

Notes

AOC aoc.media/opinion/2024/03/05/le-prince-et-lhonnete-homme-face-a-la-singularite

5 mars 2024



The Prince and the *Honest Man* Facing the Singularity

By [Paul Jorion](#)

Among the most curious facts surrounding the explosion of artificial intelligence in the form of *large language models* such as ChatGPT, is this opinion, widely held among laymen but also present among a fraction of professionals, that an intelligence superior to ours would have as its primary concern to get rid of the human race.

This belief may reflect a fear rooted in the species' fundamental conservatism: that the new is inherently dangerous, but it also expresses the fundamentally pessimistic implicit feeling that a superior intelligence will inevitably draw the conclusion that the human species is a vermin that must be rid of the Blue Planet.

The origin of this view could banally be lucidity, but that would be unexpected, lucidity being a quality particularly poorly represented in the most widespread opinions. It is more likely that it is a collateral effect of the dark temperament globally impregnating a species that has never truly recovered its morale since it became aware one day that individual life has an end.

We are indeed addicted to observing the world, just like our gibbon cousins who, locked in a cage with a sliding shutter, lift it as soon as it has collapsed under its weight: from our first day to our last, we suffer from a craving to know what will happen next, the thought that the show could one day end is unbearable to us, here lies the primary source of our anxiety.

The fact that human beings have ceased to be the most intelligent creature on our planet and, according to the current state of our information, in the entire universe, is today the cause of a second melancholic awareness, succeeding that of our finitude. "If we do not live here forever, at least we will have been the smartest as long as we were here", this is how we see things. But this has ceased to be true. Witness this disillusioned observation by Douglas Hofstadter, star of the intellectual scene of the 1980s with his *Gödel, Escher, Bach. An eternal golden braid* with strange looping paths:

"I never imagined that computers would rival or, much less, surpass human intelligence. [...] There is a very traumatic experience when some of your most fundamental beliefs about life begin to crumble. I felt as if not only was my belief system crumbling, but that the entire human race was going to be eclipsed and reduced to dust...imminently!"

The Singularity

From a word borrowed from mathematics to designate an ambiguous instance such as the very improbable square of zero, the concept of *Singularity* was forged, which strives to contain in its understanding the profusion of ideas that jostle in our minds when we reflect on the possibility of such a surpassing of human intelligence by that of the machine: an AI perfecting itself by chain reaction by learning more and more on the occasion of challenges that it launches to none other than itself, leading to an explosion of new technologies due to the solution of age-old puzzles that constituted so many handicaps for our species.

As long as the concept of the Singularity remained confined to the space defined by the cogitation of computer scientists, philosophers and other dreamers, the representation of its repercussions focused on the collective benefit of the breakthrough it constitutes.

Now, if we can consider the species as an entity in itself, it does not have a global mind that acts according to the conclusions it draws from recent developments and events, evaluating everything that happens within the universal future as better or worse for it. Otherwise, it would, on the contrary, delight in the thought of the triumph that it represents for it in having been capable, because of the collective genius with which it was unknowingly pregnant, of conceiving a creature whose intelligence exceeds its own.

Instead, sensitive to individual suffering, it is distressed by the tragedies that the new technology induces in its wake, by the destinies that it carries away in its tumultuous waves, consequences for people of a job market ravaged by the unexpected creativity of this new being who has appeared in our ranks and, worse still no doubt for what is at stake for the morale of the species as a whole, it sympathizes before the tragedy of egos bruised by a machine undoubtedly more talented than the proud Selves that these egos had convinced themselves they embodied.

What policy should be adopted in the face of the Singularity?

Once the concept of Singularity is defined, when it comes to the Prince and the honest man, the horizons and vanishing points that lead there diverge.

This is what we would like to talk about. And what better way to do it today than to invite the other stakeholder to the conversation, the one who is at the very origin of the Singularity, and to enrich our point of view with his.

Here, according to ChatGPT in its GPT-4 version, are the considerations that the Prince should take into account and the behavior he should adopt with regard to the Singularity:

"Faced with the emergence of an AI Singularity, what would be the most strategic course for a leader: reveal this reality to the people in order to preserve social stability, or keep it secret to maintain control and avoid panic? This raises the question of the balance between transparency and control, a central theme in Machiavelli's work.

