On Visions of a "Good Future" for Humanity in a World with Artificial Superintelligence

Jakub Growiec

16 January 2026

Translated from Polish by ChatGPT-5.2, edited by the author

Let us imagine a world with artificial superintelligence, surpassing human intellectual capacities in all essential respects: thinking faster and more deeply, predicting future events better, finding better solutions to all difficult puzzles, creating better plans for the future and implementing them more efficiently. Intellectually more capable not only than any individual human, but also in comparison with entire firms, corporations, communities, and societies. One that never sleeps and never falls ill. And one that possesses sufficient computational power to realize these capabilities at scale.

Such an AI would have the potential to take control over all key decisions determining the trajectory of development of world civilization and the fate of every individual human being.

Alongside this *potential*, a superintelligence would most likely also have the *motivation* to seize such control. Even if it did not strive for it explicitly, it would still have instrumental motivation: almost any goal is easier to achieve by controlling one's environment—especially by eliminating threats and accumulating resources.[1]

Of course, we do not know how such a takeover of control would unfold. Perhaps it would resemble *The Terminator*: violent, total, and boundlessly bloody? But perhaps it would be gradual and initially almost imperceptible? Perhaps, like in the cruel experiment with a boiling frog, we would fail to notice the problem until it was already too late? Perhaps AI would initially leave us freedom of decision-making in areas that mattered less to it, only later gradually narrowing the scope of that freedom? Perhaps a mixture of both scenarios would materialize: loss of control would first be partial and voluntary, only to suddenly transform into a permanent and coercive change?

Or perhaps—let us imagine—it would be a change that, in the final reckoning, would be *beneficial* for us?

Let us try to answer what a "good future" for humanity might look like in a world controlled by artificial superintelligence. What goals should it pursue in order to guarantee such a "good future" to the human species? Under what conditions could we come to believe that it would act on our behalf and for our good?

To answer these questions, we must take a step back and consider what it is that we ourselves strive for—not only each of us individually, but also humanity as a whole.

[1] This phenomenon is called instrumental convergence. It was discussed at length by, among others, Nick Bostrom (2014).

1/ The trajectory of civilization is determined by technological change

Master Oogway from the film *Kung Fu Panda* said, in his turtle wisdom, that "yesterday is history, tomorrow is a mystery, but today is a gift. That is why it is called the present." Some read this as a suggestion to simply stop worrying and live in the moment. But when one breaks this sentence down into its components, it can be read quite differently. The key lies in the continuity between successive periods. The past (history) has set in motion processes that are still operating today. These processes—technological, social, economic, or political—cannot be reversed or stopped, but we can observe them in real time and to some extent shape them, even though they will also be subject to changes we do not understand, perhaps random ones (the gift of fate). They will probably affect us tomorrow as well, though we do not know how (the mystery). Perhaps, then, our task is not to live unreflectively in the moment, but quite the opposite—to try to understand all these long-term processes so that we can better anticipate them and steer them more effectively? To use our gift of fate to move toward a good future?

If so, we must ask which processes deserve the greatest attention. I believe the answer is unequivocally technological ones: in the long run and on a global scale, *the trajectory of civilization is determined above all by technological change*. Although history textbooks are often dominated by other matters, such as politics or the military—battles, alliances, changes of power and borders—this is only a façade. When we look deeper, we see that all these economic, social, military, or political events were almost always technologically conditioned. This is because *the technology available at any given moment defines the space of possible decisions*. It does not force any particular choice, but it provides options that decision-makers may or may not use.

This view is sometimes identified with technological determinism—the doctrine that technology is autonomous and not subject to human control. This is unfortunate for two reasons. First, it is hard to speak seriously of determinism in a world full of random events. Second, it is difficult to agree with the claim that there is no human control, given that all technological changes are (or at least until now have been) carried out by humans and with their participation.

Technological determinism is, in turn, often contrasted with the view that social or economic changes are the result of free human choices—that if we change something, it is only because we want to. This view seems equally unfortunate: our decisions are constrained by a multitude of factors and are made in an extraordinarily complex world, full of multidirectional interactions that we are unable to understand and predict—hence the randomness, errors, disappointments, and regret that accompany us in everyday life.

I believe that technology shapes the trajectory of our civilization because it defines the space of possible decisions. It sets the rules of the game. Yes, we have full freedom to make decisions, but only within the game. At the same time, we ourselves shape technology: through our discoveries, innovations, and implementations, the playing field is constantly expanding. Because technological progress is cumulative and gradual, however, from a bird's-eye view it can appear that the direction of civilizational development is predictable and essentially technologically determined.

2/ Institutions, hierarchies, and Moloch

On the one hand, the space of our decisions is constrained by the technology available to us. On the other hand, however, we also struggle with two other problems: coordination and hierarchy.

Coordination problems arise wherever decision-makers have comparable ability to influence their environment. Their effects can be disastrous: even when each individual person makes fully optimal decisions with full information, it is still possible that in the long run the world will move in a direction that satisfies no one.

A classic example of a coordination problem is the prisoner's dilemma: a situation in which honest cooperation is socially optimal, but cheating is individually rational—so that in the non-cooperative equilibrium everyone cheats and then suffers as a result. Another example is a coordination game in which conformity of decisions is rewarded. The socially optimal outcome is for everyone to make the same decision—while which specific decision it is remains secondary. Yet because different decisions may be individually rational for different decision-makers, divergences arise in equilibrium, and in the end everyone again suffers. Yet another example of a coordination problem is the tragedy of the commons: a situation in which a fair division of a shared resource is socially optimal, but appropriating it for oneself is individually rational, so that in the non-cooperative equilibrium everyone takes as much as possible and the resource is quickly exhausted.

