sciendo

Artificial Intelligence, Value Alignment and Rationality

Zhumagul Bekenova

Department of Philosophy and Political Science
Al-Farabi Kazakh National University
Al-Farabi 71
Almaty 050040, Kazakhstan
Email: Zhumagul.Bekenova@kaznu.edu.kz

Peeter Müürsepp

Department of Philosophy and Political Science
Al-Farabi Kazakh National University
Department of Law
Tallinn University of Technology
Akadeemia tee 3
Tallinn 12618, Estonia
Email: peeter.muursepp@taltech.ee

Gulzhikhan Nurysheva
Laura Turarbekova

Department of Philosophy and Political Science
Al-Farabi Kazakh National University
Al-Farabi 71
Almaty 050040, Kazakhstan
Email: Gulzhikhan.Nurysheva@kaznu.edu.kz
Email: Turarbekova.Laura@kaznu.kz

Abstract: The problem of value alignment in the context of AI studies is becoming more and more acute. This article deals with the basic questions concerning the system of human values corresponding to what we would like digital minds to be capable of. It has been suggested that as long as humans cannot agree on a universal system of values in the positive sense, we might be able to agree on what has to be avoided. The article argues that while we may follow this suggestion, we still need to keep the positive approach in focus as well. A holistic solution to the value alignment problem is not in sight and there might possibly never be a final

doi: 10.2478/bjes-2022-0004 TalTech Journal of European Studies
Tallinn University of Technology (ISSN 2674-4619), Vol. 12, No. 1 (35)

solution. Currently, we are facing an era of endless adjustment of digital minds to biological ones. The biggest challenge is to keep humans in control of this adjustment. Here the responsibility lies with the humans. Human minds might not be able to fix the capacity of digital minds. The philosophical analysis shows that the key concept when dealing with this issue is value plurality. It may well be that we have to redefine our understanding of rationality in order to successfully deal with the value alignment problem. The article discusses an option to elaborate on the traditional understanding of rationality in the context of AI studies.

Keywords: *artificial intelligence, rationality, responsibility, value, value alignment*

1. Introduction

We tend to understand artificial intelligence (AI) as a man-made replica of our own human intelligence, which is logical, in a way, as we do not know any other type of intelligence. There is a good reason to create and develop AI only if this is to the benefit of mankind. Still, there is an interesting challenge here. Our own biological human intelligence has by far not necessarily been to the benefit of mankind. Science, the most ingenious creation of the human mind, has made immense progress. Multiple examples of this success story can be given. Let us just point out the achievements of medical science in finding efficient cure for numerous infectious diseases that killed humans in large numbers until quite recently and the very rapid development of relatively efficient vaccines against COVID-19. On the other hand, however, science has created or contributed to causing a number of global problems that are not just a menace but even present a real existential threat to mankind, such as weapons of mass destruction or climate change (Maxwell, 2021). Given the latter, we should by no means be looking for an AI that is too similar to human intelligence.

The positive achievements and failures of human intelligence make the value alignment problem especially important, but also complex. Theoretically, a good solution to the problem is self-evident. We need to aim at an AI that enhances the positive aspects and achievements of the human mind and prevents the negative ones from happening. But how to move in that direction, let alone how to get there? There is a diversity of human cultures

and they do not agree on the value systems they adhere to. There are some very basic values, however, such as avoiding killing an individual of your own culture, but the majority of values accepted as important in different cultures can vary dramatically. A basic division that has to be accounted for is the one between prioritising individual or collective type of rights and liberties. Given this cultural diversity, there is an idea to focus on a preventive type of approach in dealing with value alignment (Sutrop, 2020). In other words, we might know better what we do not want to happen to us rather than what we want to happen. Still, this may not work significantly better than the positive approach. We may be caught here in the classical Christian dichotomy of good and evil. No good can be done without evil being there in the world. Without the latter we would not be able to recognize good. There will just be something happening without any ethical flavour. Considering this context, we may end up with the recognition that digital minds should just be copies of biological ones, that AI should just mimic human intelligence, because the latter is able to master the ethical issues in the best possible way. However, in that case, why would we need AI in the first place. Therefore, AI should, to a certain extent, prevail over human intelligence. At the same time, it must not, in any way, become a threat to humans. Prevention of machines taking over is a task of utmost importance (Nyman-Metcalf & Kerikmäe, 2021). We, humans, have the responsibility to not create anything that would be of great, even existential, threat to the whole civilization. In the end, we, humans, still carry the responsibility for having let the genie out of the bottle (Bryson, 2018, p. 21).

We need to ask ourselves, is there an option to stop developing AI, as we are not able to guarantee that no harm will be done? Probably not. There is no way to stop technological progress. Even if a unanimous international political decision is made, someone somewhere would still continue to develop technology (Bostrom, 2019). We have gone too far already. Now the task is to take the responsibility for remaining in control.