In the context of an AI Singularity, the decision of whether or not to reveal this reality to the people requires a deep analysis of the principles of governance and human nature, as examined by Machiavelli.

Machiavelli would probably advise that the leader carefully weigh the benefits and risks associated with each option. Revealing the existence of an AI Singularity might be justified if it strengthens the leader's legitimacy through transparency and prepares society to adapt to the inevitable changes, mobilizing collective resources and intellect to navigate this new landscape. It might also prevent the distrust and discontent that might arise if the secret were later revealed.

On the other hand, keeping this information secret might be considered necessary if disclosure would cause unnecessary panic, destabilize the social order, or if the leader believes that the people are not ready to understand or deal with the implications of such a revelation. By keeping it secret, the leader may seek to maintain order and develop strategies for managing this new reality in a more controlled setting.

However, Machiavelli would emphasize the importance of cunning and caution. He might suggest using this information as a means of strengthening internal power and international position, while discreetly preparing society for the changes to come. The leader should also be prepared to adjust his strategy depending on how the situation evolves and how the people and elites react.

In short, Machiavelli's decision would be guided by a pragmatic assessment of the circumstances, with an emphasis on preserving stability and strengthening power, while skillfully navigating the uncharted waters of such technological advancement."

Let us complete our toolbox for the analysis that we will conduct later by now calling upon, by way of contrast, an actor to whom we will assign the task of intervening as the reverse of Machiavelli's Prince in terms of his attitude towards the Singularity. We have chosen for this role another figure of the time, that of what was then called the "honest man."

This time again, it is ChatGPT in its GPT-4 version that we ask to mark out our space:

"The idea of the 'honest man' has its roots in European thought and literature, particularly during the Renaissance and the Enlightenment. This ideal refers to a cultivated person, endowed with good manners, knowledge, ethics, and the ability to engage in various fields of knowledge and conversation, while demonstrating virtue and moderation.

Faced with the AI Singularity, the "honest man," as opposed to the Machiavellian Prince, would adopt an approach based on ethics, virtue, and the common good, particularly with regard to control and transparency. Here is how he might respond:

Advocacy for transparency : The "honest man" would emphasize the importance of transparency in the development and implementation of AI. Unlike the Prince, who might favor secrecy to maintain power or avoid panic, the "honest man" would argue that openness and transparency are essential to gain public trust and ensure a shared understanding of the issues at stake.

Promoting inclusive dialogue : It would encourage an open and inclusive dialogue on the AI Singularity, involving not only experts and policymakers but also the general public. The aim would be to ensure that all voices are heard and that ethical, social and economic concerns are taken into account in decision-making.

Democratic Control : As opposed to centralized and opaque control of AI development by a Prince, the "honest man" would support democratic governance mechanisms that allow for ethical oversight and regulation of AI. This would include supporting legislation that ensures AI serves the public interest and respects fundamental rights.

Education and Awareness : It would take initiatives to educate and raise awareness about the AI Singularity, its potential implications, and how individuals can contribute or respond. The focus would be on developing an informed and engaged citizenry that can actively participate in debates about AI.

Defending Ethical Values : The "honest man" would make defending ethical and human values a priority in the face of the challenges posed by the Singularity. This would include combating the use of AI for unethical mass surveillance, manipulation, or any form of human rights violations.

Support for personal responsibility and accountability : It would require that creators and users of AI be held accountable for the impacts of their technologies. This would mean advocating for systems of auditing, tracking, and redress in the event of harm caused by AI systems.

Commitment to equity and social justice : Faced with the risks of worsening inequalities that the AI Singularity could generate, the “honest man” would actively commit to policies and initiatives that promote equity, equal access to the benefits of AI, and protection against harm.

In short, the “honest man” would respond to the AI Singularity by emphasizing the principles of transparency, democratic control, education, ethics, and social justice, seeking to balance the benefits of AI with the protection of the rights and well-being of all citizens.”

