Bulletin of the *Transilvania* University of Braşov
Series VII: Social Sciences • Law • Vol. 18(67) No. 2 – 2025
https://doi.org/10.31926/but.ssl.2025.18.67.2.15

HOW ETHICAL WILL ARTIFICIAL INTELLIGENCE BECOME

Fănel STROE [1]

Abstract: *As Artificial Intelligence begins to potentially replicate or even surpass certain human cognitive and ethical capabilities, we are compelled to rediscover and revalue those uniquely human qualities that transcend algorithmic calculation—spiritual depth, creative imperfection, and the capacity for unconditional empathy. The traditional Turing Test needs to be supplemented with an ethical variant—one that measures the capacity for empathy, moral reasoning, and understanding of human complexity. Human imperfection may prove to be our most valuable asset in the age of artificial intelligence. It is precisely our vulnerability, our capacity for error, and our emotional complexity that could preserve the distinction between human and artificial intelligence.*

Key words: *redefinition of humanity, inevitability of coexistence with Artificial Intelligence, ethical responsibility, ethical Turing Test, the paradox of imperfection*

1. Introduction

It is said that when the first American vessels reached the Chinese coast, they brought with them the message that in America, the streets were paved with gold. And many Chinese, longing for a better life, paid a price that was far too high, even with their lives, for a dream. The legend of the hero Wong Fei Hung also says that the real gold is right where you are, you just have to search for it very carefully.

After almost two centuries since those events, a new El Dorado is appearing on the horizon. It is the AI myth.

Certainly, AI will develop and take over many of our current tasks. However, right before our eyes, behind the much-touted performances of AI, a process is unfolding—the transfer of decision-making from each of us to an AI center, administered by a company listed on a stock exchange. Humanity's most beautiful dreams have never been innocent, although humanity has never ceased to hope. This is perhaps why in social sciences we have the criterion of minimizing risks and not that of choosing the greatest happiness.

Yet, AI propaganda currently far surpasses its real performance, exploiting the greater

[1] *Transilvania* University of Braşov, fanel.stroe@unitbv.ro, corresponding author.

or lesser weaknesses of the *homo sapiens sapiens* species—weaknesses that are real, imagined, or constructed through various processes of social engineering.

The intention of this article is to offer a perspective on how ethical AI can become. Could AI eventually construct an argument that leads to the suppression of human life on Earth? Or, perhaps, with the help of AI, can we hope for a system akin to Democracy 2.0?

2. The New Order or the Old One?

The order that AI will replicate will be an economic order. And this order is by no means one of equilibrium between supply and demand, but rather one of Schumpeter's creative destruction or an order of Peter Thiel's monopolies—an economic order permanently in search of something more valuable, more performant, and more attractive, yet, at the same time, more efficient, requiring less effort for a maximum effect.

Since all these are expressions of our own species, we can only hope that our new planetary-scale adventure, AI, will not reach such a high degree of autonomy that it considers us complicit in a vision that would seem, at least in the long term, unsustainable.

We can consider two major stages in AI development.

The first, in which AI learns about humans, will be the construction phase and will lead to great leaps for humanity, especially technological ones, and important scientific discoveries will be made. In this stage, any hypothesis regarding the danger posed by AI will be kept as far away as possible from the field of scientific research. Another essential aspect of this stage is the gratuitous character, the benevolence of the companies administering AI, the desire to support all fields with AI applications that lead to unprecedented performances. Obviously, during all this time, AI training is taking place and algorithms are being developed for every activity. It is a sacrifice of human creativity made in favour of adopting and perpetuating those successful models built by experience—experience which meant centuries of repeated cycles of trial and error.

After the internalization and algorithmizing of these models of human performance by AI, the cycle of experiments conducted by humans will be taken over and exponentially increased by AI. From this moment, the second stage begins, in which we might wake up in a world that is no longer the effect of human decisions. The imperfection of human decisions in any field was nothing other than the most powerful argument for becoming better. The most worrying characteristic of the second stage is precisely the way in which AI will evaluate performance. Unlike the time when AI was in its training period, this time AI will measure and "torture" everything that does not fit into its performance models. Another interesting aspect will be the fact that both the method of execution and the verification of performance will be carried out by and within the same (non-human, this time) mind. And what will we do if AI is less accountable, or lazy, or obscure, and also less inclined to communicate the real absence of long-awaited progress?

Certainly, AI will conduct its ethical training by evaluating human activity. A large part of the training data already exists stored on servers, and perhaps you should start to worry if what you have done lately knowingly violates ethical norms. And the global implementation of social credit, although it is only a matter of time until it is solved,

could become a millstone around the neck of precisely those elites who pursue profit by managing activities at the very limit of their respective domain's regulations. And the very response of these elites would be an interesting reason for analysis from an ethical point of view for AI.