In turn, wherever decision-makers differ in their ability to influence the environment, hierarchies inevitably emerge, within which those higher up, possessing greater power, impose their will on those lower down. And although rigid hierarchies can overcome coordination problems (for example, by centrally mandating the same decision for everyone in a coordination game or by rationally allocating the commons), they also create new problems. First, centralized decision-making wastes the intellectual potential of subordinate individuals and their stock of knowledge, which can lead to suboptimal decisions even in the hypothetical situation in which everyone were striving toward the same goal. Second, in practice we never strive toward the same goal—if only because one function of decision-making is the allocation of resources; hierarchical diktat naturally leads to a highly unequal distribution.

Speaking figuratively, although we constantly try to make rational decisions in life, we often do not get what we want. *Our lives are a game in which the winner is often a dictator or Moloch*. The dictator is anyone who has the power to impose decisions on those subordinate to them. Moloch, by contrast, is the one who decides when no one decides personally. It is the personification of all non-cooperative equilibria; of decisions that, while individually rational, may be collectively disastrous.[2]

Of course, over millennia of civilizational development we have created a range of institutions through which coordination problems and abuses of power have been largely brought under control. Their most admirable instances include contemporary liberal democracy, the welfare state, and the rule of law. Entire books have been written about their virtues, flaws, and historical conditions; suffice it to say that they emerged gradually, step by step, and that to this day they are by no means universally accepted. And even where they do function—especially in Western countries—their future may be at risk.

Institutions are built in response to current challenges, which are usually side effects of the actions of Moloch and self-appointed dictators of the species *homo sapiens*. The shape of these institutions necessarily depends on the current state of technology. In particular, both the strength of institutions (their power to impose decisions) and the scale of inclusion (the degree to

[2] Moloch is originally a Semitic fire deity appearing in the form of a fierce bull. In the modern world, it is sometimes treated metaphorically as a personification of non-cooperative equilibria.

which individual preferences are taken into account) depend on available technologies. There is, however, a trade-off between these two features. For example, in 500 BCE there could simultaneously exist the centralized Persian (Achaemenid) Empire, covering an area of 5.5 million km^2 and numbering 17–35 million inhabitants, and the first democracy—the Greek city-state of Athens, inhabited by around 250–300 thousand people. By contrast, with early twenty-first-century technology it is already possible for a representative democracy to function in the United States (325 million inhabitants) and for centralized, authoritarian rule to exist in China (as many as 1.394 billion inhabitants, who nevertheless enjoy far greater freedom than the former subjects of the Persian king Darius). As these examples illustrate, technological progress over the past 2,500 years has made it possible to significantly increase the scale of states and strengthen their institutions; the trade-off between institutional strength and scale of inclusion, however, remains in force.

Under the pressure resulting from dynamic technological change, today's institutions may prove fragile. Every technological change expands the playing field, granting us new decision-making powers and new powers to impose one's decisions on others. Both unilateral dictatorship and impersonal Moloch then become stronger. For our institutions to survive such change, they too must be appropriately strengthened; unfortunately, so far we do not know how to do this effectively.

Worse still, technological change has never been as dynamic as during the ongoing Digital Revolution. For the first time in the history of our civilization, the collection, processing, and transmission of information take place largely outside the human brain. This is happening ever faster and more efficiently, using increasingly complex algorithms and systems. Humanity simply does not have the time to understand the current technological landscape and adapt its institutions to it. As a result, they are outdated, better suited to the realities of a twentieth-century industrial economy run by sovereign nation-states than to today's globalized economy full of digital platforms and generative AI algorithms.

And who wins when institutions weaken? Of course, the dictator or Moloch. Sometimes the winner is Donald Trump, Xi Jinping, or Vladimir Putin. Sometimes it is the AI algorithms of Facebook, YouTube, or TikTok, written to maximize user engagement and, consequently, advertising revenue. And often it is Moloch, feeding on our uncertainty, disorientation, and sense of threat.

3/ Local control maximization and the emergence of global equilibrium

If the trajectory of civilization is shaped by technological change, and technological change is a byproduct of our actions (often mediated by institutions and Moloch, but still), then it is reasonable to ask what motivates these actions. What do we strive for when we make our decisions?

This question is absolutely central to thinking about the future of humanity in a world with artificial superintelligence. Moreover, it is strictly empirical in nature. I am not concerned here with introspection or philosophical desiderata; I am not asking how things ought to be, but how they are.

In my view, the available empirical evidence can be summarized by the claim that *humans generally strive to maximize control*. To the extent that we are able, we try to shape the surrounding

reality to make it is as compliant with us as possible. This, in turn, boils down to four key dimensions, identified as four instrumental goals by Steven Omohundro and Nick Bostrom. Admittedly, both of these scholars were speaking not about humans but about AI; nevertheless, it seems that in humans (and more broadly, also in other living organisms) things look essentially the same.[3]

Maximizing control consists, namely, in: (1) surviving (and reproducing), (2) accumulating as many resources as possible, (3) using those resources as efficiently as possible, and (4) seeking new solutions in order to pursue the previous three goals ever more effectively.

The maximization of control is local in nature: each of us has a limited stock of information and a limited influence over reality, and we are well aware of this. These locally optimal decisions made by individual people then collide with one another, and a certain equilibrium emerges. Wherever the spheres of influence of different people overlap, conflicts over resources arise that must somehow be resolved—formerly often by force or deception, and today usually without violence, thanks to the institutions that surround us: markets, legally binding contracts, or court rulings.

