# TOWARDS AN AIMLESS EXISTENCE – A DIALOGUE ABOUT AI'S POTENTIAL TO RADICALLY CHANGE THE HUMAN CONDITION

**Mikael Laaksoharju, Iordanis Kavathatzopoulos**

Department of Information Technology, Uppsala University (Sweden)

Mikael.Laaksoharju@it.uu.se; Iordanis.Kavathatzopoulos@it.uu.se

**EXTENDED ABSTRACT**

This presentation will be given in the form of a dialogue between Laaksoharju and Kavathatzopoulos. Let us start with an introduction.

In the recent years, dystopian prophecies regarding artificial intelligence (AI) have garnered public attention. For instance, the risk of AI becoming exponentially more powerful than all human intelligence combined, acquiring an independent existence of itself, transforming us into something we do not want to be, evolving in a radically different way, even affect the whole universe, etc. (Kurzweil, 2006; Bostrom, 2014; O'Neil, 2016; Harari, 2016; Reese, 2018; Tegmark, 2017; see also Future of Life Institute, 2023). Although not everyone agrees on whether any of these things will happen, or when they might happen, these prophets have in common that they focus mainly on the technical aspects of the issue or on AI itself and its purported potential. There is, however, another interesting – and in our view more relevant – angle to AI as a phenomenon, namely the effects that even weak but well-functioning AI could have on human nature, or life in general.

The complexity and opacity of many AI algorithms is often called out as a great risk of potentially losing control over the algorithms. Consequently, large research efforts have been invested in what is called "Explainable AI". We do not wish to dispute the importance of this research area but perhaps additional considerations can be added to nuance the ambition.

First of all, complexity and opacity in themselves do not necessarily imply loss of control, if the algorithms are completely predictable. For instance, most people do not know how the technology in a regular car works and the trend has been that in every new generation of cars even more of the underlying technology is hidden from car owners. As long as the car functions as expected, this is likely net beneficial for the cognitively overloaded (post)modern individual.

Some will argue that the trend of hiding technological complexity threatens human autonomy and indeed there is a sacrifice of autonomy in, e.g., giving up some control over your vehicle. Nevertheless, most seem to consider that the benefits outweigh the minor harm in giving up low-level control. Judging from the public attitude, this will be a likely fate of algorithms as well when they start producing consistently reliable results. Most of us will have no interest in scrutinizing the process behind an algorithm's recommendation or classification when it is perceived as accurate.

However, this means a cognitive cost. When algorithms will be perceived as reliable, they will start entering *the fabric of truth production*, much like calculators have been elevated to determining the correct result of arithmetic calculations and how statistical tests have come to represent the existence of correlations. The difference here is that AI algorithms can tell the "truth" about so much more than statistical tests. They will be able to decide for us whether we

are looking at a picture of a sloth or a chocolate croissant (see e.g. Zack 2016, Alasadi 2019). One day we may have become so used to trusting the algorithm so that instead of musing over how similar some sloths and some chocolate croissants look in some photos, we will be fascinated with how some photos of croissants actually could have been photos of sloths, and vice versa, if the computer did not tell us the truth. When algorithms are accurate almost every time, we will quickly lose our current skepticism towards them, simply because skepticism is unnecessarily burdensome. In other words, what the algorithms tell us is the truth will be the most convenient belief. Solomon Asch (1956) would not become surprised.

Here it is time to introduce the different positions of Laaksoharju and Kavathatzopoulos. Kavathatzopoulos (2024) claims that potent (weak) AI could become an existential threat by fulfilling our needs to the extent that the human capability to reason will eventually wane. After all, if any human goal can be fulfilled by AI, why would we ever need to practice our reasoning ability? In a sense, this is a philosophically and psychologically founded Wall-E prophecy of the future.

Laaksoharju claims that this prediction is based on an overly teleological assumption of both human behaviour and of AI. When it comes to humans, goal fulfillment, as a construct, is not a necessary condition for activating thinking, and when it comes to AI, the goals that are currently formulated and assessed are in the form of tasks for which there exist ground truths that have been somewhat arbitrarily decided by humans. The implication of this is that the current success of algorithms is more of a social construct than something that corresponds to any actual human need; a self-selected group of arbiters' have chosen what problems are relevant to be processed by AI and then deemed these problems as solved to some extent.

