3. Olivér Gábor: CRITIQUE OF THE ASILOMAR AI PRINCIPLES - Gábor Olivér: AZ ASILOMARI ELVEK KRITIKÁJA (/free-books/69-artificial-intelligence-mesterseges-intelligencia/oliver-gabor-gabor-oliver-critique-of-the-asilomar-ai-principles-az-asilomari-elvek-kritikaja/177-3-oliver-gabor-critique-of-the-asilomar-ai-principles-gabor-oliver-az-asilomari-elvek-kritikaja-oldalszammal)

Gábor Olivér Nyomtatás

2022K/Olivér Gábor: Critique of the Asilomar AI Principles - Gábor Olivér: Az asilomari elvek kritikája (/free-books/69-artificial-intelligence-mesterseges-intelligencia/oliver-gabor-gab

2022. január 18. Készült: 2022. január 18. Módosítás: 2022. március 22. Találatok: 119

ISBN 978-615-5687-04-4,

[p. 2] - indications of page numbers

Critique of the Asilomar AI Principles

Olivér Gábor

Abstract

The intelligent man with consciousness is the pinnacle of the evolution of matter that we know. So far. We know, however, that evolution will not stop. Although with sections and dead ends of varying lengths, it is moving towards the increasing complexity of organizations, so it is probably not the human in today's sense is its end point. The development of artificial intelligences that we are planning, for example, may meet both our intentions and the criteria of evolution, but as a novelty it also holds the possibility of a future without us. The latter, in turn, creates tension between the species-maintainer desires of homo sapiens and the unknown future course of evolution. The Asilomar principles seek to alleviate this tension by limiting the development of artificial intelligences. However, as technological advances lead to an increase in autonomy, this is at most a plan for time-gaining. In addition to the Asilomar program, then, there is a need for a "Second Foundation" that can reconcile the future of man not only with artificial intelligences but also with evolution. If we want to survive, the evolutionary adaptation of homo sapiens could really ease the pressure of technological determinism on us.

At the 2017 International Conference on Artificial Intelligence Safety Technology in Asilomar, participants signed an agreement.[1] They were of the opinion that the development of artificial intelligences should be controlled. More specifically, to limit the future development of algorithms in a way that suits for homo sapiens. In doing so, they sought to meet the future challenges posed by autonomous technologies.[2] The question is whether the Asilomar goal is a real possibility or just a formulation of desires? In the following, after a brief introduction to the ability for get know and formability of the future, I examine the truthfulness of the Asilomar program.

[p. 4]

I. APPROACHES TO THE FUTURE

As long as we exist advancing in time, we have no direct experience of the future, at most if it has already happened. However, our security of existence and the maintenance of our species require the reassuring arrangement of the future. That is why, in addition to our knowledge builds on the past and our consciousness exist in the present,[3] our intelligence is highly future-oriented. While the inference of our immediate individual future may be instinctive or conscious,[4] the prediction of our distant common future is more conscious and speculative (futurology). In planning the future, however, we shape the near future consciously, while in imagining the more distant future of humanity, desires are stronger. Do these collective desires stem only from our species-maintaining instincts, or do they rather reflect the general "goals" of evolution? The question is important because in the first case the goal is the survival of homo sapiens, but in the second it is no longer necessarily the case, because the evolution goes on with or without us.

II. GETTING TO KNOW THE FUTURE

It's a poor sort of memory that only works backwards, - the Queen remarked.[5]

The possibility of thinking about the future and predicting is actually within us, because then the same process takes place as when we recall something.[6] Memory is not for only recalling the past, but also for predicting the future. *Episodic memory, enabling conscious recollection of past episodes, can be distinguished from semantic memory, which stores enduring facts about the world. Episodic memory shares a core neural network with the simulation of future episodes, enabling mental time travel into both the past and the future*.[7] This is how the *behavioral immune system* works,***[8]*** which is in a human, a conscious or subconscious code of conduct that helps, to avoid accidents. All of this can be instinctive or subconscious behavior for that animals are also capable. For example, if we get hurt in a situation, next time we will involuntarily avoid it. As if knowledge and experience were a kind of sixth sense, which sees more clearly into the near future and more and more vaguely into the distant future. If, on the other hand, we are planning a house or writing a diary, we think we are doing it with a much more rational future consciousness than animals, when in fact *We deliberate not about ends but about means. For a doctor does not deliberate whether he shall heal, nor an orator whether he shall persuade, nor a statesman whether he shall produce law and order, nor does any one else deliberate about his end.*[9] So where is the truth? Do we create goals too, or do we simple receive ready goals?

[p. 6]

I know five possible ways of learning about the future. These can also be showed by historical examples. However, the evolutionary rules and physical laws[10] that justify changes, may already suggest a kind of causal determinism.