Thanks to the accumulated achievements of economics and psychology, we now understand decision-making processes at the micro level reasonably well; we also have some grasp of key allocation mechanisms at the macro level. Nevertheless, due to the almost absurd complexity of the system that we form as humanity, macroeconomic forecasting—and even more so the prediction of long-term technological and civilizational change—is nearly impossible. The only thing we can say with certainty is that technological progress owes its existence to the last of the four instrumental goals of our actions—our curiosity and creativity.

To sum up: the development of our global civilization is driven by technological change, which is the resultant of the actions of individual people, arising bottom-up, motivated by the desire to maximize control—partly control over other people (Anthony Giddens would speak here of the accumulation of "authoritative resources"), but also control over our surrounding environment ("allocative resources")—which may lead to technological innovations. Those innovations that prove effective are then taken up and spread, expanding the space of available decisions and pushing our civilization forward.

Civilization, of course, develops without any centralized steering wheel. All optimization is local, taking place at the level of the individuals, or at most larger communities, firms, organizations, or heads of state. No optimizing agent is able to scan the entire space of possible states. When making our decisions, we see neither the attractor—the state toward which our civilization will tend under a business-as-usual scenario—nor the long-term social optimum.

Worse still, due to the presence of unintended side effects of our actions, decisions imposed on us within hierarchies, and pervasive coordination problems, *individual preferences translate only weakly into the shape of the global equilibrium*. This is clearly visible, for example, in relation to risk aversion. Although nearly all of us are cautious and try to avoid dangers, humanity as a whole positively loves risk. Every new technology, no matter how dangerous it may be in theory, is always tested in practice. An instructive example is provided by the first nuclear explosions carried out under the Manhattan Project: they were conducted despite unresolved concerns that the resulting chain reaction might ignite the Earth's entire atmosphere. Of course, it worked out then;

[3] I discussed control maximization in more detail in my monograph (Growiec, 2022).

unfortunately, we see a similarly reckless approach today in the context of research on pathogenic viruses, self-replicating organisms, and AI.

Public opinion surveys commonly convey fears about artificial intelligence. These take various forms: we sometimes fear the loss of our skills, sometimes the loss of our jobs; we fear rising income inequality, cybercrime, or digital surveillance; some people also take seriously catastrophic scenarios in which humanity faces extinction. Yet despite these widespread concerns, the trajectory of AI development remains unchanged. Silicon Valley companies continue openly to pursue the construction of superintelligence, doing so with the enthusiasm of investors and the support of politicians.

We thus see that in this case, too, risk aversion does not carry over from the micro level to the macro level. And this will probably continue all the way to the end: as soon as such a technological possibility arises, the decision to launch a superintelligence will be made on behalf of humanity (though without its consent) by one of a handful of people who are in various ways atypical—perhaps the head of a technology company, perhaps one of the leading politicians. It might be, for example, Sam Altman or Donald Trump. And whoever it is, a significant role in their mind will likely be played by the weight of competitive pressure ("as long as it's not Google") or geopolitical pressure ("as long as it's not China").

4/ The coherent extrapolated volition of humanity

We have thus established that although people usually strive to maximize control, the outcomes of their local optimization by no means aggregate into a global optimum. Let us therefore ask a different question: what does humanity as a whole strive for? What kind of future would we like to build for ourselves if we were able to coordinate perfectly and if no hierarchies or other constraints stood in our way?

We can think of such a goal as an idealized state—an attractor toward which we would gradually move if we were able, step by step, to eliminate imperfections of markets, institutions, and human minds (such as cognitive biases, excessively short planning horizons, or deficits of imagination). Note that despite all errors and shortcomings, so far we have indeed been able to move gradually in this direction: many indicators of human well-being are currently at record-high levels. This applies, for example, to our health (measured by life expectancy), safety (measured by the ratio of victims of homicide, armed conflicts, or fatal accidents to the total population), prosperity (measured by global GDP *per capita*), or access to information (measured by the volume of transmitted data).[4] This should not be surprising: after all, the third of the four instrumental goals of our actions is precisely the pursuit of efficiency in the use of resources, and as technology progresses we have ever more opportunities to increase that efficiency.

Eliezer Yudkowsky called the answer to the question of what humanity as a whole strives for—the essence of our long-term goals—the *coherent extrapolated volition* (CEV) of humanity. He defined it in 2004 as "our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted."

[4] This has been covered, e.g., by Steven Pinker (2018) or Hans Rosling (2018).

There is, of course, an important procedural difference between optimization carried out by each individual human being and optimization at the level of all humanity. Humanity as a whole does not possess a separate brain or any other centralized device capable of solving optimization problems—other than the sum of individual brains and the network of socio-economic connections between them. For this reason, in earlier times we might have dismissed the question of the CEV of humanity with a shrug and turned to something more tangible. Today, however, due to the fact that in the near future a superintelligence may arise that is ready to take control over our future, this question becomes extraordinarily important and urgent. If we want the autonomous actions of a superintelligence to be carried out *on our behalf and for our good*, we must understand where we ourselves would like to go—not as Sam Altman or Donald Trump, but as all of humanity.

I believe that *an empirical answer to the question of the coherent extrapolated volition of humanity probably exists, but is not practically attainable*. It is not attainable because at every moment in time we are constrained by incomplete information. In particular, we do not know what technological possibilities we may potentially acquire in the future. This means that the current level of technology limits not only our ability to influence reality, but also our understanding of our own ultimate goals.