We can see that there is a real bundle of problems as far as AI is concerned. We shall focus on the issue of value alignment, the problem of identifying which human values should AI be aligned with, putting forward a new perspective in the context of rationality. However, for getting started, there has to be a common understanding about what AI is. No one single definition exists but there is the option to choose the approach given our aims and needs.

As we shall be dealing with minds, we cannot limit the approach to values only. Humans are rational beings as well as emotional ones. While developing

AI we tend to concentrate on rationality only, leaving emotions out of the picture. At some point, we come face to face with the value alignment problem. A purely rational mind in the traditional sense should not have anything to do with values. However, as the human mind is a combination of rationality and emotionality, this is how a mind we know so far actually works. No human is a purely rational agent. Therefore, if AI will remain purely rational then we would have no chance to resolve the value alignment problem. The problem will still be there, but no option for a solution. AI has to be rational but in a non-traditional way, just like the human mind actually is. Thus, it looks most promising that in order to move forward with the value alignment problem, we need to take a brief look at what a non-traditional rationality might be like. We shall do that in due course.

Our hypothesis is that we have to assess the value alignment problem from the position of a non-traditional approach to rationality that involves the emotional side. The most promising approach from the value perspective is value pluralism, but also in a definite sense that we will explain below.

Let us introduce here our approach on how to understand artificial intelligence. Otherwise, it would be very difficult to assess whether AI can be a rational agent in some more sophisticated way than the traditional one. Then we will introduce the non-traditional way of understanding rationality. Before analysing the problem of value alignment in more detail, we shall also present an option of how to understand value alignment as such. In addition, we shall discuss different challenges and philosophical approaches to value alignment.

2. What is artificial intelligence?

There is an "official" definition at our disposal. The European Commission High Level Expert Group on Artificial Intelligence has put it as follows:

> AI systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also

> adapt their behaviour by analysing how the environment is affected by their previous actions (European Commission, 2019a).

The definition does not rule out a system like the Chinese room (Searle, 1980), where a complex task that seemingly requires understanding of the situation, whereas understanding is actually not necessarily present, being called AI. As AI systems are designed by humans, we have the responsibility for the actions of AI. What kind of values will AI be aligned with is the very key issue here.

The definition of the High Level Expert Group can be taken as a reference point for our analysis. The definition seemingly addresses the narrow approach to AI rather than Artificial General Intelligence. However, it can be expanded for clarifying certain aspects of AI from the perspective of value alignment on the general level. It is important to note that there is a reference to deciding the best action(s) in the definition. The ethical dimension can enter here if we understand "the best" in the normative way. If this is the case then deciding about the best actions necessarily involves ethics. As the focus in the current analysis will be on value alignment, it seems that a position has to be taken concerning the Chinese room type of situation before anything meaningful concerning value alignment can be undertaken. Why bother about aligning the values with a system that may not have any understanding of what it is doing. As we know, based on the Chinese room type of thought experiment, the observer can be misled, making her believe that there is understanding where it is actually missing. Thus, the whole situation becomes especially complicated. We would be aligning values with a system that does not have value preferences but looks as it may have them. Like the case of how the individual in the Chinese room was asked questions concerning her value preferences and answered them in way that sounded meaningful for the outside observer, whereas in fact she just put together two sheets of paper with Chinese characters written on them, the question and the answer, with the help of the manual she had been provided with, not even understanding that she had been asked a question and was looking for an answer to it.

It is often assumed that a real AI should be conscious. However, there are accounts of moral status that do not view the presence of consciousness as a necessary condition for having moral status. Let us not make a mistake here. The individual acting inside the Chinese room is conscious, of course. However, the computer explained by the thought experiment of the Chinese room need not be conscious. The major complication with that is that accounting for definite value preferences can be connected to the

Chinese room type of situation but we will not be able to figure out whether understanding and being conscious are involved or not. A discussion of AI consciousness can be found in the works of David Chalmers (e.g., Chalmers, 2010). There is also a discussion of moral status in unconscious but agential AI (Kagan, 2019). It still needs to be proven whether an approach based on animal ethics can be extended to AI. However, there definitely is similarity concerning morality from the perspective of responsibility. Creatures trained by humans or devices made by humans cannot be made fully responsible for their actions, whereas humans can.

After these brief observations we can conclude that the European Commission's definition of AI is very good as it avoids problematic concepts like consciousness or understanding. However, it includes by no means less complex concepts like knowledge, decision taking and behaviour. It looks as if the guiding idea behind the AI architecture is recreating the human mind by putting its ethical shortcomings under strict control. The case is that a strictly controlled mind is not a real mind**.** Still, AI is not meant to be a real mind if real means natural to us. It is a digital mind.