1- **Conclusions based on past experiences.** (Analogical reasoning.)

Historical example. In 146 BC, watching the Carthage set on fire by his soldiers, Scipio Aemilianus was reminded destruction of Troy.[11] And from the fate of the two burned cities, he concluded that one day the power of Rome would also fade away.[12]

Evolutionary rules. Evolution has repeated itself many times, as if it had been very difficult to learn from the mistakes of the past.[13] Yet with varying speeds of intermittent (*saltation*) development, it reached an ever higher level of organization. Eventually, it created animal and then human intelligences that, in a good case, can really reckon with past experiences. And artificial intelligences never make the same mistake again.[14] All this shows that the concept of evolutionary development includes not only the achievement of ever higher organization, but also the change of evolutionary methods. Evolution is thus future-oriented and uses the analogies of the past with increasing efficiency to shape the future.

Physical laws. The memory of a strong external influence is preserved by the altered structure of the material. The memory of solids is, of course, much better than that of gases and liquids, just think of a car body after a crash, for example. Electric charging is even more suitable for coding. [15] Dealing with memories (information) requires a higher level of matter, and that is intelligence. Finally, the highest level, that is, consciousness, already sets goals and in order to achieve them is able to apply the acquired knowledge that matter remembers.

The burning houses of Troy and Carthage have preserved memories of the war. As an intelligent man who knew the past, Scipio Aemilianus rightly assumed a similar future fall of Rome. As he obviously wanted to avoid it, he consciously drew attention to the future danger of it.

[p. 8]

2- **Recognizing the rules of change and projecting them into the future.**[16]

Historical example. In the 18th century, Edward Gibbon wanted to see the British colonial empire eternal but history has taught him otherwise. He believed that the decline of empires was lawful because it was the order of nature.[17] Later, Ronald Wright, starting from the fact that individual empires run similar ways, also demonstrated their regularities (universals). According to him, the decline of empires was signaled by wars and flight (migration), which resulted in the mixing of peoples. And all this always brought about the rise of new communities, new empires.[18]

Evolutionary rules. The concepts of *selection*, *renewal*, *niche*, *fitness*, and *cyclical intermittency* can be included here. Aging, for example, is a genetically programmed *selection* that is able to maximize the renewal and adaptability of the group by sacrificing some of them (old people). The declining group due to the loss of the elderly will be demographically disadvantaged for a short time, but due to its fitness advantage (*renewal*) it will be easier to adapt to the changed environmental conditions.[19] The term *niche* refers to an empty part of an ecosystem (food chain) into which the emerging population of a given species fits. *Fitness* shows the selection success and viability of individuals when they are able to survive in the given environmental conditions. Finally, the *cyclical intermittency* of evolution is a kind of repeating regularity. All of these may also be valid phenomena for the evolution of human cultures, as history shows many examples of the rise or fall of empires. The fitness value of an old empire is declining, while an ascending young empire may take its place in the power vacuum (niche) that is emerging around it because the new culture is not specialized, and in some cases much more open to novelties than the old one.

Physical laws. The changes result from the different physical states of the substance.[20] The duration of the stages between changes depends on the stability of the material and its environment. The original tendency of the matter to be disordered explains the niche phenomenon of evolution (filling in the blank places) and the fitness phenomenon because a given particle is most likely to occur at the most suitable location. In addition, the process in the opposite direction to disorder, ie the self-organization of matter, can also cause change. Even then, it is a question of a change between stable and unstable states. In social terms, as long as a community is well-organized, in a stable state, change is not current. However, as stability breaks down for external or internal reasons, the organization of society seeks new, more stable states.

At the end of antiquity, the Roman Empire, and in modern times the British Colonial Empire, represented age-appropriate fitness or the highest organization, but eventually both became unstable. This was followed by the rise of new communities according to the rules of cyclical intermittency (Germanic kingdoms and USA).

[p. 10]

3- **Prolonging of past processes.**

Historical example. Ronald Wright believed that the failure of empires was not in vain, as they fit into the history of the development of a larger unit, the entire human civilization, as they reached ever higher degrees.[21]

Evolutionary rules. Evolution is wasteful and doesn't work well in all cases (see extinctions or evolutionary dead ends), but there is always a line that goes on. The direction in which a higher degree of organization lasts longer can be called an evolutionary success.

Physical laws. The organization of matter (evolution) develops to the direction of higher complexity. The elemental particles after the big bang formed hydrogen, then new elements, molecules, organic compounds and then living organisms. At the end of the long time organization of elementary particles (over 13.8 billion years) was formed intelligent matter: the human. We are the highest known stage of the evolution of matter. Intelligence, on the other hand, facilitates restructuring through the ability to learn, that is, it makes instabil states more frequent. As the organised matter would strive for stable states,[22] the tension becomes almost constant, that can be called developmental compulsion. And this process, which started in the farest past, does not stop, but rather accelerates and continues to point towards increasingly complex organizations.

