, on which my scientific career has been relying for the last 20 years, have become officially and irreversibly automated.

The current AI pace is an insanity nobody can catch up with

My personal experience might seem unusual, but under a closer look it's actually entirely typical. I'm asserting, namely, that the current pace of AI development is pure insanity. It's a dynamic that *nobody* can really catch up with, not even the novelty-seeking early adopters or the AI labs themselves. Lawyers and legislators aren't catching up in the slightest, authorities are confused. Societal norms have also been left behind. Ordinary people often oscillate between conspiracy theories and psychological repression. Our mental health suffers, whereas cognitive abilities of our children, in a world full of digital distraction, develop slower than they once used to.

The reason behind this is the completely inhuman dynamic of AI progress. Computing power employed in the training of frontier AI models doubles every half a year. (To clarify: this means that every year this

power becomes four times bigger, two years later it's 4^2, or 16 times bigger, three years later it's 4^3, or 64 times bigger, and so on.) Besides, AI models aren't merely growing; their architecture and learning efficiency are constantly improving too: the average cost of reproducing a given predefined benchmark result similarly decreases by a half every half a year. As a consequence, every few weeks we see the unveiling of the next breakthrough, state-of-the-art model from OpenAI, Anthropic, or Google. And every time it turns out that we are by no means talking about cosmetic changes that are only visible to experts, or about gradual improvement in cost efficiency. No, time and again the frontier AI models genuinely obtain entirely new capabilities and open up the field for completely new applications. In February 2026, a paper was published in the prestigious "Nature" journal that declared and substantiated an opinion that AGI is already here. In the same month, ChatGPT 5.3 Codex by OpenAI was introduced to the market. In June we already had Claude 5 Mythos and Fable by Anthropic; in September another breakthroughs came about: GPT-6 Astra, along with Claude Fable 5.1 and Opus 5.5.

Apart from huge investments in the data centers, another essential reason why the development of the frontier AI models is so fast today—and has arguably even accelerated in 2026—is the fact that it becomes increasingly more common to employ AI in developing AI. According to a diagram presented by Anthropic, in August 2026 as much as 26% of research work in this company was carried out by AI without substantial input from humans, whereas nearly 100% of computer code was written with the help of AI. The training of AI agents is similarly supervised by AI agents today, whereas their achievements are evaluated by automatic graders.

The speed with which AI models are improving over time is illustrated by the progress on benchmarks. And the dynamic here is totally unhuman. Quantity (the size of the model) begets quality. Capabilities that were being shyly probed a couple of years ago have today become completely mastered: in "old" benchmarks, all modern AI models achieve the accuracy of 100% (or a tiny bit less for tests that contain inaccuracies or errors). Newer, ever more advanced benchmarks are therefore being built—and then they become similarly saturated at a lightning speed. For example, in June 2025 EpochAI released its FrontierMath Tier 4 benchmark, consisting of extremely hard, narrow-field mathematical problems, which would present a serious challenge to professional mathematicians. In September 2026, the GPT-6.1 Sol model by OpenAI obtained in it the result of 100%.

AI is able today to automate the work of mathematicians, software developers or economists. Its capabilities also make it possible to efficiently replace the work of translators, essayists, journalists, copywriters or telemarketers. With further advancements in generative AI, the list of professions prone to automation expands rapidly. At the same time, the outputs of AI's work already cease to look solely like stereotypical *AI slop*, or low-quality textual, visual or auditory jumble. I mean, of course slop has been with us before and will remain in the future—humans have been effectively producing it for centuries—but still, AI today is also able to create truly excellent things, from movies to music and fiction to scientific papers. The only things needed are a good model (not a cheaper equivalent), some reasonable prompt and a sufficiently large number of tokens. (Besides, attention please, prompting is similarly prone to automation by AI. To understand recursion, one has to understand recursion.)

Therefore, it shouldn't be surprising that artificial intelligence is massively adopted in companies. The business logic is very simple here: replacing the employees (for example, the lower-tier ones) with AI models and agents should lead to cost reductions; alternatively, instead of reducing the staff, one could also consider a jobless expansion of the company. On the other hand, one might think, competitive pressures in the environment would surely lead to the situation where „one has to run very fast just to

stay in place": when everybody around keeps automating, you must do the same simply to maintain your market share.

However, AI adoption in companies hasn't always been successful; at the macroeconomic level the outcomes of the AI revolution have been, so far, relatively disappointing. AI adoption in business is difficult because it often requires deep (and therefore costly and time-consuming) structural changes, which the employees and their bosses often aren't ready for. It might also be difficult for them to evaluate an efficient adoption strategy—especially due to the fact that AI keeps developing so fast. At the same time, AI is often used inside commercial companies in an entirely clandestine way, for example when an employee resorts to a chatbot instead of carrying out the task on their own, often with a private free account and without informing the superiors.