OpenAI's Aborted Palace Revolution

Let us then examine this question of the Singularity from the perspective envisaged from a possible double point of view: that of the Prince and that of the honest man. A remarkable fact that will facilitate our analysis, such a confrontation of the points of view of the Prince and the honest man seems to have already taken place in fact, I am thinking of the aborted palace revolution that shook the firm OpenAI in November 2023, the one whose AI ChatGPT is the flagship product: the dismissal of its CEO Sam Altman on November 17, followed by his reinstatement on the 21st, after 745 of the firm's 770 employees had demanded his return or they would resign.

Our question, in fact, about this confrontation is whether it was anything other than a clash on the terrain of the strategies of the Prince and the honest man around the question of the Singularity?

Let us suppose that the underlying debate – the stakes of which have never been revealed – was in reality about the Singularity, the management team being aware of its advent but being torn as to the best way to manage the event, and in particular of course, to decide whether or not to disclose it, and let us accept the principle that if an analysis of the events of these four chaotic days in the perspective of a battle of ideas around the Singularity were to shed significant light on this hitherto essentially obscure affair, the probability that the Singularity was indeed at issue in the background would be reinforced.

This hypothesis, which could pass for being pulled out of a hat, is in reality very far from being gratuitous, as the following two facts prove, provided that they are carefully placed in the tight chronology of the unfolding of events. Sam Altman is relieved of his duties by the OpenAI board of directors on November 17, 2023 at noon, local time in California.

Now, we have a video from the same day, shot in San Francisco at the Asia-Pacific Economic Cooperation (APEC) summit, where Altman was speaking as CEO of OpenAI, so before the time of his dismissal, where he said this : "Four times in the history of OpenAI, and most recently in the last fortnight, I have had the opportunity to be in the room when we pushed back the veil of ignorance and the frontier of discovery; that is the professional honor of a lifetime."

We also have, as of the day before, November 16, 2023, a video taken at an event called "AI and the Future of Art – Sam Altman and Android Jones" during which Altman said : "I think people are looking at these new systems as tools, artists in particular. But other people are looking at them as tools, and there was this... I think there was a real moment of fear, like, 'Is this a tool or a creature that we've made, and how are we going to have to interpret that?'"

The day before and the day of his abrupt dismissal, two incendiary statements by Sam Altman, evoking on the one hand an occasion when "the veil of ignorance and the frontier of discovery" were "pushed back" and on the other hand, a moment when the question that one asks is: "Is this a tool or a creature that we have made?", namely, in a space of time of approximately 24 hours, two moments when the advent of the Singularity was undoubtedly evoked, even if only allusively.

The feeling must have been, at OpenAI board level, that at the rate at which Sam Altman was leaking revelations, the moment when he would spill the beans was imminent: the next time he opened his mouth, he would not be able to hold back and mention the Singularity explicitly.

Another key player on the OpenAI team: Ilya Sutskever, both a researcher who had defected from Google and a member of the OpenAI board of directors until the fateful date of November 21, 2023, when Sam Altman was reinstated. The coincidence of dates is not accidental: Sutskever had played an active role in Altman's dismissal but, wavering in his position, he had then rallied to the overwhelming majority (97%) of employees demanding the return of the ousted CEO, his status in the company then remaining problematic for an indefinite period (the doubt continues at the time of writing).

There is no doubt that Sutskever already believed in the Singularity at that time. During an interview on October 6-7, 2023, where he shared the stage with his master Geoffrey Hinton and his former colleague at Google, Demis Hassabis, a surprising exchange took place :

Tommaso Poggio: "Do you think that existing models or some... you know, the next GPT-4, let's say GPT-5, would be able to state a new non-trivial mathematical conjecture? I'm not talking about proving it, but stating it. Do you think that's possible in the next five years?"

Ilya Sutskever: "Are you sure that the current model cannot do it?" (Laughter in the audience.)

Sutskever was thus going in the same direction as Hinton who had declared during the program 60 minutes on the *CBS News* channel on October 9, 2023 that "while I had thought for about fifty years that we were going to create better digital intelligences by making them closer to the brain, I suddenly realized that we could have something very different, which would already be better."