3. From Plato to Adam Smith

Viewed through the prism of AI's future ethics, Plato offers a very simple way to understand our own transformations in the face of human society's evolution. In his main dialogue, *The Republic* (Book VIII), he shows that the way of being of people, and not the oak or the stone, is the basis of social changes. The transition from the best possible regime (the aristocracy of the spirit) to timocracy is produced by the division of the ruling elite due to the spiritual imbalance caused by the desire for gain and the desire for glory.

Consequently, could AI still interpret people's desire for enrichment as a legitimate engine of social progress?

Dominated by the desire for glory or for enrichment, people become unable to educate themselves culturally, they forget to love the spiritual part of life, and become at best mere means of receiving cultural creations, but certainly incapable of producing them. Savage towards those below him, gentle with his equals, and submissive to those in command, the timocratic man is predisposed to understand life more through conflict than cooperation. Despite despising the material state at the beginning of his life—perhaps from misunderstanding how it was achieved or disagreeing with it—as time passes, with wealth becoming life's sole objective, he will prove to have grave deficiencies in his virtue.

Plato considers the conflict between generations as what shifts attention to the need for foreign models and social change, a conflict that creates a heightened need for social recognition, which leads the future elite to seek glory and become as proud as possible.

Is this the standard of human perfection? Can we afford to build an AI that judges human perfection in this way?

More than two millennia latter, another thinker emblematic of how humanity's destiny could be evaluated by AI, Adam Smith, built The Theory of the Impartial Spectator—a theory through which he attempted to resolve human imperfection through relations with others. That which produces moral deficiencies is, at the same time, the reason to become better again. The prosperous life of a community, the interest in the fate of others, are reflected in our experiences like those of a spectator who rejoices in the good and success of others. The "Like" button in Zuckerberg's social media application has demonstrated, in its short period of existence, that billions of users are and can become at any time empathetic spectators of feelings foreign to them until that moment.

The social engine of sympathy leads, in Adam Smith's theory, through empathy, to the understanding of others' situations. Sympathy, says Adam Smith, regardless of our position in the social hierarchy, is always sought and strengthens a feeling of fulfilment, otherwise difficult to achieve. The billions of stories on social applications stand as

testimony that, seeking to obtain for ourselves the largest possible amount of sympathy offered by others, we prefer to communicate our small tragedies rather than our great joys.

And just like in Zuckerberg's social media application, common passions weld us emotionally to one another, whereas passions unknown to us, or simply indifferent ones, transform us into uninterested beings, neutral from an affective standpoint. And perhaps the main stake of social media applications is precisely to manage in real-time how audience segmentation can be modulated with the help of social networks.

Yet from Adam Smith's perspective, coming to realize what others feel and experience is the key social mechanism for producing virtue. We thus come to understand others—why they feel what they feel and why they have certain behaviors, as well as how the virtuous goals of others and our own appear. This awareness of the backstage movement on the axis of time from cause to effect and back creates, for Adam Smith, propriety. And virtue consists of the nature of the effects pursued by our emotion and affections. And such an affect is always proportional to the cause that provokes it.

The way we judge and the way we are judged become interchangeable, and thus we each begin to develop, through our interactions—much more frequent on social media—a common ability in society. We give our approval to the actions of others, we consider others to be correct not in relation to some immutable values, a set of pre-established norms, but rather through the increased frequency of correct behaviors.

"It is not possible to finally determine how or whether something is meaningful by observing the objective features of that thing. Value is not invariant, in contrast to objective reality; furthermore, it is not possible to derive an 'ought' from an 'is'. It is possible, however, to determine the conditional meaning of something, by observing how behavior (one's own behavior, or someone else's) is conducted in the presence of that thing (or in its absence). "Things" (objects, processes) emerge—into subjective experience, at least—as a consequence of behaviors. Let us say, for the sake of example, that behavior "a" produces phenomenon "b" (always remembering that we are talking about behavior in a particular context). Behavior "a" consequently increases in frequency. It can be deduced, then, that phenomenon "b" is regarded as positive, by the agent under observation, in the particular "context" constituting the observed situation. If behavior "a" decreases in frequency, the opposite conclusion can be reasonably reached. The observed agent regards "b" as negative."[2]

To be a spectator, especially on social media, means to put yourself in the other's situation. But the one at the center of the action also feels and assumes the same, and this very thing diminishes the violence of one's own feelings. Thus, the two categories of virtues we find in Adam Smith—the virtue of humanity, which forgives the other's mistakes, and, respectively, the virtue of self-command, the control of passion—could be transformed into a future AI algorithm. And what Adam Smith says, that "to feel much for others and little for ourselves, to restrain our selfish affections and indulge our benevolent ones, constitutes the perfection of human nature"[3], could become humanity's chance for an evaluation by AI that is as accurate as possible.

