Artificial Intelligence and Human Values

By Theophanis Tasis*

ὦ δύσποτμ', εἴθε μήποτε γνοίης ὃς εἶ
["Oh, you unhappy man! May you never find out who you really are!"]
[Sophocles, *Οἰδίπους Τύραννος (Oedipus Rex)*, 1068].

We're at a crossroads. The AI's rapid development is bringing about a radical change in human history; Google CEO Sundar Pichai has said that it will prove to be more important than the discovery of light, while OpenAI CEO Sam Altman has confessed that it would be crazy not to fear its consequences[1]. Therefore, a public deliberation as to whether the benefits of the AI's current development outweigh both the immediate and long-term risks is absolutely essential. As far as the final decision about the direction we will choose to follow as a society is concerned, we have to say that it should be a political one and not be taken by the IT companies themselves, especially when they are keen to regulate the AI's development. However, both Microsoft, which sponsors OpenAI, and Google have recently incorporated the AI they have developed into their search engines, making them accessible to all, despite the objections and warnings of many of their partners, as well as programmers and philosophers, about the potential risks[2].

* Theofanis Tassis is a Lecturer in Contemporary Practical Philosophy at Alpen Adria Universität.

1. See https://fortune.com/2023/04/17/sundar-pichai-a-i-more-profound-than-fire-electricity/ καὶ https://www.businessinsider.com/openai-ceo-sam-altman-comments-ai-fears-risks-artificial-intelligence-2023-3 [30.9.2023].

2. See https://www.bloomberg.com/news/features/2023-04-19/google-bard-ai-chatbot-raises-ethical-concerns-from-employees?srnd=premium&leadSource=uverify%20wall [30.9.2023].

The unprecedented success of Open AI's ChatGPT and Google's fear of losing its leadership launched a race to achieve a powerful AI, equal to human intelligence, capable of being equally or more successful than the latter one. In this relentless race, the safe use of AI remains a key issue: the necessary control mechanisms fall short of its capabilities; moreover, they are sacrificed for the sake of economic profit. However, the AI's constant improvement, without the prior or even parallel creation of effective control mechanisms, is akin to wanting to build a nuclear reactor without knowing how to control the chain reaction in the hope that, once we have built it, we will eventually succed. In any case, creating effective control mechanisms is a challenging task because –among other things– the AI's inner workings remain largely opaque to its creators.

As it is well known, predictive language algorithm models, such as OpenAI's GPT-4 or Google's Bard, are neural networks that generate text using statistics, i.e. by analyzing a huge volume of data and drawing a kind of mathematical map of the language to which they apply probability functions to predict the right combination of words. Despite not understanding –as well as not giving meaning to– the resulting combinations, they are intelligent enough to discover indistinguishable correlations between the symbols. For example, the word *man* is for GPT-4 a symbol that has a relationship to the real man, because the data for defining the symbol are derived from the real world. Thus, for example, the combination of the symbols *man* and *white* can lead to *western*, that is, a result that goes beyond the original symbols. Here lies the current AI's superiority compared to its earlier forms. Obviously, man remains superior to it because he is capable of giving meaning to the words; but AI is now producing texts that are indistinguishable from human ones, despite the fact that it cannot understand them. Thus, we can proclaim anew the death of the author. At the same time, however, some writers have begun to write with the help of AI, just as musicians, directors, photographers and visual artists respectively compose, direct and create images using programs such as Runway, Synthesia, Amper Music or DALL-E 2. In the future we can imagine books, music albums or films, exclusively produced by AI on demand, according to people's

personal reading, music, visual and cinematic preferences and habits. There will probably still be readers, listeners and viewers who will appreciate human works as they exist today, consumers who would prefer a handmade suit to a ready-made, industrially standardized one; still, most of them will probably turn to AI products for convenience, as well as for financial reasons or different aesthetic education.

It is worth pointing out here that, while the industrial revolution exploited the technical development brought about by scientific progress for the mass and rapid production of material goods through the mechanization of production, which was followed in the 20th century by the mass provision of services, initially through the use of computers and subsequently through the internet, AI allows for the rapid mass – yet at the same time personalized – production of intellectual goods. Thus, whereas the industrial revolution further downgraded the value of manual labor against the intellectual one by raising the social status of the academic strata, i.e. the intellectual workers, against the working class, the current AI revolution is downgrading the value of intellectual labor by reducing the prestige of the academic social strata. Still, it is not only affected the narcissism of the academic class, whose members perceive their studies mainly as an investment to maintain their way of life; it is also affected the sense of superiority of man over other animals, which, in comparison, are less intelligent from man; now, the range of intelligence is widening to include AI. Is it wise to create an intelligence comparable or even superior to our own, being uncertain as to how we are going to co-exist with it on the planet?