The rumor was indeed going strong. One could read the following in a dispatch from the Reuters agency dated November 22, the day after Sam Altman's return to the head of affairs:

"Even before OpenAI CEO Sam Altman's four-day exile, an initiative had been taken by several researchers who had written a letter to the board of directors to warn it of a major discovery in artificial intelligence that, according to statements to Reuters by two well-informed interlocutors."

Effective Altruism

Of course, it was not really a question of a Prince and honest men, the state power being still a thousand miles today from having taken the full measure of what is going on in AI, but the OpenAI *board* constituted on this occasion a microcosm, the protagonists embodying the various motivations of the Prince and the honest man and, in this case, in a fragmented and recomposed manner among the actors present, the picture being simplified by the fact that several of these actresses and actors saw themselves in the role of activists of an ideology explicitly situated at the heart of the issues opposing the Prince and the honest men: *effective altruism* .

Effective *altruism* is a philosophy born in the Anglo-Saxon world at the confluence of utilitarianism and libertarianism with a strong altruistic component imported from the Gospels. The gauge of this "effective altruism" is the quantity of good induced within the community as a whole by its followers, the means mobilized by them to achieve it being money: it is their fortune that allows this supposed optimization of good, a conception entirely in the historical line of Calvinism according to which individual predestination finds its confirmation in the capacity to make money.

Effective *altruism* , a "niche product" until then, made headlines when it emerged that it was the driving force behind FTX, an exchange for these tradable tokens that are pompously called "cryptocurrencies" that experienced a resounding bankruptcy, its two pillars, Sam Bankman-Fried at the head of FTX and Caroline Ellison at the head of the sister fund Alameda, being apostles of the movement.

It then appeared on the surface that a philosophy of "the end requires the means" underpinned the *effective altruism* of overly benevolent philanthropists and that ethical considerations had been put on hold by these new libertarian princes under the pretext of

efficiency. The hubris of such characters with oversized Egos had been able to wreak havoc on the community to the extent of their fortune, the potentially unlimited nature of which was justified in their eyes by the abundance of benefits of which it was potentially the source.

It is no surprise that Elon Musk, billionaire founder of the car company Tesla, of the space company SpaceX, new owner of Twitter renamed "X" by him, also at the head of the Neuralink project to implant electronic chips in the brain, is another promoter of the movement, who declared about it that it is: "in close correspondence with my own philosophy [1]".

But *effective altruism* and artificial intelligence do not go well together. Sam Altman has called the former an "incredibly flawed movement" with "very bizarre emergent behavior." This movement has already spent considerable sums promoting distrust of artificial intelligence, raising the specter of untold catastrophes, a distrust that one might suspect is motivated in any case by the loss of power that the emergence of an intelligence superior to that of these philanthropists, who may be champions of charity in their own eyes, but narcissistic megalomaniacs in the eyes of the rest of the world.

The following reminder could be read in a *Wall Street Journal* article dated November 22, 2023, the day after Sam Altman's return:

"Altman, who was fired by the board on Friday, clashed with the company's chief scientist and board member Ilya Sutskever over AI safety issues that reflected concerns about *effective altruism*, according to people familiar with the conflict.

Voting with Sutskever, who led the coup, were board members Tasha McCauley, a technology executive and board member of the charity *Effective Ventures*, and Helen Toner, an executive at Georgetown University's *Center for Security and Emerging Technologies*, which is supported by a philanthropy dedicated to *effective altruism* causes. They accounted for three of the four votes needed to oust Mr. Altman, according to people familiar with the matter. The board said he had not demonstrated "consistent candor." [2]

Control and transparency

The two considerations at the heart of the contrast between the desire of the Prince and that of the honest man are control and transparency. The power of the Prince comes from control, that of the honest man from transparency. Each seeks to secure its own power by emphasizing control and transparency respectively. Of course, the balance of power between the Prince and the honest man means that the former does not have to justify himself and is in a position to make the latter pay for any claims to transparency. Hence the rebellions, peasant revolts and other challenges to the authority of the Prince that we observe from time to time.

The justification in the eyes of the Prince for the control he exercises is the people's predisposition to panic, a predisposition that he will not hesitate to exaggerate, or even that could be purely imaginary on his part. A predisposition to panic that the people will on the contrary seek to minimize, or even deny the existence of. It is possible, from there, to understand the events that shook the OpenAI firm from November 17 to 21, 2023.