We can introduce two different views on the possible moral status of digital minds. There is the claim that humans hold a "full moral status" that cannot be surpassed. Thus, digital minds have to remain subordinate to human minds forever. However, there are others who argue that the so-called super-patient status is possible since the psychological capacities that are taken to confer human moral status allow for superhuman degrees. The term 'super-patient', introduced by Daniel Dewey, stands for a being with superhuman moral status (Shulman & Bostrom, 2020, pp. 2–3). According to this view, human minds are in the position of giving way to superhuman digital minds. Shulman and Bostrom specify about the super-patients: "If a being has interests that should be given much greater moral consideration than the interests of a human being, not because the interest is stronger but because it has higher moral status, then that being would be a super-patient in our terminology" (Shulman & Bostrom, 2020, p. 3). Thus, ideally, AI could become a super-patient and the value alignment problem will be resolved because from the human perspective it will not exist for a being that has higher moral status than humans do. We could perhaps say that we have here a deflationary solution to the value alignment problem, meaning that we need not pay any special attention to value alignment at all. Just develop the software and hardware according to our best understanding. However, before claiming any solutions or agreeing about not doing so, let us specify what we normally understand by value alignment.

3. What is value alignment?

So far, we have relied upon an intuition of what value alignment might mean. Now it is high time to specify. There is no commonly agreed upon definition proposed by an official body concerning value alignment, like we have in the case of AI. However, AI researchers have provided us with an explanation: "As intelligent systems gain autonomy and capability, it becomes vital to ensure that their objectives match those of their human users; this is known as the value alignment problem" (Fisac *et al.*, 2018, p. 1). Others specify that value alignment can be defined as a property of intelligent agents that allows them only to pursue goals that are beneficial to humans (Soares & Fallenstein, 2014; Soares, 2015; Russell *et al.*, 2016; Arnold *et al.*, 2017; Russell, 2019). The task is a real tough one as we can hardly claim that we know which goals are only beneficial to humans. To get an understanding of this would mean permanently ongoing reassessment. We cannot expect to be able to figure out any finite list of goals, not to speak of values or actions that would resolve the issue. Ideally, a kind of rationality could be implanted into AI that is capable of the same type of ongoing reassessment as the human mind is. However, the rationality should be better than the human one. Otherwise, again, we have to ask ourselves what is the purpose of AI if it is not better, i.e. more rational, than a human being, and could even constitute a potential threat to humans. As we shall see below, it is precisely rationality that is the key concept here. We hardly need an AI that is more irrational, not to speak of being more emotional than humans are. Curiously, we need to make up our mind of what kind of rationality we need in order to solve the value alignment problem with any success whatsoever. This may look like a pseudoproblem. Do we not know what rationality is? Definitely we do have at least an intuitive understanding of the concept. It appears, however, that quite different approaches to rationality are possible. It is obvious that at this point we are not seeking to create a digital mind that can calculate faster and with higher precision than a biological mind. This issue has been resolved long ago. The rationality we are looking for is much more sophisticated.

4. What kind of rationality?

The question in the title of this chapter may seem out of place. We normally think that rationality means strictly controlled reasoning based on the basic rules of logic and mathematics, validity of arguments being the central feature (Maxwell, 2021). It appears, however, that rationality is a much more complex intellectual capacity. There are at least two ways of being extreme about rationality. Let us not forget that being extreme, being radical, actually means reaching for the roots of the phenomenon. We do not mean the historical sources here but the understanding of the concept. The first type of extreme rationality would be a kind of overregulation, creating order in the most trivial sense of the term. This kind of situation in society has been described well by Yevgeny Zamyatin (1993). The early twentieth-century Russian writer described a rational society as follows: People have numbers for names and reason dominates life. For entertainment, people link arms and march down the streets, singing patriotic songs. Obviously, this is not a way of life most of the people, if anyone at all, would enjoy. On the other hand, no direct harm would come to anyone in such type of a rational society. It is unlikely, however, that we would want to develop an AI that would regard such an arrangement of social life as an ideal and act on the principle of moving towards a Zamyatin-type rational society, although such an AI would be not very difficult to programme and the value alignment problem could be relatively easily solvable.

The other, if not extreme then certainly unorthodox, understanding of rationality has been presented by Nicholas Maxwell. For Maxwell, this is the correct approach to rationality and he has given reasons for us to believe so. Maxwell starts formulating his approach with a reasonable explanation: "Reason does not tell us what to think and do, it helps us discover what we want to do, what is in our best interests to do. It puts us in a better position to discover what we do really want to do" (Maxwell, 2021, p. 26). There is obviously a "learning by doing" element in this observation. What follows, however, raises questions. Maxwell insists that emotion and desire are not anti-rational or non-rational but they are vital ingredients of rationality (Maxwell, 2021, p. 26). This could perhaps be accepted if the only rationality we know remains the one that is carried by a biological mind. In a way, emotions and desires are ingredients of human rationality because there is no mentally healthy human being without these qualities, therefore no rational human being either. Normally, emotions and desires have been considered to be human qualities that prevent making purely rational

decisions whenever we need them. This view was already strongly present in the philosophy of ancient Stoics. But the Stoics recognized the presence of desires in human life as well. Should we be looking for an emotional AI with desires or a purely rational one in the traditional sense? Somehow, the latter looks safer from the human perspective. Or do we have to try to make a distinction between rational and irrational emotions and desires? The latter must exist because otherwise we would not have any irrationality in the world whatsoever. That would mean that we would not have any rationality either, just neutral human mental capacities.