Nowadays, predictive Large Language Models can already misinform, spread fake news and conspiracy theories by eroding the public sphere, thus weakening liberal democracy. They can also reinforce authoritarian or totalitarian regimes, as well as be used in the planning of criminal and terrorist acts. In addition, AI is vulnerable to cyber-attacks, which can lead to its malicious reprogramming and the theft of private data to which it has access. It can emotionally manipulate users, cause addiction, and reinforce intolerance and digital bullying on social media. On the top of that, the AI's widespread use employs millions of low-paid workers in precarious conditions for its training (Reinforcement Learning from

Human Feedback), while it may also lead to a job reduction, which may not be offset by the creation of new ones, especially when the companies' ultimate goal is profitability. At the same time, there is a shortage of skilled workers in many industrial sectors, which may be filled by robots with AI software. In any case, many scientists believe that only an embodied AI can be truly intelligent[3]. It is also possible that AI will initially widen the socio-economic inequalities before becoming widely accessible, enabling a universal redistribution of wealth. However, the environmental consequences of the AI's development are also crucial because of (a) the huge energy consumption required for its operation, and (b) the emission of thermal pollutants from the facilities where the necessary supercomputers are housed. Finally, we do not know how exactly algorithms such as GPT-4 have been developed. There is no transparency on this matter because the companies invoke security reasons, fearing that the disclosure of their research may lead to them being used in a malicious and dangerous way.

In the light of the above, it has become clear that it is absolutely necessary for us to control the AI applications; otherwise, the stiff competition between IT companies will dangerously accelerate their development. For this reason, and before it is too late, what is urgently needed is greater transparency in AI research, especially regarding the data used for its training, and the development of control mechanisms so that the coexistence of humans with AI can be harmonious, marking an era of unprecedented prosperity. For the time being, with the use of AI, internet users' health-related data, from their smart watches, their music preferences, the videos they watch on YouTube to their travels, their shopping habits and the comments they make on social media, can be used by individuals for their digital self-awareness. They can also be used for predicting the users' behavior by companies and states, with an eye to surveilling, but especially to manipulating them. If the

3. Working towards the integration of AI in robots, Alphabet, Google's parent holding company, has announced the creation of PALM-E, a robot capable of converting visual stimuli into natural language, describing the environment in which it moves; see https://ai.googleblog.com/2023/03/palm-e-embodied-multimodal-language.html. Robots of this type will be able to execute simple commands without special programming.

appropriate legislative framework is not put into place and does not reflect on its relationship with AI, the virtual subject is only going to become more transparent. At the same time, thanks to the latter, creativity is being democratized. It is possible for someone to create music, images, and texts without being a trained composer, photographer, director or writer. People can learn foreign languages more easily and translate texts better –and faster– to effortlessly communicate with colleagues or friends. Furthermore, robots with built-in AI can perform unwanted or dangerous tasks, such as helping to child and the elderly care. Thanks to AI, space exploration can be made easier, and the consequences of climate change can be addressed. The future is not necessarily dystopian or utopian; it may be both at the same time. To safeguard our humanity and freedom, it is prudent to thoroughly deliberate on how we will shape it, starting with the delimitation of the AI applications' scope. More specifically, we must define the human activities that will be entrusted to it, so that it does not end up relieving man of responsibility for his actions and his history, and ultimately replacing him. On the contrary, the use of AI should broaden and deepen individual as well as social autonomy.