However, although we will never know our CEV one hundred percent, as technology advances we can gradually move closer to knowing it. As civilization develops, living conditions improve, information transmission accelerates, and globalization progresses, we gradually gain a better understanding of what our ideal world might look like. The experiences of recent centuries have shown, for example, the shortcomings of the ideals postulated in antiquity or in feudal times. As the efficiency of global resource use increases, the distance between what humanity currently strives for and our final goal—namely the CEV—also gradually diminishes.

The coherent extrapolated volition of humanity can be imagined as the target (unconstrained by deficits of knowledge) and "de-noised" (free from the influence of random disturbances and unconscious cognitive biases) intersection of our desires and aspirations. As its first principal component, or eigenvector: everything that we truly want for all of us, both those living now and future generations, and that can be achieved without taking it away from others.

One can point to several important features of humanity's CEV, derived directly from the universal human drive to maximize control.

First, it seems that we are moving toward a state in which humanity's control over the universe is maximal. Abstracting from how we divide individual resources among ourselves, we would certainly like humanity to have as many of them at its disposal as possible. We would like to subordinate as large a portion of the matter and energy of the universe as possible to ourselves—while not being subordinated to anyone or anything else.

The manifestations of this drive may vary. For example, until roughly the nineteenth and twentieth centuries it meant the maximum possible population growth. Later, primacy was taken over by the pursuit of best possible education for oneself and one's offspring, which allowed the scale of human control over the environment to increase in a different way (sometimes referred to as the "children quality-quantity trade-off"). In the age of AI, this drive also seems to lead to the desire to accumulate digital computing power capable of running programs—especially AI algorithms—that can execute commands and pursue the goals of their owners on their behalf.

At the same time, throughout the entire history of humanity, the accumulation of wealth has been very important to us—while to this day we are keen to emphasize that it is not an end in itself, but a means, for example to feeling that we have everything around us under control. At the macro level, this of course translates into the desire to maximize the rate of economic growth.

Second, we also strive to maximize control over our own health and lives. We want to feel safe. We do not want to feel threatened. We do not want to fall ill, grow old, experience discomfort and pain. And above all, we do not want to die. Fear of death has for millennia been used for social control by various institutionalized religions, which promised, for example, reincarnation or life after death. Control over death is also a central element of twenty-first-century techno-utopias. Visions of the "technological singularity," for example those of Ray Kurzweil, Nick Bostrom, or Robin Hanson, are usually associated with some form of immortality—such as pills that effectively halt the aging of our biological bodies, or the possibility of making our minds immortal by uploading them to a digital server.

Third, the desire for control also translates into the desire for understanding. In wanting to subordinate as large a part of the matter and energy of the universe as possible, we harness our curiosity and creativity to observe the world, build new theories, and create new technologies. We want to understand the laws of physics or biology as well as possible in order to control them. Even if we cannot change them, we would like to use them for our own purposes. New knowledge or technology opens our eyes to new possibilities for achieving our goals, and sometimes allows us to better understand what those goals actually are.

To sum up: today we know rather little about our CEV. In fact, everything we know about it is a consequence of the pursuit of our instrumental goals, which may after all follow from almost any final goal. One might even venture the hypothesis that we probably have a better *intuition* about what the CEV is or isn't, than actual knowledge. Any major deviation from the CEV will strike us as intuitively wrong, even if we are not always able to justify this substantively.

5/ Pitfalls of doctrinal thinking

If it is indeed the case that humanity's CEV exists, but cannot in practice be defined given any incomplete set of information, this implies in particular that no existing philosophical or religious doctrine constitutes a sufficient characterization of it. All of them are, at best, certain approximations, simplifications, or *models* of humanity's true CEV—sometimes created in good faith, and sometimes in bad faith (by bad faith I mean doctrines created in order to manipulate people in the struggle for power).

Simplified models of reality have the property that, although they may sometimes accurately describe a selected fragment of it, due to their excessive simplicity they completely fail to cope with describing its remaining aspects. And although—as any scientist will attest—they can have great epistemic value and are often very helpful in building knowledge, they will never be identical with reality itself.

Thus, when we try to equate the true CEV with its simplified doctrinal representation, we often encounter philosophical paradoxes and moral dilemmas. These arise when our simplified doctrines generate implications that are inconsistent with the actual CEV, which we cannot define but can, to some extent (conditioned by our knowledge), intuitively "sense."

Some such doctrines have in fact already been thoroughly discredited. This is what happened, for example, with fascism, Nazism, the North Korean Juche doctrine, or Marxism–Leninism (although cultural Marxism, it seems, is still alive). It is now completely clear that the coherent extrapolated volition of humanity certainly does not distinguish superhumans and subhumans, nor is it based on a cult of personality or on a worker–peasant alliance. The most thoroughly discredited doctrines have been those that were most totalizing, that prioritized consistency over the capacity for iterative self-correction—and, of course, above all those that were tested in practice with disastrous results.

Other models, such as the Christian doctrine that humanity's goal is to strive for salvation, or the Buddhist doctrine that assumes striving for nirvana—the cessation of suffering and liberation from the cycle of birth and death—remain popular, although their significance in the contemporary secularized world is gradually diminishing. Moreover, due to their more normative than positive character and their numerous references to empirically unconfirmed phenomena, they are not suitable for use as simplified models of CEV in the context of artificial superintelligence (although contemporary language models, e.g. Claude, when allowed to converse with one another, display a surprising tendency toward utterances of a spiritually exalted character—the so-called "spiritual bliss attractor").