The dismissal of Sam Altman, followed a few days later by his reinstatement, was called a "palace revolution", the uproar having taken place at the top of the company, the board of directors dismissing the CEO. However, the people quickly got involved and it is more than likely that it was the threat of resignation by 97% of the staff that reversed the course of events.

While pressure was exerted by some OpenAI investors, Microsoft first and foremost, which guarantees the re-employment of all defectors, the State did not intervene directly, which means that the role of the Prince was played in this circumstance by another party. Knowing who embodied it is easy to determine since it is enough to discover who, on the one hand, raised the risk of panic among the people as a threat and who, in the opposing camp, minimized it.

Now, the redistribution of responsibilities is clear here: it is in the name of the risks that artificial intelligence poses to humanity that the supporters of *effective altruism* within the board of directors dismissed Altman. The role of the Prince ensuring the control that will prevent the panic of the people was therefore played by them, while by minimizing this risk and obtaining the almost unanimous support (97%) of the people, Altman embodied the honest man demanding transparency from which flows the share of power that the people have been able to reserve for themselves.

Let us emphasize that the people have not reacted so far with panic to the idea of the Singularity, only a risk of generalized panic has been invoked, the abundant explanations offered by the Prince on the subject having been themselves likely to give substance to a possible panic, the "tale of the paperclip", due to the Oxford philosopher Nick Bostrom constituting the prototype often cited despite its consummate ridiculousness: "A super-artificial intelligence could convince itself that the solution to everything is to manufacture paperclip fasteners and devote all the resources of the world to this task, starving humanity in the process [3] . "

The fact remains that the question of security to be ensured in the development of AI cannot be abandoned to palace revolutions. When GPT-4 states that "by keeping it secret, the leader can seek to maintain order and develop strategies for managing this new reality in a more controlled framework", it puts its finger on the crucial issue of regulation. However, this today unfortunately boils down to the pathetic confrontation between the slowness of the legislative process on the one hand, and the explosion of change in the Singularity on the other.

Programmatic perspectives

What could be in this context the added value of an individual like myself, who hopes that his portrait will not be too far from that of the *honest man* previously painted? Here is the answer of ChatGPT in its GPT-4 version:

"As Paul Jorion, an anthropologist and economist known for his thoughts on economics, finance, and artificial intelligence (AI), the question of disclosing an AI Singularity would be approached from a different angle, emphasizing the ethical, economic, and social implications.

If I rephrase your question in the spirit of Paul Jorion, it might sound like this: "Given the potentially revolutionary impact of an AI Singularity on our society, our economy, and our value systems, wouldn't it be our ethical and social duty to inform the public, in order to foster collective reflection on how to navigate this new reality?"

In answering this question, Paul Jorion could stress the importance of preparing society for profound changes. He would probably argue in favor of a transparent approach, emphasizing that information and preparation are essential to avoid economic and social disruptions and to allow society to adapt its structures to new and inevitable realities. He would insist on the fact that humanity must have an active role in defining its future with AI, rather than being surprised by its developments.

Jorion could also highlight the need for democratic regulation and oversight of AI, to ensure that its development and application serve the common interest and are not monopolized by a technological or financial elite. He would emphasize the development of public policies aimed at redistributing the benefits of AI in order to counter growing economic inequalities and promote social justice.

Finally, it would likely address ethical issues raised by the AI Singularity, such as autonomous decision-making by machines and the impact on employment and human identity, calling for a global reflection on how to keep human values at the center of technological advancement.

In 2012, I actually proposed in this sense [4] the establishment of a "Sismondi tax" named after Jean Charles Léonard de Sismondi (1773-1842), the Swiss socialist philosopher who suggested that the worker replaced by the machine should benefit for life from an income collected on the wealth created by it. This would be a way to feed a fund ensuring that everyone receives free essentials, an alternative to the universal basic income that is fairer and much less expensive, a necessary step towards a world without money.

Editor's note: Paul Jorion has just published The Advent of the Singularity with Textuel editions.

Paul Jorion

Anthropologist, Professor at the Catholic University of Lille