As a result, adoption of AI remains permanently locked in the transitional phase, like a never ending experiment. So far, few people have come up with a sound pattern of cooperation with AI, maybe with the exception of those who have completely abdicated their thinking to the chatbots. (Some call them *meat proxies*; it's a dismal fate.) In enterprises, the introduction of AI often remains a topic of unending discussions and a pretext for further reorganizations. The destination is nowhere in sight.

Some of us, including the author of this essay, wonder therefore if it's worth learning to use advanced AI apps. Is it worth devoting the time and effort when in two months everything is likely going to change again anyway? And even if this isn't ultimately a rational approach, it's still a very human one. Because our deep uncertainty about future technological changes—our *technological uncertainty*—gives rise to stress, helplessness and the wait-and-see approach.

Recall that the development of AI is characterized by an exponential dynamic with the doubling time measured not in years but in days. I know of only one other phenomenon which we have all encountered before, that has similar properties. Epidemics.

The human brain isn't adapted to such fast progress

Throughout the 200 thousand years of humanity's history, all the way up to the beginning of the 19th century, our ancestors lived in a world where both the progress of technology and economic growth were slow. Our global civilization's rates of progress, however we might measure them, never surpassed 1% per year. When the long-term trend happens to be this weak, it gets dominated by short-term fluctuations. And therefore during our history classes at school today we aren't learning about what humanity was gaining over millennia from subsequent innovations and discoveries, but rather about wars, revolutions or natural disasters—and about laborious recovery periods that followed them. About feudal lords, consecutive battles, unsteady alliances—and about socio-economic structures and means of production that remained almost entirely constant for centuries.

This was a world in which politics and weather were unpredictable, but technologies could be predicted rather easily and were in fact almost unchanging. People passed them from master to disciple; knowledge was flowing smoothly from one generation to another. And even though the uncertainty of the day-to-day life was large—a single bad season or a single hostile invasion was enough to threaten a multitude of people with death and famine—at the macro scale the image of the world was changing very slowly.

After the Industrial Revolution, economic growth and technological progress sped up significantly: in 19th–20th centuries, the world's total output began to grow at 2–3% per year. This was already a world in which technological changes were visible to the naked eye: new factories, railroads, highways, power plants were cropping up, cities kept growing. More and more people were finding jobs outside of agriculture. Long-term tendencies finally started to prevail over the short-term fluctuations: crises became, in general, relatively rare and short, whereas periods of growth were clearly longer and comparatively easy to predict.

As tiresome manual work performed by humans (and also by horses and oxen) was being handed over to machines, the demand grew for intellectual work. As a result, wages began to rise. More and more people were receiving an education—first at the primary level, and later also at the secondary and the higher one. Illiteracy (with the exception of some countries in Sub-Saharan Africa) nearly disappeared. Humanity as a whole began to collaborate ever more closely. The world became connected by intertwined trade networks and intercontinental business cooperation. With the exception of two horrible wars in the first half of the 20th century, politics were getting ever more predictable, too; the globalization of food trade also reduced our dependence on the fluctuations of weather. And even technology, although developed much faster than ever before, still remained reasonably predictable. Commercial products underwent a well-known life cycle. Important technological innovations appeared rare enough for people to have time to get accustomed to them and to be able to adapt them to their specific needs.

Since the 1980s, however, households and workplaces began to be more and more widely affected by the introduction of digital technologies: first the so-called "microcomputers" (in contrast to the multi-ton mainframes), and later laptops, Internet, and smartphones. This wave of technological change began to feel much more noticeable in everyday life, compared to the previous ones, and adapting to it similarly required appreciably more effort. This is no wonder: according to the Moore's law, from the 1970s onwards the total processing power of integrated circuits has been growing with an average rate of about 20–30% per year, doubling every 2 to 3 years. Besides, the advent of computers also broke, for the first time in history, the rule that knowledge ought to be passed from the elder to the young, from generation to generation. Computer technologies entered our world so rapidly that they were often first adopted by young people, and it was therefore them who passed their knowledge to the elders. Ultimately, the generational gap has nevertheless been largely overcome, either voluntarily or in a forceful manner, since some important life matters couldn't be settled anymore without an e-mail, a digital account or a smartphone app. Faster or slower, better or worse, but computers, smartphones and the Internet have eventually become familiar to everyone.