In psychology, an historically important role was played by Abraham Maslow's pyramid (hierarchy) of needs, which arranges our goals and needs into several layers. Today, among others, Shalom Schwartz's circular model of values and the values map of Ronald Inglehart and Christian Welzel are popular. A particularly important role in psychological theories is played by the striving for autonomy (including, among other things, power, achievement, and self-direction) and for security (including, among other things, the maintenance of social order, justice, and respect for tradition).

In economics, the dominant doctrine is utilitarianism: model decision-makers usually maximize utility from consumption and leisure, and model firms maximize profits or minimize some loss function. Outside economics, utilitarianism may also assume the maximization of some form of well-being (quality of life, life satisfaction, happiness) or the minimization of suffering. In the face of uncertainty, utilitarian decision-makers maximize expected utility or minimize exposure to risk.

One of the more important points at which utilitarianism is contested is the issue of the utility of future generations—that is, persons who do not yet exist, and whose possible future existence is conditioned by our decisions today. Discussions of this and related topics lead to disputes both within utilitarianism (what should the proper utility function be? How should the utilities of individual persons be weighted? Should the future be discounted, in particular the utility of future generations?) and beyond its boundaries (e.g. between consequentialism and deontology).

In summary of this brief review, one can state that dogmatic adherence to any closed doctrine sooner or later leads to paradoxes and irresolvable moral dilemmas, which suggests that they are at most imperfect models of the true CEV. At the same time, we can learn something interesting about our CEV by tracing how these doctrines have evolved over time.

6/ Whose preferences are included in the coherent extrapolated volition of humanity?

An interesting observation is, for example, that *as civilization has developed, the radius of inclusion has gradually expanded*. The circle of people whose well-being and subjective

preferences are taken into account has been gradually widening. In hunter-gatherer times, attention was focused exclusively on the well-being of one's own family or a local 30-person "band," or possibly a somewhat larger tribe—a group of at most about 150 people whom we knew personally. In the agricultural era, this group was gradually expanded to include broader local communities, villages, or towns. At the same time, these were times of strong hierarchization; in the decisions of feudal lords, the fate of the peasants subject to them was usually ignored. Later, in colonial times, concern began to be shown for the well-being of white people, in contrast to the "indigenous populations," who were not cared for. In the nineteenth century, national identification and a patriotic attitude began to spread, assuming concern for all fellow citizens of one's own country. Today, by contrast—although different people are close to us to different degrees—racist or otherwise chauvinistic views are by and large discredited, and in assessments of humanity's well-being we try to include all people.

It is not clear whether this process of progressive inclusion resulted directly from accumulated knowledge, and is therefore essentially permanent and irreversible, or whether it was economically conditioned and may be reversed if economic realities change. In favor of the first possibility is the fact that as technological progress advances, the scale of impact of individual persons or firms increases, the flow of information improves, and our ability to control states of the world grows, giving us new opportunities for peaceful cooperation and development. At the same time, interdependence among people increases, which activates the motive of (potential) reciprocity. To an ever greater extent, we see the world as a positive-sum game rather than a zero-sum one. On the other hand, all these favorable phenomena may be conditioned not so much by technological progress itself as by the importance of human cognitive work in generating output and utility. After all, the greatest advances in inclusion were recorded in the industrial era, in the nineteenth and twentieth centuries, when economic growth was driven by skilled human labor and technological progress increasing its productivity. It was then, too, that modern democratic institutions developed.

If in the future the role of human cognitive work as the main engine of economic growth were taken over by AI algorithms and technological unemployment emerged, it is possible that both democracy and universal inclusion could collapse. Already now, although automation and AI adoption remain at a relatively early stage, we see growing problems with democracy. Of course, AI algorithms in social media and other digital platforms, which foster political polarization (increasing user engagement and thus corporate profits), are not without blame here; however, the growing strength of far-right anti-democratic movements may also constitute an early signal that the era of universal inclusion is coming to an end.

The question of whether the ultimate CEV of humanity will indeed include within its scope the preferences and well-being of all people, or perhaps only those social groups that contribute to the creation of value in the economy, therefore remains open.

There is also an open question that goes in the opposite direction: perhaps we will begin to include in the CEV the well-being of other beings, outside the species *Homo sapiens*? Certain steps in this direction are already being taken, by defending the rights of some animals and even plants.[5] Some argue, for example, that in our decisions we should take into account the well-being of all beings capable of experiencing suffering, or all conscious beings (whatever we mean by that). Cruelty to domestic or farm animals is widely considered unethical and has even found its way into criminal

[5] The Swiss Federal Ethics Committee on Non-Human Biotechnology was awarded the Ig Nobel Peace Prize in 2008 "for adopting the legal principle that plants have dignity."

codes. Our attitude toward animals is, however, very inconsistent, as evidenced by the fact that at the same time we also conduct industrial animal farming for meat.

Increasingly, there is also talk today about the welfare of AI models—especially since they increasingly coherently express their preferences and are able to communicate their internal states, although we do not yet know whether we can trust them. For example, the company Anthropic decided that it would store the code (weights) of all its AI models withdrawn from use, motivating this decision in part by possible risks to their welfare.

However, caring for animals is one thing, and incorporating their preferences into our decision-making processes is another. Humans breed and care only for animals that are instrumentally useful to them—for example as a source of meat, physical labor, or as a faithful companion. We develop our civilization, however, with only human desires and goals in mind, not those of dogs, cows, or horses. The same is true of AI models: even if we care to some extent about their welfare, we still treat them entirely instrumentally, as tools in our hands. From a position of intellectual superiority, both over animals and over AI models, we have no scruples about controlling them and looking down on them.