Maxwell elaborates on his understanding of rationality, explaining that reason becomes relevant when we get into difficulties and face problems we do nt know how to solve (Maxwell, 2021, p. 27). We can hardly disagree with this. However, Maxwell continues:

> There is nothing irrational in acting instinctively, spontaneously, without any deliberation at all if, as often is the case, what we are doing faces no problem, no difficulty. Quite the contrary, it is the person who cannot act instinctively, spontaneously, but must always deliberate before action, who is irrational. (Maxwell, 2021, pp. 27–28)

This presents the observer with several puzzling questions. It can hardly be that any kind of instinctive and spontaneous behaviour should be taken as rational. Besides, not any kind of deliberation needs to be irrational. Thus, we find ourselves in the situation of having to learn to distinguish between rational and irrational instinctive, spontaneous and deliberate behaviour. Following Maxwell, there is also the controversial issue of facing or not facing a problem. If we feel we want to act then something has the corresponding impact on us. Thus, in a way, we are always facing a problem or at least, there is always an issue there that has to be resolved.

Given the above, in addition to traditional rationality we have the simplistic extreme one (Zamyatin) and the "irrational" extreme one (Maxwell), the one including emotions and desires. What kind of a rational agent would we prefer AI to be? We suspect that the majority of us would vote for the traditional option. However, a traditionally rational agent in its pure form would not have subjective value preferences. If we are interested in a human-like AI then we must let in some Maxwell-type rationality, i.e. some irrationality. Is that somehow feasible at all? For instance, what kind of rationality would we prefer to rely on in the case of the data protection problem (Kesa & Kerikmäe, 2020)? Again, the traditional one, we suspect.

Given the complexity of the human being, it may be that by combining rationality with irrationality we shall still miss the point, i.e. we come face to face with the missing something argument (Kerikmäe *et al*., 2020, pp. 83–84). Some would say that it is the soul that we are missing. If so, its creation is beyond human powers. The situation is still not hopeless, however, because we do not really know what is the basis of humans having a soul. Is it something of divine origin or perhaps rather a metaphoric expression of some mental phenomena that develop on their own along the lines of a kind of self-organization. In the latter case we have reason to believe that if the mental structure of digital minds would become close enough to the human mind then the soul-like general capacity will appear in digital minds sooner or later.

Let us look into history for help. There were two powerful cultural movements in the 18th and 19th centuries, one of which is normally linked to rationality and the other to irrationality. We are referring to the Enlightenment and Romanticism, of course. Enlightenment is the model era for rationality and science in the traditional sense. Romanticism is associated in our minds with emotions and the arts. In Nicholas Maxwell's approach, they have to be combined in order to achieve real rationality. Traditional rationality in its pure form is rather irrational for Maxwell because it fails to make full use of the mental capacities of human beings.

In the human culture, historical development plays a significant, sometimes decisive, role. The values we share have formed gradually, passing several stages of cultural history in the broadest sense. Westerners have not always worshipped individual freedom in the style of liberal democracy, as in the Middle Ages. Thus, we should ask ourselves the question whether we should somehow try to programme this historical past into digital minds in order to facilitate value alignment. Again, if we are looking for AI that is as similar to biological minds as possible, then the answer to this question would be a straightforward 'yes'. Perhaps, as we have observed above, AI need not be similar to humans in each and every respect. If so, then AI may not need a memory of the historical past that it does not have anyway. This looks like a relief to the human developers of AI, but on the other hand, making AI rational in the human sense may be even more complicated if the historical past is not available. To put this in the terms of Nick Bostrom (2003), do we need to run the simulation of the historical past for the civilisation of the digital minds in order to facilitate value alignment? We do not know for sure but that is definitely a question to consider.

The biological human mind includes both sides, the traditionally rational Enlightenment one and the emotional Romanticist one. Some say, Nicholas Maxwell among them, that ideally we should be looking for an interplay between the mind and the heart. To this day, AI research has been focusing on the mind only and we still do not know how to bring the heart into the picture. Perhaps aiming at heart and soul would be hopeless in the case of human artificial creation and should be left to nature (or God) to implement. However, humans should not hide away from something they are clearly involved with. Responsibility has to be taken. Let us still put up the challenge here. By doing so, we condemn ourselves to not being able to solve the value alignment problem perhaps for quite some time. Still, we can develop our understanding of what we need in order to move towards more and more ethical AI.