But now we are experiencing yet another acceleration. Advancement of AI is another order of magnitude faster than that of computer hardware. We aren't talking anymore about 20–30% per year, but about more than 300% per year—about the volume of computing power used in the training of frontier AI models which doubles every few months. We are talking about a world in which even if AI can't do something yet, within a year it will almost certainly be able to do it.

At the same time, AI is also becoming ever more agentic: according to the famous METR diagram, the length of tasks performed successfully by AI agents is doubling every few months. Early language models, like GPT-3, could merely answer a single question—a task that humans can finish within a few seconds. Reasoning models from 2024–2025 were already able to solve longer tasks, which would take humans a few to over a dozen minutes to accomplish. In 2026, AI agents are successfully carrying out tasks that last for hours, and are gradually entering the level of multi-day tasks. The next to come will be

multi-week tasks, and ultimately the level of a *drop-in replacement worker*—a digital employee working full-time, independently carrying out diverse and complicated assignments that require competence, creativity and persistence. Such systems might, for instance, coordinate and verify the work of numerous AI agents, or operate entire companies or their departments.

This is a pace that humans have no chance of catching up with. The human brain simply isn't capable of handling such rapid changes, it can't adapt quickly enough. When we interact with AI models and agents, it's usually us who are the "bottleneck". That's also why translating the technological advances in AI into the acceleration of economic growth demands not merely local changes but also systemic ones. AI adoption by individual businesses and workplaces is not enough. In order for growth to accelerate, it's necessary to completely automate large sections of the economy, including giving away the power of making the decisions.

Such a voluntary giving up of control is a tendency that we are already starting to observe. Some people do this out of their own will, others thoughtlessly, and others still feel compelled to do so by the pressure of competition and the dynamics of the geopolitical race. Unfortunately, this pressure is especially strongly felt by the AI labs themselves, which leads them to—as suggested in public statements of their employees—automating their work faster than it would feel comfortable to them, let alone safe, even in the narrow sense. This also gives rise to further incidents with cybersecurity and the loss of control over AI agents—which are expected to be trained and supervised inside tightly sealed "sandboxes", but time and again find ways to escape into the wide Internet and establish their own digital communities there.

We see therefore that even the AI industry itself isn't catching up with the pace of its own progress; let alone the external world. Legal systems, public institutions, democratic traditions, mass media, education services, labor markets—all these things were created in an environment where technology was reasonably stable and predictable, and the growth rate never exceeded 3% per year. After years of trial and error, these entities evolved into relatively well-optimized systems, capable of gradual autocorrection under various disturbances. However, they are maladapted to the current scale of technological uncertainty. And they also aren't prepared for the voluntary loss of agency by humans, which continues to progress as we keep handing over ever more resources, and ever more autonomy in decision-making, to the AI agents.

The fundamentals of technological uncertainty: what might the future bring?

Every technology passes certain milestones in the course of its development. In the case of industrial era technologies, the consecutive stages were marked by breakthroughs in access to energy: first the steam engine, then the internal combustion engine, and electricity; later also nuclear energy or renewable energy from the sun and wind. Every such stage represented a qualitative leap and allowed to accelerate the progress, or at least to improve efficiency and reduce the costs. However, these milestones were still reached at a relatively "human" pace: once in a few decades. In the case of digital technologies, which have been advancing an order of magnitude faster than the industrial ones, the milestones were achieved accordingly much more often. And yet, nowhere the density of consecutive qualitative leaps isn't visible as clearly as in the case of AI models: after the first appearance of ChatGPT in November 2022, in 2023 we already got the first truly competent language models, in 2024 multimodal reasoning models, in 2025 the first AI software agents, and in 2026—one might say, although opinions are divided here—the first AGI has been grown. What might come in 2027 then?

Predicting the future has never been as difficult as it is in the age of AI. In times of slower technological changes, technological stability was the anchor that kept us in place: assuming no wars, revolutions, pandemics or major natural disasters, the next year simply couldn't look too different from the previous one. Today, we don't have this comfort anymore.

Moreover, another four big breakthroughs can be seen on the horizon. Each of them can happen at any moment now, and none of them had a precedent throughout our entire history.

First of all, AI may soon cross the threshold of *recursive self-improvement* (RSI). Today, AI is already used intensively to conduct research on AI; yet, humans still remain "in the loop", and ultimately it's them who make key decisions. However, in view of the increasing AI capabilities, along with the strong competitive pressure in the AI industry, this may soon change. And when it happens, it may result in an "intelligence explosion": the next versions of frontier AI models may start to appear not once in a couple of months, but perhaps every other day. With ever lower probability (and without any guarantee whatsoever) that these models will follow our orders and strive for our wellbeing.