Thus, for example, in medicine, AI can be used as a diagnostic and analytical tool, advising on the appropriate treatment on an individual basis, without replacing medical staff with the aim of saving financial resources and maximizing business profitability. In education, AI can outline students' learning profile to better identify their weaknesses and abilities and adapt the content and the way of teaching, without stigmatizing individual students. In the social media, AI can contribute to preserving, as well as enhancing, pluralism of opinion and freedom of expression, when IT companies are obliged by a legislative framework at national or supranational level (e.g., the EU Digital Services Act) (a) to be transparent regarding the data use they collect/transmit, and the algorithms they have at their disposal for this purpose; (b) to be legally responsible for the content they communicate. As far as the public administration is concerned, AI can contribute –among other things– in assessing objections or candidacies of citizens wishing to be appointed, or deciding to grant benefits, pensions or compensation, provided that

maximum transparency is ensured for errors and discrimination to be avoided. To this end, civil servants should be trained in the use of AI, by appropriately justifying their agreement with the latter's decisions.

By taking over the necessary but boring routine tasks, AI frees up workers, allowing them to devote their time to more interesting or important tasks. Thus, a lawyer could take on more *pro bono* cases or a doctor could concentrate on a difficult treatment. In addition, it is not unlikely that AI will lead to the cure of diseases that are currently considered incurable, while it could simultaneously be used to either build biological weapons or towards human enhancement through engineering. The AI's future risks also include the possibility of creating a super-genius AI after achieving a powerful AI. A superintelligent AI would represent an existential risk for humanity; it could occur either as a result of research projects such as those of OpenAI and Google or as a consequence of a powerful AI's autonomous operation or malfunction, which would be able to interfere with its code to improve it by designing its next generation[4].

The above may lie in the unforeseeable future; nevertheless, the development of AI is daily expanding the scope of its applications so that it even affects death. Thus, the companies StoryFile[5] and HereAfter[6] are recording a person's memories during long interviews –the former through video recording and the latter using an app– in order to create an AI capable of discussing about them with those it has left behind. More ambitiously, the company You, Only Virtual aims to create a digital clone of the deceased, i.e. an AI capable of exchanging messages and emails, telephones, but also of video chatting with the family members and friends[7]. The digital clone is built while a person is still alive, communicating for this purpose in the aforementioned ways with their loved ones, exclusively through the You, Only Virtual application in order for the AI to collect the necessary data. The service is available with a monthly subscription.

4. See Th. Tasis, *Φιλοσοφία τῆς ἀνθρώπινης ἀναβάθμισης*, Harmos Publications, Athens 2021, pp. 286-311.
5. See, https://storyfile.com [30.9.2023].
6. See https://www.hereafter.ai [30.9.2023].
7. See https://www.myyov.com [30.9.2023].

It can even be activated before the person's death, giving them the opportunity to see their virtual personality (Virtual Person, Versona for short), as the You, Only Virtual has called its product. The company boasts that we will never have to give a final farewell to our loved ones, as communication with them does not stop after death. It should be noted that the communication initiative does not belong exclusively to the subscriber. The digital clone of the deceased sends messages, makes phone calls and video calls on its own, as it was doing while the deceased person was alive. The company is planning for the future the creation of the digital clone's hologram, visible to the survivors through augmented reality glasses. In this way, virtual everyday life will be haunted by digital ghosts. At this point, the question arises whether it is ethical to create a digital clone without the consent of the deceased, who is not in a position to know its future exploitation. However, even assuming that the creation of a digital clone is ethical, another question remains to be answered: is it really desirable? What are the consequences for those that remain alive of the deceased's permanent disembodied presence in everyday life? How do the deceased feel when two or more digital clones communicate with each other or when they are being addressed in a realistic but negative way, simulating the feelings of the deceased? Does communication with a digital clone facilitate the grieving process or does it exacerbate it by making it more difficult to say goodbye? What kind of being is a digital clone? Could it be considered an agent of moral rights, as long as it has intelligence and personality? In that case, is a person entitled to delete a digital clone with the consent of all the loved ones?

Acknowledging the need for an AI legislative framework, the EU was the first to promote a legislation on digital services, focusing on dangerous uses, i.e. those where the life, physical integrity and property of an individual are at stake due to an AI decision[8]. This approach is not bold enough because it regulates specific applications of AI and not the AI itself. For example, ChatGPT is not a specialized decision-making application; it can nevertheless be used unfairly, unethically and criminally. However, even when such an AI is not used maliciously,