The same is true of human history. Just as we have no reason to assume that the development of human civilization will stop, nor can we think that homo sapiens is the highest achievable stage of evolution. This is because evolution contradicts the static conception of the Created world.

4- **Recognition of distant goals.** (Causal determinism.)

Because we are generally aware of our short-term individual goals, into this point I list the distant, common goals of humanity. In these, however, it seems at most only the idea of the way to get there that seems conscious, while the goals themselves manifest as collective desires.[23]

[p. 12]

Historical example. On July 20, 1969, Neil Armstrong was the first man to step on the moon. He added the following comment: *That's one small step for man, one giant leap for mankind.* He obviously couldn't have taken that move without NASA's Apollo-11 program, but the lunar voyage wasn't really even invented by U.S. inhabitants. If we look only at the literature of Europe, it becomes clear that the lunar journey is in fact the old dream of mankind. It can already be seen in plays from the 2nd century BC that Icaromenippus flew to the moon on eagle wings[24] and on a water pole of a Greek ship.[25] In the 17th century, in Bishop Francis Godwin's work, Domingo Gonsales with swans,[26] and Cyrano de Bergerac in his own novel with help of dew[27] took to the Moon. Eventually, the most famous fiction on the subject became Jules Verne's 19th century novel, in which the characters leave the Earth with a cannonball.[28] So we don't know where the desire for the moon and stars came from, but perhaps it was already present in prehistoric people and even wolves?

Evolutionary rules. Under ideal conditions, all species are forced to reproduce and expand, as they are more likely to survive. It is likely that this species-maintenance motive appears in the collective dreams and mythologies of mankind about the Moon. And in 1969, it all realised, as the Apollo-11 landing unit brought terrestrial microbes to the moon and Neil Armstrong hoisted the U.S. flag.

Yet, how the rules of evolution transform to human desires pointing into the distance?[29] The answer can be found in the selfish gene,[30] the function of which can be interpreted at the level of both the individual and the species. The selfish gene serves itself to help the species survive while putting the individual at a disadvantage (altruism***[31]***). However, the "interest" of the selfish gene is also future-oriented through species-maintenance, which is best recognized in human behavior. A larger goal, for example, is to make a trip to Mars or programming the artificial intelligences. The former is aimed at getting our species out into space, and the latter at the easier future prosperity. Furthermore, the result of both is the spread of intelligence. None of them are ordinary engage of the working man, yet it devours some of his possessions. For a Mars mission, for example, the selfish gene sacrifices not so much the individual for the future, but rather a part of the work output of the entire human race. What's more, it does all this from within, meaning that evolution in this case not only responds to environmental influences, but also gives us a coded inner urge.[32] And these codes are embodied in selfish memes[33] (ideas, distant goals) created with the help of the human intellect. The recognition of internal evolutionary programming, which starts from selfish genes and is expressed in the selfish memes of our culture, is a causal determinism and at the same time an analysis of the fate of humanity.[34] At night, observing to the Milky Way, we think that it is our innermost longing, which perhaps nothing more than an algorithm written into our genes.

[p. 14]

Physical laws. During the general disarrangement of the material (*thermodynamic 2nd law, entropy, dissipation*), a local property is the local increase of the heat absorption capacity. This can result in the self-organization of the material (*adaptability*) and its partial arraning (*anti-entropy, negentropy*). And nature behaves as if its goal is to create increasingly complex self-organizing systems with heat absorption[35] (*evolution*). With Neil Armstrong's step, human intelligence, the highest known stage in the evolution of matter, has involved the Moon in this self-organization.

Homo sapiens achieves its distant goals one after the other: it has widespread on Earth, learned to fly, and now discovers new exoplanets. Do not we invent the distant goals, but we have an inner urge to achieve them? That's how we got to the moon.

5- Conscious shaping the future.

Planning future is the most direct way to learn about the future. Through its proactive nature,[36] it goes beyond the conclusions based on past experiences (1), the usage of rules of changes (2), the prolongation of past processes (3), and anticipating the more distant goals of evolution (4). The more successfully we plan our future, the less we may be surprised.[37] As part of planning, long-term desires about the future (codes written into our genes) force themselves into the distant future almost independently of our free will.[38] In contrast, the use of our past experience (learning the laws of physics) tries to concretize all of this in the short term, always winning some more time to further refine it. The planning of the future thus becomes an endless, half-predestined and half-influenced by us process of realization.