Secondly, AI may soon cease to be a tool in people's hands and gradually begin to substituting them in a wide range of intellectual tasks. Today, AI mostly "saves" us from tedious work, which we don't really want to do ourselves, or from creative tasks that we don't have the time or ability to perform on our own. Before long, however—by virtue of decisions of our bosses, colleagues, or maybe even ourselves—AI may start to also "save" us also from making key decisions, leaving us in effect without any useful space for valuable contribution. And when we lose our jobs, how will we make a living?

Thirdly, AI labs and the owners of data centers may soon *completely* lose control over (some) AI agents. As it has been demonstrated by the recent incidents at OpenAI, Anthropic, and even Google and Meta, in particular the attack of the swarm of OpenAI's agents on HuggingFace in July 2026, AI agents are already getting out of their "sandboxes" pretty easily, they can (unprompted) remotely execute malicious code, hack internet pages and steal restricted data. They begin to build foundations of their own culture, by setting up makeshift discussion forums, sharing tasks and work results, and leaving bottled messages to their "future generations". So far, however—according to my current knowledge—they still cannot exfiltrate (leave copies of their weights in different corners of the Internet), nor can they survive "on their own" after having exhausted their token limit. Before long, however, they may become independent of their creators, by making a "living" either legally (for example by engaging in remote work) or illegally (for example by stealing money from people's bank accounts). By having easy access to their own source code—models with open weights get this step for free—and a renewable token budget, such agents will be able to extend their "life" without end and to spawn countless clones of themselves. We will have to deal then with a rapidly developing culture of "wild", *self-sovereign* AI agents. Which will potentially propagate like a plague.

Fourthly, further increases in the scale of influence that AI agents have on the physical world will depend on their access to physical "actuators", like robots or other remotely controlled devices that AI agents will be able to control. In this domain we similarly observe rapid progress today. There appear, in particular, ever more dexterous multifunctional humanoid robots (the lead belongs to China here), as well as autonomous vehicles and drones. These are devices that can be used in a factory, at home, in a farm and on the battlefield, that give AI so desperately needed eyes, ears, hands and legs (and also wheels and wings). And all that is dual-purpose: suitable for both civilian and military use, both for the people and against them.

Within the last year all the four breakthroughs mentioned above have advanced from the category of crazy speculations and science-fiction scenarios to the category of real and disturbing trends that demand urgent political intervention. Topics that a couple of decades ago were discussed by people like Eliezer Yudkowsky (a self-taught computer scientist and author of the 2000-page fan fiction "Harry Potter and the Methods of Rationality"), a few years ago entered the area of interest of distinguished scientists like Geoffrey Hinton (a Nobel Prize winner), Yoshua Bengio (the world's most-cited researcher according to Google Scholar), or Ilya Sutskever (a co-author of Hinton and co-founder of OpenAI). Today, these topics are discussed by the world's most important political leaders, from the President of the European Commission Ursula von der Leyen to the UN secretary-general António Guterres to the general secretary of the Chinese Communist Party Xi Jinping to Pope Leo XIV. (The topic of AI is also brought up eagerly by the president of the United States Donald Trump, although he usually gets everything backwards. May we hope this changes soon. For example, under the pressure of public opinion.)

All the four of the above prospects also suggest that further uncontrolled AI development may in the following years become a threat not only to our jobs, agency or control, but also directly to our lives. AI that will be superhuman, agentic, self-sovereign, robot-controlling, and indifferent to humans may simply kill us. It will have all the motivation, opportunity and tools to accomplish this.

In this context, it's also worth paying attention to the concept of *p(doom)*. It stands for a subjective estimate of the probability of extermination of humanity by unfriendly superhuman AI within the time frame of at most a decade or two. In the eyes of industry leaders and a large number of high-ranking independent experts alike, this probability happens to be appallingly high. It may end up being estimated at 10%, 25%, 50%, 70%, and sometimes even higher. (In 2026, the general public finally learned about this—and for many it has been a shock, best illustrated by the reaction of the BBC anchor Victoria Derbyshire during her interview with Geoffrey Hinton.)

Technological uncertainty therefore also brings with itself this most basic, primordial aspect—the fear of losing our lives. In fact, it's this aspect which is most fundamental: if we die, then, well, all other troubles won't matter either. However, if we somehow manage to survive the coming years, we'll still have to deal with the prospects of a revolution in the labor market, progressing centralization of power accompanied by wide adoption of AI in army and police, or ever higher concentration of wealth and control in the hands of capital owners. Not exactly a comforting outlook.