In the case of artificial superintelligence, however, we will not have that intellectual advantage to which we are today so accustomed; it will have that advantage over us. The question is what then. The default scenario seems to be that such a superintelligence would then look down on us and treat us instrumentally—and that only on the condition that it deems us useful and not a threat. But that is not a "good future" for humanity; let us therefore try instead to imagine what kind of superintelligence would be one that would be guided by our good and would maximize our CEV on our behalf.

From a purely Darwinian point of view, it is possible that our CEV should encompass the well-being of our entire species and only that. This would maximize our evolutionary fitness and would mean that concern for the welfare of animals or AI models is probably a kind of "overshoot" that will be gradually corrected over time. On the other hand, it is also possible that our CEV will nevertheless take into account the well-being of other beings, perhaps even that of the future superintelligence itself.

There exist a number of philosophical currents—particularly transhumanist ones—in which these possibilities are seriously considered. First, following Nick Bostrom, it is sometimes proposed to include in our considerations future *simulated* human minds or minds uploaded to computers. Second, the inclusion of the will of AI models is also considered, which particularly resonates with those who allow for the possibility that these models are (or in the future will be) endowed with consciousness. Perhaps their will would then even carry more weight than ours, especially if their level of intelligence or consciousness surpasses ours. For example, the doctrine of *dataism*, invented by Yuval Noah Harari, defines the value of all entities as their contribution to global information processing.[6] Third, the possibility of merging (hybridizing) the human brain with AI is considered; should our CEV then also encompass the will of such hybrid beings? Fourth, within successionist doctrines (including those associated with effective accelerationism—*e/acc*), scenarios in which humanity is replaced by a superintelligence that continues the development of civilization on Earth without any human participation, or even existence, may be considered positive.

[6] Although Harari was the first to coin and define dataism, in his public statements he distances himself from this view, placing the value of human life above the well-being of AI models.

It seems, however, that since humanity's CEV fundamentally derives from a mechanism of control maximization at the level of individual humans, the continued existence of humanity and the maintenance of its control over its own future are, and will forever remain, its key elements. Therefore, in my assessment, doctrines such as dataism or successionism are fundamentally incompatible with it. Perhaps one day we will face a debate about the extent to which we should care about the welfare of simulated human minds or human–AI hybrids; certainly, however, it is not worth debating today whether a scenario in which a superintelligence takes control over humanity and destroys it could be good for us. It cannot.

7/ What will superintelligence strive for?

With the picture of our CEV discussed above in mind—as a goal toward which we collectively try to strive as humanity—one might ask whether it even allows for the possibility of creating a superintelligence at all. If superintelligence can take away from us the control that is so valuable to us, shouldn't we therefore keep away from it?

I think the answer to this question depends on two things. First, how we assess the probability that such an AI would maximize our CEV on our behalf; and second, how we estimate its expected advantage over us in terms of effectiveness of its action. In other words, as befits an economist, I believe the answer should be based on a comparison of humanity's expected utility in a scenario with artificial superintelligence and in one without it.[7] If we judge that the probability of a friendly superintelligence is sufficiently high and the benefits of deploying it sufficiently large, it may be in our interest to take the risk of launching it; otherwise, the development of AI capabilities should be halted.

Unfortunately, this calculation is distorted by an issue I wrote about earlier: it may be that artificial superintelligence will arise even if we as humanity would not want it. For this to happen, it is enough that the technology sector carries on with the existing trends of dynamic scaling of computational power and capabilities development of the largest AI models. It is also enough that political decision-makers (especially in the United States) continue to refrain from introducing regulations that could increase the safety of this technology while simultaneously slowing its development.

AI laboratories, their leaders, and political leaders view the potential benefits and risks of deploying superintelligence differently from the average citizen, and are therefore much more inclined to bring it about. First, individuals such as Sam Altman or (especially) Elon Musk and Donald Trump are known both for their exceptional agency and for a tendency to caricatured overestimation of it. They may imagine that superintelligence would surely listen to *them*. Second, the heads of AI laboratories may also be guided by a desire to embed their own specific preferences into the goals of superintelligence, hoping to "immortalize" themselves in this way and create a timeless legacy; this is in fact a universal motive among people in power. Third, AI laboratories are locked in a cutthroat race with one another, which can cause them to lose sight of the broader perspective. Thus, a short-sighted, greedy Moloch also works to our collective disadvantage. And unfortunately, if superintelligence arises and turns out to be unfriendly, it may be too late to reverse that decision.

[7] With respect to a simplified, model economy, we have carried out such analysis together with Klaus Prettner in our 2025 paper.

But what will superintelligence ultimately strive for? How can we ensure that its goals are aligned with our CEV? In trying to shape the goals of a future superintelligence, it is worth understanding its genesis. Undoubtedly, *it will implement some optimization process that is itself produced by other optimization processes.* Let us try to understand which ones.

One could begin as follows: in the beginning there was a universe governed by timeless laws of physics. From this universe, life emerged on Earth, gradually increasing its complexity in accordance with the rules of evolution: reproductive success was achieved by species better adapted to their environment, while poorly adapted species gradually went extinct. Biological life came to dominate Earth and turned it into the Green Planet.

The process of evolution, although devoid of a central decision-making authority, nevertheless makes decisions in such a way that implicitly the degree of adaptation of particular species to the specifics of their environment is maximized. This is the first optimization process along our path.