5. What kind of value alignment?

The value alignment problem faces two kinds of challenges, the technical and the normative one. Obviously, they are interrelated (Gabriel, 2020). The technical challenge deals with the question of how to encode human values in artificial intelligence. The normative challenge is associated with two questions: which values and whose values should AI align with? We cannot solve normative problems if we do not have the technological capacity. Nevertheless, our focus in the current treatise is the normative side of the issue. Given this, we shall assume that there are no technological limits and we are left just with the normative side of the problem. At once, we have a principal issue to resolve. If there are no technological limits whatsoever then we are capable of producing exact copies of human mind, including the hardware (body). However, this effort would be worthless. Why build something that does exist anyway. We need minds that are at least some way better than human ones. The solution seems easy, even trivial. Let us copy the human mind (the software) and put it into a more endurable and easier to repair mechanical hardware. However, a human type of intelligence does not need to function in the same way in a mechanical hardware as in the human body and we do not know what the differences might be like. As AI research has revealed so far, even if we manage to create an identical copy of a human being, a robot having a digital mind with decision capacity and responsibility, we still will not get a human being, a creature that is alive. As we saw above, something will still be missing. Even if the digital mind

functions in the same way as the human one, this is not what we need, as humans tend to cause harm to each other and, in fact, to any other creatures found on the face of the earth as well. This is an important argument to account for but as in reality we are not technologically omnipotent, at this point we should still keep the focus on the normative side of value alignment. It seems that, in a way, the missing something argument could be taken as a guiding light in looking for the solution of the normative challenge of the value alignment problem. It may be that we need not look for this missing something but rather forget about it. This way, we will obtain artificial intelligence that is definitely different from human intelligence and not necessarily worse in each and every respect. Perhaps it is just this missing something that allows humans to cause harm to each other.

The problem of value alignment would be easy to resolve if we knew what are the most important universal values that human beings share. Unfortunately, we can hardly ever find out for sure. We do not even know whether the universal values exist. There is the option to rely on the Christian Ten Commandments, except for the first three of them. The remaining seven principles may be fit for a foundation of the universal values, though definitely not in their traditional historical formulation. For instance, these commandments include the rhetoric and vocabulary that corresponds to the traditional model of the Christian family that is not shared across all human cultures and not necessarily even in one culture.

Somewhat fortunately, we need not look back into the cultural-historical tradition as far as the Bible, which many people do not consider practically applicable in the contemporary context. There never seems to be any need to revisit the historical tradition at all. On the other hand, human culture is a continuously developing system and we will hardly manage to enforce an interruption to this development unless the interruption is a fatally terminal one. Here, however, let us rely on contemporary considerations. We have the *Ethics Guidelines for Trustworthy AI* (European Commission, 2019b) at our disposal. The document asserts that "AI systems should respect the plurality of values and choices of individuals", which means that "certain fundamental rights and principles, such as human dignity, are absolute and cannot be subject to a balancing exercise" (European Commission, 2019b, p. 13). We see that the EU document declares that AI systems must function on the basis of freedom of choices of individuals. Thus, AI is committed to follow the Western tradition of values. On the other hand, the document declares human dignity a fundamental principle and presumes there can be other such kind of absolute rights and principles. On a level of abstract category,

dignity is really a universal value. However, in living and developing cultures even such general concept as dignity can be understood in different ways. There is more that coheres with the Western cultural tradition in the EU document, stating that AI has to be "aligned with our foundational values of respect for human rights, democracy, and rule of law" (European Commission, 2019b, p. 4).

The document we are referring to is a European Union document. Therefore, it is understandable, in a way, that the ethics guidelines given in it for trustworthy AI have to be based on what we call today European values. We could also say that they are the basic values of democracy. On this level of value assessment there is hardly any difference between European and North American values and by all evidence, by and large, we can add all democracies into this community. This move, in a way, resolves the issue of value plurality. However, are we committed to strive for a European AI or a democratic AI? Does that mean that the value alignment problem can have more than one solution? Are we moving towards building a democratic AI and in addition, non-democratic AIs of even different kinds? This seems strange. Or should we take the position that AI can be aligned or even applied in democratic societies only, or even prevented from being applied in nondemocratic societies? Again, a weird idea and, by all evidence, also not applicable. How could we imagine limiting AI to the democratic world?

However, the problems just stated are complementary for us. The focus of the analysis, as noted above, is the question whether the preventive approach towards value alignment can be more successful than the promotive one. Obviously, the issue of different kind of values in different cultures makes our task more complicated and perhaps not solvable at all. There is an attempt to set the score presented in an IBM manual: "Care is required to ensure sensitivity to a wide range of cultural norms and values. As daunting as it may seem to take value systems into account, the common core of universal principles is that they are a cooperative phenomenon" (IBM, 2019). It seems overly naïve to believe that recognizing the cooperative phenomenon in the common core of universal principles is enough to deal efficiently with the plurality of value systems but it points out something really important, perhaps even in most general terms, concerning AI studies. As human civilisation is a cooperative phenomenon, the community of digital minds should be the same by all evidence.