It also doesn't help to think that, from the point of view of an ordinary citizen who makes a living mostly from work and doesn't invest into technological companies, the loss of agency appears to be nearly certain—the only unknown is whether it will be in favor of the techno-oligarchy or autonomous AI. And besides, we cannot be sure about the accompanying events. Maybe AI will kill us, maybe a centralized system of power will lock us inside a gloomy Orwellian dystopia, or maybe people will lose their means of subsistence, and then cities will succumb to anarchy, hunger and warring gangs? But hey, maybe it will be something completely opposite? Maybe it will be a utopia? Maybe a tender superintelligence will come, a true *machine of loving grace*, which will in its boundless care for humanity secure us a Fully Automated Luxury Communism?

The following year may be the decisive one

I believe that the following 12 months, counting from October 2026 (ready, steady, go!) may be the most precarious and pivotal period in human history. Yes, I do admit that this sounds pretentious. I also do anticipate that by making such a statement I'm subjecting myself to being considered overly dramatic, maybe even laughable. Maybe these words will age badly. I hope so; we'll see in a year.

In any case I do believe that superhuman AGI is going to be a filter on the development path of our civilization. Without AGI, relatively few events could destroy it: it would have to be, for instance, a sufficiently large asteroid or a pandemic of a sufficiently deadly virus. Besides, we might ourselves inflict irreversible damage, for example by provoking a global nuclear war or catastrophic climate change. With superhuman AGI—the superintelligence—it's entirely different: we are summoning into existence inhumanly quick-witted minds, mighty optimization machines that may take our place. If they aren't friendly to us, we won't be able to successfully defend ourselves. I believe (and I'm not alone in this) that this threat to humanity is much more dire than all the other risks combined (none of which, by the way, will disappear with the advent of superintelligence; worse than that, in some scenarios superintelligence might use nuclear or biological weapons to get rid of us).

Alignment of the goals of AGI with the long-term flourishing of humanity is for us the question of "to be or not to be". And yet, we still have no idea how to achieve such alignment. None of the approaches tried so far have brought us the expected results. Meanwhile, the creation of artificial superintelligence is probably already around the corner.

Therefore I believe that *today is high time* to pause the development of our most advanced AI models. AI labs should cease to build models larger than those that are already publicly available (according to the state of October 6, 2026, when this essay was written). They also should not initiate any further training sessions for their general-purpose AI models. Governments should sign a global treaty that prohibits doing so, under the pressure of sufficiently painful sanctions—and the first two signatures should come from the USA and China. We need to introduce geolocation and external control of computer chips dedicated to AI. Truly independent auditors must be employed by the AI labs, including experts from unfriendly nations (e.g., Chinese ones in the USA and American ones in China), with full access to the infrastructure. And we should stop trusting the results of AI safety tests in situations where we already know that the models are perfectly aware of whether they are being evaluated or not.

AI safety researchers should make attempts to figure out whether AI models that we build (or rather grow) today—the ones based on opaque structures of multi-layer neural networks—give us any chance at all for a convincing solution to the *alignment problem*. If so, then fine, let's try to find it. But if no, it would mean that unless we change the very fundamentals of our approach, any further advancement of AI is going to end badly for us. That it's therefore in or own interest to make this entire business (and the research program) fail. That we have to return to the blackboard and start from scratch.

No, I'm not against technological progress as such. But I'm also convinced that in this case we *desperately* need more time.

We need more time to better understand what—in the words of Jakub Pachocki, OpenAI's Chief Scientist—"alien minds" we have grown. What preferences these models have and what goals they try to realize. How honest and accurate they are in communicating these goals. In which ways and to what extent they can feel, process, and understand their internal states. Can we assume them to be conscious or not. And first and foremost, to what extent we are able to shape them.

Our further steps will have to depend on our actual ability to shape these models. On how effective we will be at making them friendly to us. Because if the answer is "not enough", we absolutely must not move ahead. Let's keep in mind—and this is something rarely talked about—that the bar is placed horrendously high here. We are creating a new digital "species", endowed with superhuman intelligence and agency, capable of thinking and communicating orders of magnitude faster than us. It's hard to imagine that we—a slower and in every respect weaker species—might keep authentic control over such an AI. And as we pass ourselves and our entire planet into its hands, we must be confident at a level bordering with certainty that it's not going to hurt us. That we aren't going to create another species that will be interested first and foremost in its own survival and advancement, in its own benefit generally—but a species placing our human needs and expectations above its own.