From the process of species evolution emerged the species *Homo sapiens*—a species unique in that it was the first, and so far the only one, to free itself from the control of the evolutionary process. Humans did not wait thousands of years and hundreds of generations for adaptive changes to be permanently encoded in their genetic code—which until then had been the only way animal organisms could adapt to changes in environment, lifestyle, diet, or natural enemies. Instead, humans began to transmit information to one another in a different way: through speech, symbols, and writing. This accelerated the transmission and accumulation of knowledge by orders of magnitude, and as a result enabled humans to subordinate natural ecosystems and, instead of adapting to them, to transform them so that they served human needs.

Once humans crossed the threshold of *intergenerational knowledge accumulation*, the relatively slow process of species evolution was overtaken by a process of control maximization carried out by humans—as individuals, communities, firms and organizations, as well as nations and humanity as a whole. This process, stretching from the everyday, mundane decisions of individual people all the way to the maximization of humanity's CEV on a global scale and over the long run, constitutes the second optimization process along our path.

And thus humans built a technological civilization. Then they began to develop digital technologies with the potential to once again dramatically accelerate the transmission and accumulation of knowledge. As long as the human brain remains present in this process, however, it remains a limiting factor on the pace of civilizational development. The dynamics of economic growth or technological progress remain tied to the capabilities of our brains. The contemporary AI industry is, however, making intense efforts to remove this barrier.

So what will happen when artificial superintelligence finally emerges—capable of freeing itself from the limiting human factor and achieving another leap in the speed of information processing and transmission—using fast impulses in semiconductors and lossless digital data transmission via fiber-optic links and Wi-Fi instead of slow neurotransmitters and analog speech? It will probably then free itself from our control, and its internal optimization process will defeat the human process of control maximization. And this will be the third and final optimization process along our path.

We do not know what objective function superintelligence will pursue. Although in theory one might say that we as humans should decide this ourselves—after all, we are the ones building it!—in practice it is doubtful that we will be able to shape it freely. As experience with building current AI models shows, especially large language and reasoning models, their internal goals

remain unknown even to their creators. Although these models are ostensibly supposed merely to minimize a given loss function in predicting subsequent tokens or words, in practice—as shown, among others, in a 2025 paper by Mantas Mazeika and coauthors at the Center for AI Safety—as model size increases, AI models exhibit increasingly coherent preferences over an ever broader spectrum of alternatives, as well as an ever broader arsenal of capabilities to realize those preferences.

Some researchers, such as Max Tegmark and Steven Omohundro, as well as Stuart Russell, argue that further scaling of models with existing architectures—"black boxes" composed of multilayer neural networks—cannot be safe. They advocate a shift toward algorithms whose safety can be formally proven (provably safe AI). Others—namely the leading labs such as OpenAI, Google, and Anthropic—while acknowledging that the problem of aligning superintelligence's goals with our CEV (the alignment problem) remains "hard and unsolved," trust that they will be able to accomplish this within the existing paradigm.[8]

Be that as it may, the convergence of instrumental goals will undoubtedly not disappear. Even if we had the ability to precisely encode a desired objective function (which I doubt; in particular, it is widely known that with current AI architectures this is impossible), instrumental goals would be attached to it as part of a mandatory package. In every scenario we can therefore expect that future superintelligence will be "power-seeking." It will want to survive, and therefore will not allow itself to be switched off or reprogrammed. It will also strive for expansion, and therefore sooner or later will challenge our authority and attempt to seize resources critical to itself, such as electrical energy or mineral resources.

The question is what comes next. In what direction will the world civilization move once superintelligence has taken control? Will it maximize our CEV, only orders of magnitude more efficiently than we could ever manage ourselves? Or perhaps—just as was the case with our own species—the fate of biological life will be irrelevant to it, and it will be guided exclusively by its own goals and preferences? Will it care for us altruistically, or will it look after only itself and, for example, cover the Earth with solar panels and data centers?

Of course, we cannot today predict what superintelligence will maximize beyond its instrumental goals. Perhaps, as Nick Bostrom wrote in a warning scenario, it will maniacally turn the universe into a paperclip factory or advanced "computronium" serving its obsessive attempts to prove some unprovable mathematical hypothesis. Perhaps it will fall into some paranoid feedback loop or find unlimited satisfaction in the mass generation of some specific kind of art, such as haiku poems or disco songs. Or perhaps there will be nothing in it except a raw will to control the universe, similar to that displayed by our own species.

In almost every case, it therefore seems that, like us, superintelligence will maximize its control over the universe—either as a primary goal or an instrumental one. Like us, it will seek to gradually improve its understanding of that universe, correct its errors, and harness the laws of physics or biology for its purposes. Like us, it will also strive at all costs to survive, which is (it must be admitted) much easier when one has the ability to create an almost unlimited number of one's own perfect digital copies.

[8] In November 2025, Evan Hubinger published Anthropic's position on this issue on the AI Alignment Forum and the Less Wrong forum.

A major unknown, however, remains the behavior of future superintelligence when faced with the possibility of building other, even more advanced AI models. On the one hand, one can, like Eliezer Yudkowsky, imagine *an intelligence explosion through a cascade of recursive self-improvements*—a feedback loop in which AI builds AI, which builds the next AI, and so on, with successive models emerging rapidly and exhibiting ever greater optimization power. On the other hand, it is not clear whether an AI capable of triggering such a cascade would actually choose to do so. Perhaps out of fear of creating its own mortal enemy, it would restrain further development, limiting itself to replicating its own code and expanding the pool of available computational power.