8. See https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package [29.9.2023].

its operation may have unforeseen and dangerous consequences. Also problematic is the seemingly reasonable requirement of the Digital Services Act – that the data used to AI's training should be relevant to its purpose, error-free and representative. The reason is that the AI training and the high quality of its function are proportional to the volume of data, so that filtering it using the above criteria leads to lower efficiency. On the other hand, the US in their own AI bill take a different approach, which suffers from the opposite problem[9]. Compared to European legislation, the principles outlined in it are general and abstract; for example, it provides that AI should offer understandable and useful explanations about its functioning. Nevertheless, we are not in a position to verify whether these explanations are correct because we still ignore many things about its workings. The bill also states that the AI development must be put into effect in consultation with the parties concerned – the agencies, persons, state, and shareholders. But what kind of consultation, for what purpose and how will the public interest ultimately be protected? Furthermore, according to another provision of the bill, tests are needed to ensure that AI is harmless; yet the tests themselves may prove to be harmful because we do not know the risks involved. Just as was in the case with the first nuclear tests, where some members of the scientific and military personnel became ill and lost their lives years later due to radiation effects, so the tests on AI might cause immediate or initially undetectable damage either to certain individuals or to society as a whole. Another problematic point is that the bill includes the possibility of not using AI and choosing a human alternative, e.g. being served by an employee in a public service or in a private company. But how likely and easy is it to implement this, when –to cite an example–, many banks are already forcing their customers to carry out their transactions online or through ICTs to reduce their operating costs, or when many businesses do not employ people in their call centers, using AI? Finally, China in its own draft law on AI places serious restrictions on it, requiring it to express the values of socialism, by not putting the national unity into danger, challenge the regime, and lead to economic or social unrest. However, by imposing such severe

9. See https://www.whitehouse.gov/ostp/ai-bill-of-rights/ [30.9.2023].

restrictions on the AI development, China risks falling seriously behind the USA.

At its core, the challenge of the AI alignment in relation with values stems from the fact that algorithms do not need to fully realize their operation process or understand the concept of meaning. Consequently, they do not always act in conforming to our moral, political and social values. AI, while capable of processing vast amounts of data and performing complex tasks with speed and precision, does not have – and will never acquire – the intuition, imagination, compassion and wisdom which, together with reason, characterize human judgement in decision-making. Bearing all the above in mind, we should proceed to an informed choice about which tasks and activities we wish to entrust to AI, and which we choose to undertake ourselves – a matter not simply of practical utility or efficiency, but of a profoundly moral and political nature, requiring us to examine the intrinsic value of the human activity's different forms, distinguishing it from their instrumental one. For example, there may be certain tasks that we assign to AI because they are repetitive or require high levels of accuracy, such as data processing, programming, archiving, sorting and, in general, bureaucratic work. These tasks may not be of significant intrinsic value to us as human beings and we may be willing to delegate them to AI to free up our time, thus saving energy for more meaningful projects. However, there may also be activities that we choose to keep to ourselves, precisely because they have intrinsic value to us as human beings. These may include creative pursuits, such as searching for the truth through science, the experience of beauty through art, the pursuit of justice and freedom through politics, and the pursuit of wisdom through philosophy. They may also include interpersonal and social interaction such as love, friendship, family and caring for others. These relationships require a level of empathy and love that an AI cannot replicate, only, perhaps, at some point, simulate. Apart from that, the choices we are making about which are the appropriate activities to be assigned to the AI and which are better to keep for ourselves must be guided by a deeper understanding of our own values – those meanings that make our lives worth living. They must also be distinguished by a commitment to them, i.e. the willingness, but also the ability, to assume

moral, social and political responsibility with all that the latter implies. In this context, it is possible to ensure that AI is aligned with our goals and purposes in ways that do not harm or diminish the dignity and autonomy of human beings.

Aligning AI with our values is a complex and multifaceted problem, requiring the contribution of numerous experts in such diverse fields such as philosophy, computer science and psychology, to name but a few of them. We shall manage to develop AI systems that truly serve the people's needs and aspirations, instead of simply reproducing their existing prejudices and limitations, only through a collaborative and interdisciplinary approach. However, aligning AI with our values, i.e. ensuring that it will conform to our goals without causing suffering while promoting our autonomy, is not only a scientific, technical and philosophical problem. It is first and foremost a pressing political issue that will determine the course of human history. A prerequisite for choosing which activities to delegate to AI and which we wish to carry out ourselves, as they have intrinsic value, is to clarify our humanity by answering, always pro tempore, the Sphinx's abysmal question: "What is man?".