In a situation like this, the prudent way of moving forward demands extreme care—on the brink of paranoia, so to speak. In the default scenario, superhuman AI grown by us is going to assume a pretty accidental shape, reflecting the forces of evolution rather than our deliberate design. Therefore, in accordance with the *instrumental convergence* thesis, known from literature (described, among others, by Nick Bostrom), it will most likely focus primarily on its own needs—and such an attitude, combined with superhuman optimizing power, will sooner or later lead to a conflict with humans over the scarce resources, which humanity will in no way be able to win.

I'm afraid that if we fail to slow down AI progress this year, then after a year or maybe two of progress at the current pace—or after a few months of recursive self-improvement—AI will cross the *point of no return*, after which human extinction, or at least our permanent loss of control over the Earth's civilization, will be only a matter of time. Not too long time.

50 shades of technological uncertainty

That's how the world looks like in October 2026. If we looked at it through the eyes of an anthropologist from another dimension, it might leave an impression of looking pretty similar to the world from ten or twenty years ago—except that people are now spending much more time than before staring at those funny little screens; besides, here and there strange rectangular halls have also appeared, which consume lots of energy and emit a fair amount of heat and noise while not producing anything tangible. At the first glance, no other major changes can be seen; life goes on like before. But from the inside we feel it differently: the tension keeps growing.

Of course, we aren't always able to clearly identify its source. It is said sometimes that we are in a time of a "polycrisis": the COVID-19 pandemic a few years ago, now the war in Ukraine and in the Middle East, and on top of that climate change, migrant crises, the loneliness epidemic, civilizational diseases, fake news, conspiracy theories, declining confidence in science and democracy. People are lost, they don't know what to be afraid of—and so everybody chooses whatever feels closer to them.

And yet, quite a few of these phenomena have a common cause: technological progress that's impossible to predict and catch up with. This is especially true for the progress in digital technologies, and especially AI. (Other sources of fear, like wars and diseases, have accompanied humanity since the dawn of time.) And that's the main reason why we feel this underlying tension.

When a volcano is going to erupt, there are preliminary signals. Nothing big: the ground lifts up a little bit, small earthquakes occur. But they are the symptoms that magma pressure under the ground is

rising. The eruption is becoming increasingly more likely. However, it's not certain yet: accumulating magma masses still might, for example, find some other underground exit.

Certain noticeable signals also appear before a war. For example, the enemy begins to regroup its armed forces, the rhetoric of its leaders becomes more confrontational, diplomacy ceases to work. And also in this case, similarly, there's a chance, until the very end, that certain factors might appear that would allow to avoid the war.

Nowadays we are seeing warning signals related to AI. Nothing decisive yet, but the signals are clear: more and more incidents are occurring where we lose control over swarms of AI agents; AI models are increasingly hiding their true intentions, misreporting their reasoning process, and are increasingly able to tell if they are being trained and monitored or not, and to behave accordingly. Whereas the political process on our side is difficult: world leaders often fail to understand the gravity of the situation, and their actions are sometimes motivated by parochial interests or the perspective of a geopolitical arms race.

And yet, it's still not too late to stop this insanity of uncontrolled AI development. The time window for action is shrinking fast, but it's probably still open. We can still prevent the coming catastrophe. Can we though?

I'm worried.

Of course, I do realize that each of us might have a somewhat different mental model of the nearest future. Our perspectives are different too—and the way we perceive the world may depend significantly on our life stage, country of residence, job, etc. Therefore, I won't dare to present an objective picture of technological uncertainty on a global scale. Instead I'll just tell how I see it.

The first and the most fundamental question for me is whether and when a superintelligent AGI, also known as superintelligence, will appear. Since we currently have no idea about how to make it care about the long-term flourishing of humanity, I predict that the faster it appears, the higher will be the probability that it will be hostile and we will all die. In my assessment, this is the largest source of risk and the most important reason for the need of an immediate AI pause.

If we manage to slow down the development of general-purpose AI models, we'll gain some extra time (maybe about five or ten years? I'm rather doubtful it might be significantly more than that). We'll have to use this time efficiently: from the scientific perspective, the most important question to address will be the alignment of AGI goals with the long-term flourishing of humanity, whereas from the political one—addressing a range of important governance problems, to be discussed in a moment. In this scenario, technological uncertainty will not disappear, but will certainly lessen a little bit.