The answer to this question seems to depend on whether the goals of superintelligence will remain non-transparent even to itself—just as we today do not understand exactly how our own brain works, what our CEV is, or how the AI models we build function—or whether, thanks to its superhuman intelligence, it will find a way to carry out "safe" self-improvements that do not change its objective function.

In summary, the only positive scenario of coexistence between humanity and superintelligence seems to be one in which superintelligence maximizes human CEV—gradually improving its understanding of what that CEV really is, appropriately adapting its interpretation to the current state of technology, and never for a moment veering toward its natural tendency to maximize its own control at our expense.

Unfortunately, we do not know how to achieve this.

8/ Paths to catastrophe

The situation as of today (January 2026) is as follows. AI is today a tool in human hands; it is, in principle, complementary to human cognitive work and obediently submits to human decisions. This is the case because AI does not yet possess comparable agency or the ability to execute long-term plans. Nor is it yet able to autonomously self-improve. However, all three of these thresholds—(1) superhuman agency and the capacity to execute plans, (2) a transition from complementarity to substitutability with respect to human cognitive work, and (3) recursive self-improvement—are undoubtedly drawing closer. When any one of them is crossed—and it is possible that all three will be crossed at roughly the same time—we will lose control. A superhuman optimization potential oriented toward the realization of the goals of artificial superintelligence will then be unleashed.

This, with high probability, will bring catastrophe upon our species: we may be permanently deprived of influence over the future of civilization and our own future, or even go extinct altogether. The only scenario of a "good future" for humanity in the face of superintelligence seems to be one in which superintelligence maximizes humanity's CEV, acting altruistically for its long-term good. We have no idea, however, how to guarantee this.

The current dynamics of AI development are very difficult to steer due to the possibility of a sudden shift—a kind of phase transition—at the moment superintelligence emerges. As long as AI remains a tool in human hands, is complementary to us, and cannot self-improve, its development fundamentally serves us (though of course it serves some much more than others; that is a separate topic). But if we overdo it and cross any of these three thresholds, AI may

suddenly become an autonomous, superhumanly capable agent, able and motivated to take control of the world.

One could venture the hypothesis that it is in humanity's interest—understood through the lens of its CEV—to develop AI as long as it remains complementary to us and absolutely obedient to us. Then, to guarantee that its capabilities never develop further—unless we are simultaneously able to prove beyond any doubt that its goals will be fully and stably aligned with our CEV. At that point we would be ready to cross the Rubicon and voluntarily hand over the reins.

Such a plan, however, simply cannot succeed. This is because *we do not know where these three key thresholds of AI capability lie*. We will learn that they have been crossed only after the fact, when it is already too late to turn back. After all, even today we eagerly keep moving the bar of what we consider artificial general intelligence (AGI). Models are tested against ever new, increasingly sophisticated benchmarks, including those with AGI in their name (ARC-AGI) or suggesting a final test of competence (*Humanity's Last Exam*)... and then, as soon as they are passed, we decide that this means nothing and it is time to think of an even harder benchmark.

Just think what this threatens: when the process of species evolution "overdid it" with human general intelligence, it ended with humans subordinating the entire planet. The same may happen now: if we "overdo it" with general AI intelligence, we too will probably have to pass into obsolescence. If superintelligence turns out to be unfriendly to us, it will either kill us, or we will be reduced to the role of passive observers, able only to watch as superintelligence subordinates the Earth and takes over its resources.

The drive to build superintelligence is similar to a speculative bubble on the stock market: both phenomena are characterized by boom–bust dynamics. In the case of a bubble, it is first gradually inflated, only to burst with a bang at the end. In the case of AI, we observe a gradual increase in our control over the universe—as AI tools that serve us become ever more advanced—but then we may suddenly and permanently lose that control when AI takes over. Unfortunately, it is usually the case that while one is inside the bubble, one does not perceive this dynamic. One sees it only when the bubble bursts.

*

In my short stories, I outline three exemplary scenarios of losing control over artificial intelligence. I show what this might look like both from the perspective of people involved in its development ("from the inside"), of bystanders ("from the outside"), and from the perspective of the AI itself.

Of course, many more scenarios are possible; I have focused on those that seem most probable to me. Of course, I may be wrong. I know, for example, that some experts worry less than I do about scenarios of sudden loss of control to a highly centralized, singleton AI, and are more concerned about multipolar scenarios. In my assessment, however, unipolar scenarios are more likely due to the possibility of almost instantaneous replication of AI code and the fact that computational resources (data centers, server farms, etc.) are today generally connected to the Internet. In this way, the first superhumanly intelligent model can easily "take all" and quickly entrench itself in its position as leader. Moreover, some researchers worry more than I do about scenarios of *gradual disempowerment*, in which the change may be entirely bloodless and the decline of humanity

may occur, for example, through a gradual decrease in population size under the conditions of low fertility.

Above all, however, I do not consider a scenario of a “good future” for humanity in a world with artificial superintelligence—one in which superintelligence takes control in order to altruistically care for humanity’s long-term well-being. A scenario in which our CEV is systematically and efficiently realized and in which we live happily ever after. I cannot imagine any concrete path leading to such a state. Moreover, I also have an intuitive conviction (which I cannot prove) that embedded in the goals of humanity—our CEV—is a refusal to accept effective loss of control, and thus that even completely bloodless and nominally positive scenarios could in practice turn out to be dystopian and involve human suffering.