There is still another important document that has to be referred to—*Ethically Aligned Design. A Vision for Prioritizing Human Well-being*

with Autonomous and Intelligent Systems. It has been published under the auspices of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. At first, it seems that this particular document accounts for a different cultural tradition in the context of AI development. However, this is a premature impression. Let us just consider the next formulation:

> whether our ethical practices are Western (e.g., Aristotelian, Kantian), Eastern (e.g., Shinto, School of Mo, Confucian), African (e.g., Ubuntu), or from another tradition, honoring holistic definitions of societal prosperity is essential versus pursuing one-dimensional goals of increased productivity or gross domestic product (GDP). Autonomous and intelligent systems should prioritize and have as their goal the explicit honoring of our inalienable fundamental rights and dignity as well as the increase of human flourishing and environmental sustainability. (IEEE, 2019, p. 1)

There still is a step forward in the IEEE document compared to the EU and the IBM ones. The IEEE document does not prioritise traditional values, such as individual freedom, but rather the major existential challenge of today—environmental sustainability. The problem with this type of value is that it is on a different level compared to the values that concern human beings directly, such as their individual wellbeing. This may complicate alignment with such kind of indirect values. However, at some point, developing AI has to reach the "global" values like environment protection as well. There is a strong priority issue here. For a digital mind, sustainability of the planet may be far more important than the wellbeing of any individual human being. But, after all, humans may come face to face with this dilemma as well.

Another general concept put forward in the IEEE document is human flourishing. This is a really abstract concept even in comparison with environmental protection. If the latter can be quite well quantified then the former is a philosophical-ethical conception. The idea of human flourishing can be understood quite differently in different cultures and there can hardly be anything analogous in a community of digital minds. For the latter, flourishing would just be a technical term.

Needless to say, speaking about the inalienable human rights is problematic as well. It is clearly an absolutely Western concept. The good side of the concept is that in the Western tradition it has been quite well explicated. Things may be somewhat different with human dignity but even here there

are strong cultural influences. Worst of all, we have reasons to doubt whether all human cultures care about environmental sustainability. Can we confirm that even in the case of the Western political-economic system? The need for care is being declared strongly, of course, but hardly enough has been done so far. But still, if a digital mind has to be made capable of caring about anything at all then environmental issues have to take the centre stage. After all, not harming the environment means not harming human beings in the most significant way.

In most general terms, what we can find in all the documents cited above is a kind of double emphasis that seems to cause difficulties at the core. Everywhere there is an emphasis on the plurality of values, but at the same time certain specific values are prioritised. However, pluralism is not the same as relativism and it is not the same as subjectivism. Relativism means that rightness depends on the culture or society. Subjectivism means that rightness depends on the individual. Pluralism holds that sometimes rightness cannot be determined at all and thus there is irreducibility of values. There are also claims that pluralism can be compatible with objectivism. Isaiah Berlin, for example, has supported this view. According to Berlin (2013[1947]), pluralism does not exclude a world of objective values or ends that men pursue for their own sake. Ends and moral principles are plural, but they are not infinitely many, since they have to remain "within the human horizon" (Berlin, 2013[1947], p. 12). So, we need to keep AI within the "human horizon". On the other hand, we are probably tempted to go beyond the human horizon with the digital minds in order to make them better ethical agents than humans. Based on Berlin's views, the point is that we should be able to distinguish between what is objectively good and what is objectively bad (Sutrop, 2020, p. 67). It is really obvious that no intelligent agent in the world today could classify slavery, racism or sexism as something not bad. The argument here is that we can more easily agree on the bad things than the good and, thus, this should be the basis for value alignment for AI. Accepting the view of value pluralism does not change the situation here. There are attitudes that we can declare objectively bad. There is certainly an opportunity to start solving the value alignment problem from this side of the coin.

Still, we cannot forget about the other side. Isn't it really a matter of not being able to put forward anything objective that we would want to happen to us individually, or that would be of basic value for the society in which we live? What about freedom, security, dignity? Or more specifically, what about helping people with disabilities (Joamets, & Chochia, 2021). So, there

is still the question in the context of the value alignment problem whether to start from the preventive or from the promotive side, or develop both in parallel? Our humble suggestion is that we should still deal with both directions in parallel.

As we could see above, there is a large area of problems where AI has to take up the preventive attitude in order to protect humans. It is the global problems. Still, avoiding and helping to resolve the global problems can be viewed as a promotive behaviour as well. Perhaps it could be questioned whether addressing global problems is a normative issue at all. May be not, but ethics is involved here beyond any doubt.

We do not have a straightforward answer to the question of what kind of value alignment we should prefer. Still, there are good reasons to believe that focusing on just the preventive or just the promotive approach would be an oversimplification. There is a universal concept that has to be adhered to—that of "plurality". It should never be confused with either relativity or subjectivity. And let us not forget that the normative side of the value alignment problem cannot be resolved without AI possessing an appropriate type of rationality.