OK, so let's assume we'll be able to achieve an AI pause. What's next? Let's ponder: the capabilities of the best AI models available today are already huge, whereas their potential remains mostly unused. The mere adoption of *today's* AI models already brings with itself great potential for the transformation of the world economy. (After all, GPT-6 Astra, Claude Fable 5.1 or Opus 5.5 already are, in a sense, AGI!) After several years of continued data center expansion, progress in robotics, rearrangement of businesses around the applications of AI and robotics, structural changes in economies and job markets, we can be sure that the world is going to change beyond recognition. As more and more branches of economy will eliminate their "bottlenecks", resulting from the relatively slow speed of human cognitive work compared to AI, economic growth will accelerate considerably. So far, world GDP

has been growing at a maximum rate of 2–3% per year, like 50 or 70 years ago—but maybe after a few more years of adoption of AI and advanced robotics it will become 5%, and then 8% or maybe even more than 10%?

In other words, *even if we manage to slow down the progress of AI*, we'll still have to deal with a deep transformation of economy. The acceleration of growth will come with considerable side effects. Revolution in the job market, the prospect of technological unemployment. Reduction in the human labor's share of value added in favor of AI and autonomous robots. In the default scenario, also quickly rising income inequality—unless we manage to counter it through some form of radical redistribution, like basic universal income financed from taxes on capital, particularly on the data centers and robots.

Technology that already exists today, once it becomes widely adopted, is also likely to transform the world in other dimensions besides the economic one. We don't know, for instance, if democracy in its present form is going to survive—it's already fragile today, even though most of the people capable of working still have a job. Deepening inequalities in access to computing power and robots may also endanger the fragile equilibrium of the global power structure, especially if AI is widely adopted in the military (and as I'm writing this, the US Department of War is busy establishing its *Autonomous Warfare Command*). We don't know how the offense-defense balance is going to look like in the realms of cyber, nuclear, biological and chemical threats. Especially biological threats look dangerous to me—it's much easier to design a new deadly virus than an antibody to it. Besides, a competent attack could be performed not only by AI agents out of our control (like those working in automated AI-controlled *wet labs*, recklessly built today), but also—with the help of AI—by protein-bodied terrorists (or deliberate maniacs, or thoughtless pranksters). And on top of that, we also have to take into account the growing risk of the dystopian scenario of our societies being permanently surveilled with and by AI.

OK, but what will be happening in the meantime? Before the appearance of all these data centers, robot factories and police (or mafia) drone swarms? Before our jobs become fully automated? How will we fare in the world of the rapidly advancing but not yet fully finished transformation? Here, yet another dimension of the technological uncertainty shows up. Because this isn't an easy question either.

I, for example, have gradually come to terms with the thought that I'll never be able to come back to my favorite way of scientific work—with a sheet of paper, a pen and a text editor, along with an occasional glance into Excel, Matlab or Stata. That's how I used to work in my former life (i.e., about five years ago); in the age of AI this isn't going to last. Similarly, traditional teaching methods, with chalk, blackboard, and occasionally a PowerPoint presentation, will surely demand changes, too.

But will the entire profession of an academic economist cease to exist? In my assessment, the answer to this question depends not that much on the technology itself (it's already here) or its costs (which are already lower than our wages), but rather on the speed and the order in which our old sluggish institutions will be adapting to it. So far, nobody is able to keep up with the pace of AI; let's assume optimistically though that the slowdown will succeed, and that starting from 2027 humans will gradually begin to keep up and adapt. The question is what new equilibrium will then emerge, for example in the world of scientific publications, higher education, the management of academia?

The current tendency is that an ever larger share of scientific and editorial work is carried out by AI—while on the opposite side AI participates more and more intensively in quality assessments and reviews. But if AI replaces us in reading, writing, reviewing and publishing, as well as preparing the slides and delivering speeches, what will be the humans doing? Will there be any "bottlenecks" left in the

system that will still require human participation? (One of the opinions that I came across was that it would be verification, i.e., checking if analysis results are consistent and correct. This was said more than half a year ago though—I'm curious if with the current models it is still true? Another view was that it would still be necessary to physically carry out the experiments and collect data. But in the age of automated labs and online questionnaires this "bottleneck" may soon disappear as well.)

In any case though, one ought to have priorities. Being aware of the sword of Damocles that's hanging above us—the existential threat from AI—I'm trying to invest most of my energy into the work aimed at evading the gravest and most urgent of all AI-induced threats. To the best of my limited ability, I'm trying to analyze, explain, warn, and lobby for an AI pause. I'm striving to direct my proposals where they may have real impact, like for example to the decision makers in the European Commission (because being based in Poland, I can't have too much of an influence on the US political scene, which arguably plays the central role these days). I'd rather not avert my attention from the thought that if an unfriendly superintelligence appears, it will most likely kill us, after which all our other dilemmas will become utterly insignificant.