Before proceeding, it should be mentioned that the positioning of human nature outside of, or in opposition to, technology is limiting. In line with the views of Bernard Stiegler (1998), we see technology as a response to human/organizational/societal values and by that entangled with human nature. Humans do not primarily interact with technology but their interaction with other humans is augmented/mediated/supplemented by technology. With this lens, nuclear weapons, for instance, are arguments in negotiations about territorial power, and AI algorithms are arguments in negotiations about power in general.

In essence, the dialogue is revolving around the concept of goals and its importance for human existence. If Kavathatzopoulos is correct, even weak but effective AI will lead to the demise of humanity. If Laaksoharju is correct, strong AI is still a pipe dream and weak AI will be just like any technology introduction – it will change the logic of existence to some extent but humans will find new ways to compete with each other. The questions to be addressed in the dialogue are thus:

1. Do goals "exist" or not?
   Kavathatzopoulos claims that goals are an integral part of thinking/life, which emerged because of uncertainties in the kinesis of the world. However, if they do exist, they are either real (which conflicts with the perception of the world as chaotic motion, which itself is a prerequisite for the existence of goals since goals have meaning in a world of uncertainty; real goals seem to be a contradiction in terms) or the goals are an illusion of thinking (which is in accordance with the world as chaos, i.e., goals arise in uncertainty, as something to be sought, identified, and pursued, but they cannot be real because then the whole process/thinking would be "locked").

Laaksoharju simply refutes the explanatory value of “goals” for understanding human behavior and instead claims that humans largely act by following mental patterns that are activated by stimuli in their lifeworlds. However, Kavathatzopoulos means that there is no explanatory value of “goals” but they are there together with the thinking process, which is not possible to run without both of them.

2. Are self-determined goals sufficient for AI to become "autonomous"?
   Kavathatzopoulos will argue for the possibility that if goals arise in connection with uncertainty and life emerges, perhaps facilitating "goals" for AI will open up the possibility for AI to have its own "life". Perhaps the more one "confuses" AI regarding which goals to strive for or invent on its own, the harder it will be for AI to become autonomous.

   Laaksoharju will argue that goals can be useful for programming sensing systems to determine how to regulate their behaviors, but that this goal-directed behavior will not lead to anything similar to human consciousness in machines.

   The ironical conclusion of this introduction is that the predictions of both Laaksoharju and Kavathatzopoulos will lead to an understanding of human existence as aimless, with the difference that we are either experiencing it already now or we will as soon as we have perfected AI to fulfill all our desires.

**KEYWORDS:** Artificial general intelligence, motivation, existential threat, alignment problem.

**REFERENCES**

Alasadi, Z. (2019). *Is it a Sloth or a Chocolate Croissant*? Github. https://github.com/zainabalasadi/sloth-or-croissant

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and applied*, 70(9), 1.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Future of Life Institute. (2023). *Pause giant AI experiments: An open letter.* https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Harari, Y. N. (2016). *Homo Deus: A Brief History of Tomorrow.* Random House.

Kavathatzopoulos, I. (2024). Artificial Intelligence and the sustainability of thinking: How AI may destroy us, or help us. In T. T. Lennerfors and K. Murata (Eds.), Ethics and Sustainability in Digital Cultures (pp. 19–30). London: Routledge.

Kurzweil, R. (2006). *The Singularity is near: When humans transcend biology.* Penguin Books.

O’Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown.

Reese, B. (2018). *The fourth age: Smart robots, conscious computers, and the future of humanity.* Atria Books.

Stiegler, B., 1998, *Technics and Time, 1: The Fault of Epimetheus*, Stanford: Stanford University Press.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence.* Knopf Publishing Group.

Zack, K. (2016, March 10). *Chihuahua or muffin* [Tweet]. Twitter. https://twitter.com/teenybiscuit/status/707727863571582978