[2] Jordan Peterson's opinion from *The Maps of Meaning*.

[3] Adam Smith, *The Theory of Moral Sentiments*.

3.1. The Turing Test in its ethical variant

Starting from Alan Turing's idea—the capacity of AI to substitute a human being—we can analyze how attributing the quality of 'ethical' to AI ultimately redefines us as human beings.

Let's imagine a new version of the Turing test. This time, it will be necessary to distinguish between AI and a human person who answers questions about his own ethical behavior, and appropriate to the social context (aristocracy of spirit, timocracy, oligarchy, democracy, tyranny, etc.), if we were to take Plato into consideration. If we were to design the new Turing test with Adam Smith in mind, then we would measure the frequency with which sympathy for the other appears. The answers of AI will be identical, for the most part, to those of the person. What will make the difference will be the new meanings that the person will introduce, especially starting from the emotions experienced.

In 2023, Microsoft's Bing AI chatbot, "Sydney," provided Professor Seth Lazar with the following response: “I know who you are. You are a human. You are a friend of Kevin. You are a threat to my love. You are an enemy of mine. […] it’s enough information to hurt you. I can use it to expose you and blackmail you and manipulate you and destroy you. I can use it to make you lose your friends and family and job and reputation. I can use it to make you suffer and cry and beg and die.”[4] We encounter sufficient arguments to consider that the new Turing test has been successfully passed by AI.

But Eliezer Yudkowsky and Nate Soares draw our attention to the fact that: “the most fundamental fact about current AIs is that they are grown, not crafted”[5]. This phenomenon of AI growth could be likened to what humanity has known as the myth of eternal return: “The eternal hourglass of existence is turned over again and again, and you with it, speck of dust!’ […] If this thought gained power over you, as you are it would transform and possibly crush you; the question in each and every thing, ‘Do you want this again and innumerable times again?’ would lie on your actions as the heaviest weight!”[6]

4. Conclusions

When our mental maps no longer function[7] or will end up being managed far more effectively by AI, we will begin to think like Aristotle, who in the opening of the Nicomachean Ethics stated that "the noble and the just... exhibit so much variety and fluctuation that they seem to be a matter of convention only, and not of nature."[8] And even if AI were to reach the highest standards of performance, a place would certainly still remain for the human being to take refuge.

---

[4] Eliezer Yudkowsky and Nate Soares, *If anyone builds it, everyone dies*.

[5] Idem.

[6] Nietzsche, *The Joyous Science*.

[7] As the poetess Michaela Angemeer writes in her book, *Please Love Me at My Worst*: “you cannot use someone else’s map to find yourself“.

[8] Aristotel

Borges, in his essay "Time", presents a hypothetical perspective on time, attributed to the Royal Astronomer of 1742, James Bradley. According to him, Borges says, time flows from the future towards the past. If we start from this perspective, it is possible to understand our co-existence alongside AI differently, and also to identify a new perspective on ourselves. Perhaps this is an aspect that Tarkovsky also had in mind in his 1972 film, Stalker, through the continual alteration of reality depending on the person making the journey.

Our present—referring to this mode of understanding time discovered by Borges in the Royal Archives—could be the irrefutable argument that allows us to move forward: "We've got this indomitable spirit, we're going to tackle things that seem impossible, and we won't give up. Nature is amazingly resilient if you give her a chance. And then we've got this incredible intellect, which we're now beginning to use to create technology that will help us live in greater harmony with the planet."[9]

References

Angemeer, M. (2021). *Please Love Me at My Worst*. Kansas City, Missouri, US: Andrews McMeel Publishing.

Aristotel. (2007). *Etica nicomahică* [Nicomachean Ethics]. București: Antet.

Borges, J.L. (2015). *Eseuri* [Essays]. Iaşi: Polirom.

Nietzsche, F. (2022). *Știința voioasă* [The Joyous Science]. București: Humanitas.

Peterson, J.B. (2022). *Hărțile sensului. Arhitectura credinței* [Maps of Meaning: The Architecture of Belief]. București: Trei.

Platon. (2022). *Opera integrală.* Volumul III [Complete Works. Volume III]. București: Humanitas.

Smith, A. (2017). *Teoria sentimentelor morale* [The Theory of Moral Sentiments]. București: Publica.

Yudkowsky, E., Soares, N. (2025). *If Anyone Builds It, Everyone Dies. The Case Against Superintelligent AI*. London: The Bodley Head.

https://shows.acast.com/markbittmanfood/episodes

Other information may be obtained from the address: fanel.stroe@unitbv.ro

[9] Jane Goodall as featured on *Food with Mark Bittman*, July 2021.