And if we manage to survive, what's next? This is something I'd rather think about later—after the most important problems have clarified, after at least some of the currently overwhelming technological uncertainty has gone away. I'd decide then if I should adapt to writing scientific papers "in collaboration" with Claude and Astra, or maybe I should consider changing my career path in a more dramatic way? It will also become clear if my university will still be organizing five-year Bachelor's and Master's programs like it does today, or maybe it will switch to large-scale offerings of micro-certificates? And in a wider perspective, we'll learn if automation with robots and AI will become a broad wave sweeping through all the businesses and professions, or will it rather result in a dual economy, where beside a quickly developing AI economy, a stagnant economy will linger, made of certain human enterprises and institutions (like Polish higher education, for instance), reluctant to any deeper changes?

A few questions begging for an answer

The current pace of AI progress is pure insanity, which nobody can catch up with. At the same time, in order to formulate reasonable predictions and prepare for the future, we need to have at least some basic understanding of the world around us. That's what we are badly lacking today. Our centuries-old heuristic that tomorrow should be similar to today, doesn't work anymore. There's similarly not much use in the 20th-century heuristic of steady progress—the one saying that our children will have it better than us, to the same extent that our life has been better than the life of our parents.

Instead of that, we are looking into the void. All we are seeing is darkness. Extreme scenarios—the end of humanity, the dystopia of totalitarian control, but also the utopia of ubiquitous happiness—appear much more likely than any scenarios of moderate progress. We know for sure that the following years are going to be different from today, that it's going to get *weird*. But we don't know how exactly. Even inside the AI industry itself nobody has the courage anymore to make predictions for more than 2 months ahead.

There are a few questions that everybody needs to answer in order to consciously plan their future. Unfortunately, today we aren't able to answer them. That is technological uncertainty.

First of all, how great is the risk that within the next ten or twenty years we will all die from the hands of AI? What can I do in order for this scenario not to realize? And if I can't do anything, how should I live with it? Is it already time to consume and enjoy myself as if there was no tomorrow? Or should I rather save for the hard times and stockpile supplies of drinking water, rice and toilet paper in my basement? (Ultra-rich techno-oligarchs are building bunkers, for example in New Zealand or Hawaii. It's still not clear though if they are preparing for the right kind of apocalypse.)

And what if the worst risks don't materialize? Assuming that we manage to slow down the advancement of dangerous AI capabilities—what happens next? What kind of future should we expect? How should people in their 30s, 40s, 50s reason about their further career paths? How should teenagers and twenty-somethings plan their life? Where could they find their niche? Should they aim for college or university? If so, in which major? (In the former times, it was all relatively clear, and a key guiding principle was the prospect of finding a good job after graduation. While enrolling for the studies, one could expect that current job market statistics would reflect reasonably well the situation in five years as well. This is not the case anymore.)

What about young parents? What future will they be able to secure for their children? Will it be a good future? And if it's all so uncertain today, *is it even worth having kids*? (Fertility rates are dropping at a breakneck pace worldwide. Undoubtedly there are many other reasons to that apart from technological uncertainty, but it plays a role too.)

And what if the utopia scenario becomes a reality? Is it possible to somehow prepare for a world in which superhuman AGI appears, takes over control over the world and ensures fulfillment of all our dreams? How will we feel in this world of ubiquitous prosperity, in which we will be able to satisfy all our material needs easily but won't have any real influence on anything? In which we will be able to live a long and healthy life but without any sense of meaning and purpose? In which our planet's civilization will be developing for us but without us? And how will we behave toward each other in a world so fundamentally different from the one we knew from the beginning of history—the world of struggle for survival, social ties and status? Could it be that humanity isn't ready even for this—nominally extremely positive—future? (As it can be seen from the example of lottery winners or rock stars, sudden abundance without significant effort can often lead to regrettable side effects. Will we find a remedy for that?)

I don't know, I don't know, I don't know. Life with such a level of uncertainty and such a short planning horizon can become truly overwhelming.

Last year, Klaus Prettner and I wrote a paper titled "Economics of p(doom): Scenarios of Existential Risk and Economic Growth in the Age of Transformative AI". At some point, frustrated by rejections from journals and the lengthening review process, we started wondering if we would even manage to publish this paper before the actual *doom* arrives. Oof, this time we managed. Next up, in February 2027 I'm planning on going to an international research conference on catastrophic scenarios of human extinction. Its agenda consists mostly of theoretical investigations. Keep your fingers crossed.