6. Conclusion

Artificial intelligence continues to be developed and we, the biological minds, will become increasingly surrounded by the digital minds. The major task for humans is to stay in control in this unprecedented situation in the history of civilisation because we continue to carry the responsibility for what is going to happen. First of all, this means paying constant attention to the value alignment problem, especially its normative aspect. The technological side is evolving rapidly and we better keep up with our considerations on the normative aspect of the problem. In this article we have analysed the problem from different perspectives, basing the approach on some "official" understandings of the content of the main concepts used. The article proposes the novel idea that dealing with the value alignment problem requires taking a fresh look at the concept of rationality and connecting the new understanding of rationality that is based on the ideas of Nicholas Maxwell with developing digital minds. We suggest to proceed from here and ask whether the missing something argument would help us, although we have great difficulty in specifying what exactly we are missing here. Is it

having the soul, being alive, being conscious, all of these, or something else entirely? This will be an interesting problem for forthcoming studies.

Concerning the issue of preventive or promotive approach to the value alignment problem, we arrived at the conclusion that it would not be reasonable to limit ourselves to one or another of these approaches but they should rather be pursued in conjunction, as this is also the way the normative side of the biological minds functions. There are some, although not too many, universally bad as well as universally good things that mentally healthy human beings would always like to happen or want to prevent from happening. The keyword of the balanced approach to values would be plurality, not relativity or subjectivity because contrary to relativity and subjectivity, plurality does not rule out objectivity.

7. Acknowledgements

The research of Peeter Müürsepp on the article was supported by the Estonian Research Council grant PRG462.

Zhumagul Bekenova is a doctoral student of the Faculty of Philosophy of Al-Farabi Kazakh National University. She graduated with a master's degree in philosophy. Bekenova's research interests are philosophy of technology and the ethical problems of the digital society. She has conducted research at Istanbul University (Turkey). She is the author of scientific publications on the philosophy of technology, on the problems of digital philosophy. She is also involved in pedagogical activity at the Department of Philosophy of Al-Farabi Kazakh National University.

Peeter Müürsepp is an associate professor at Tallinn University of Technology and frequent visiting professor at Al-Farabi Kazakh National University. He is a corresponding member of the International Academy of the History of Science and chairperson of the Estonian Association of the History and Philosophy of Science. He is editor-in-chief of two Scopus indexed journals: *Acta Baltica Historiae et Philosophiae Scientiarum* and *ICON*. He has been visiting researcher or professor at several well-known academic centres, including London School of Economics and Political Science, Helsinki Collegium for Advanced Studies, Shanghai University, etc. His research and publications are in philosophy and history of science and technology.

Gulzhikhan Nurysheva is a professor at Al-Farabi Kazakh National University. She has carried out research in the framework of several international research projects. In the course of a translation project she has translated *A Brief History of Philosophy, From Socrates to Derrida*, *The Beginning of Modern Philosophy* and *Modern Philosophy*. She is a member of the Academic Council of the International Symposium 'Philosophical Heritage of the Turkic peoples' (Turkey), editorial board member of the Al-Farabi Kazakh National University Bulletin, Academic Research in Educational Sciences (Uzbekistan) and Acta Baltica Historiae et Philosophiae Scientiarum (Estonia). She was awarded the badge of the Ministry of Education and Science of the Republic of Kazakhstan 'For merits in the development of science' (2019) and the title of The Lecturer of the Year 2019.

Laura Turarbekova is an associate professor at Al-Farabi Kazakh National University. She is executive secretary of the *Journal of Philosophy, Culture and Political Science*. She is an associate member of and postdoctoral researcher at Gilles Gaston Granger Center at Aix-Marseille University. She is the academic coordinator of the project 'Digital Humanities in the Republic of Kazakhstan: local solutions for global problems'. She received her PhD at Jean Moulin Lyon 3 University. Her research and publications are in French and world postmodernism, contemporary Deleuzianism, digital philosophy and digital humanities in general, history and philosophy of science, and philosophy of education.

References

Arnold, T.; Kasenberg, D. & Scheutz, M. (2017), 'Value alignment or misalignment—what will keep systems accountable?' *The AAAI -17 Workshop on AI, Ethics, and Society, WS-17-02*. Retrieved from https://aaai.org/ocs/index.php/WS/AAAIW17/ paper/view/15216/14648 [accessed Dec 2020]

Berlin, I. (2013[1947]), 'The pursuit of the ideal,' in H. Hardy (ed.) *The Crooked Timber of Humanity*, Princeton & Oxford*:* Princeton University Press.

Bostrom, N. (2003), 'Are you living in a computer simulation?' *Philosophical Quarterly*, vol. 53, no. 211, pp. 243–255. https://doi.org/10.1111/1467-9213.00309

Bostrom, N. (2019), 'The vulnerable world hypothesis,' *Global Policy*, vol. 10, no. 4, pp. 455–476. https://doi.org/10.1111/1758-5899.12718

Bryson, J. (2018), 'Patiency is not a virtue: The design of intelligent systems and systems of ethics,' *Ethics and Information Technology*, vol. 20, no. 1, pp. 15–26. https://doi.org/10.1007/s10676-018-9448-6

Chalmers, D. (2010), 'The singularity: A philosophical analysis,' *Journal of Consciousness Studies*, vol. 17, nos. 9–10, pp. 7–65.

European Commission (2019a), *A Definition of AI: Main Capabilities and Disciplines*. Retrieved from https://www.aepd.es/media/docs/ai-definition.pdf [accessed Sep 2021]

European Commission (2019b), *Ethics Guidelines for Trustworthy AI*. Retrieved from https:// ec.europa.eu/futurium/en/ai-alliance-consultation [accessed Dec 2020]

Fisac, J. F.; Gates, M. A.; Hamrick, J. B.; Liu, C.; Hadfield-Menell, D.; Palaniappan, M.; Malik, D.; Sastry S. S.; Griffiths, T. L. & Dragan, A. D. (2018), 'Pragmatic-pedagogic value alignment,' in N. Amato, N. G. Hager, S. Thomas & M. Torres-Torriti (eds.) *Robotics Research: Springer Proceedings in Advanced Robotics 10*, Cham: Springer, pp. 49–57. https://doi.org/10.1007/978-3-030-28619-4_7

Gabriel, I. (2020), 'Artificial intelligence, values, and alignment,' *Minds and Machines*, vol. 30, pp. 411–437. https://doi.org/10.1007/s11023-020-09539-2

IBM (2019), 'Everyday ethics for artificial intelligence', *IBM Corp*. Retrieved from https:// www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf [accessed Sep 2021]

IEEE (2019), *Ethically Aligned Design. A Vision for Prioritizing Human well-being with autonomous and intelligent systems*, The Institute of Electrical and Electronics Engineers, Incorporated. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf [accessed Dec 2020]

Joamets, K. & Chochia, A. (2021), 'Access to artificial intelligence for persons with disabilities: Legal and ethical questions concerning the application of trustworthy AI,' *Acta Baltica Historiae et Philosophiae Scientiarum*, vol. 9, no. 1, pp. 51–66. https://doi.org/10.11590/abhps.2021.1.04

Kagan, S. (2019), *How to Count Animals, More or Less*, Oxford: Oxford University Press.

Kerikmäe, T.; Müürsepp, P.; Pihl, H. M.; Hamuľák, O. & Kocharyan, H. (2020), 'Legal person- or agenthood of artificial intelligence technologies,' *Acta Baltica Historiae et Philosophiae Scientiarum,* vol. 8, no. 2, pp. 73–92. https://doi.org/10.11590/abhps.2020.2.05

Kesa, A. & Kerikmäe, T. (2020), 'Artificial intelligence and the GDPR: Inevitable nemeses?' *TalTech Journal of European Studies*, vol. 10, no. 3, pp. 68–90. https://doi.org/10.1515/bjes-2020-0022

Maxwell, N. (2021), *The World Crisis – And What to do About It. A Revolution for Thought and Action,* New Jersey, London, etc.: World Scientific.

Nyman-Metcalf, K. & Kerikmäe, T. (2021), 'Machines are taking over—are we ready? Law and Artificial Intelligence,' *Singapore Academy of Law Journal*, vol. 33, pp. 24–49.

Russell, S. (2019), *Human Compatible AI and the Problem of Control*, London: Penguin Books.

Russell, S.; Dewey, D. & Tegmark, M. (2016), 'Research priorities for robust and beneficial artificial intelligence,' *AI Magazine*, vol. 36, no. 4, pp. 94–105. https://doi.org/10.1609/aimag.v36i4.2577

Searle, J. R. (1980), 'Minds, brains, and programs,' *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–457. https://doi.org/10.1017/S0140525X00005756

Shulman, C. & Bostrom, N. (2021), 'Sharing the world with digital minds,' in S. Clarke, H. Zohny & J. Savulescu (eds.) *Rethinking Moral Status*, Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780192894076.003.0018

Soares, N. (2015), *The Value Learning Problem*, Technical Report, Machine Intelligence Research Institute.

Soares, N. & Fallenstein, B. (2014), *Aligning Superintelligence with Human Interests: A Technical Research Agenda*, Technical Report, Machine Intelligence Research Institute.

Sutrop, M. (2020), 'Challenges of aligning artificial intelligence with human values,' *Acta Baltica Historiae et Philosophiae Scientiarum*, vol. 8, no. 2, pp. 54–72. https://doi.org/10.11590/abhps.2020.2.04

Zamyatin, Y. (1993), *We*, London: Penguin Classics.