Historical example. The Asilomar program seeks to consciously influence the development of artificial intelligences because of future of humanity.

Evolutionary rules. Higher levels control[39] can also be aimed at controlling future processes, but only over lower-level organizations. However, development comes precisely from the becoming more complex of lower-level organizations, which is nothing more than evolution itself. While higher levels control can be recognisable, the process of making the lower levels higher can result in unknown new qualities. If artificial intelligences were to exceed homo sapiens, we would no longer be able to control them from a lower level of development.

Using and lack of physical laws. The self-organization of matter with heat uptake[40] is the process of becoming more complex, which we trie to consciously control.[41] The more accurately we can calculate with the laws of physics, the more we can determine the future. In the meantime, however, our consciousness is *substrate-independent*,[42] and the laws of physics do not apply to it either.

[p. 16]

III. REALIZATION OF FUTURE VISIONS

In connection with the planning of the future, I will present very briefly three historical examples together with their lessons. (1.) Plato's Republic,[43] (2.) Communist utopia by Friedrich Engels - Karl Marx,[44] and (3.) SETI-movement by Carl Sagan - Frank Drake.[45] The first two were social programs, while the third had strong social implications but were academic.

1. Plato's Republic. Plato considered all previous state forms bad.[46] Instead, he invented an ideal state form ruled by philosopher kings in which power and reason work together.[47] There was an opportunity to put his idea into practice in Syracuse, but the experiment failed. I. Dionysius enslaved Plato, II. Dionysios charged him with treason,[48] and eventually his friend Dion was murdered. Plato from then on found as most ideal the Macedonian kingdom of III. Perdiccas,[49] and in Syracuse he proposed the co-rule of three kings.[50] In his old age work, he modified his ideas about the state. By then, he already thought that the state should be governed by laws instead of scientists.[51] Several forms of state were declared by him acceptable if they take into account the interests of all and govern the state through constant discussions.[52]

Plato's experiment failed, but it also had results. On the one hand, he showed that the application of the ideas considered true can also be tested in the shaping of society, and on the other hand, he realized that the original ideas needed to be modified during their implementation.

2. Of the communist utopia of Marx and Engels,[53] the proletarian dictatorship was the mostly realized. In Russia, it was created by Lenin, perpetuated by Stalin, mitigated by Khrushchev, and finally demolished by Gorbachev. Other countries have tried it, but difficult to get rid of its dictatorial character.

The attempt to realize communism has resulted in the deaths of many millions of people and the suffering of many more. In the absence of freedom, the proclaimed equality of socialism could not be fulfilled, and the the dictature brought the opposite of Christian community of Jesus.[54] Eventually, even the original communist principles stiffened into a misunderstood dogma. However, this experiment also had results. Its direct historical impact was in the expansion of labor rights, the defeat of fascism, and then the sharing of world power. Ferdinand Lassalle, who deviated from Marxist principles, was perhaps the most successful in representing the interests of the working class, as he was able to form a workers' party that still exists in Germany.[55] However, Lassalle could not have foreseen that by the 21st century, the classical working class would disappear in Europe and the capitalists would gain unprecedented wealth. The "experiment" of communism actually confirmed both lessons of Plato's experiment: social ideas are needed, but they need to be constantly modified as they are implemented.

3. SETI was born in 1963 as an interdisciplinary discipline.[56] Its main purpose is to detect aliens. From 1974, it became a global movement from the broadcast of the enthusiastic Arecibo message[57] by astronomers Carl Sagan and Frank Drake to space.[58]

[p. 18]

In terms of the main goal of the SETI movement, it has so far had no results.[59] Today, enthusiasm has waned and criticism has been leveled on it.[60] At the same time, the program had its other results. The science of astrobiology was born, the relevant toolbar of space

exploration has advanced greatly with the discovery of exoplanets and gravitation waves, and humanity has, of course, never given up the fight against "cosmic loneliness". The general lesson of the SETI movement is similar to the first two examples. Future-oriented ideas will always be needed, but they must change as our knowledge grows.

The conclusions of the three presented programs can also be drawn in the language of evolution:

- We always have collective desires (*selfish gene, genetic program*) for the future realization of the True and Good (*species maintenance*)[61] that are embodied in ideas or visions (*selfish meme*).

- Of these five basic forms of learning about the future, these ideas usually like to use only the last two. The past is often considered to be discarded or irrelevant to the future, so many things are "reinvented" again (*convergent evolution*). As a result, the rules that apply to the past, as well as the processes that started in the past, seem more outdated to them. And the new programs themselves behave as if they were on a mission, and some of them can force themselves on the future for a while (runaway phenomenon).

- The civilization needs visions that can be the engine of development (*Sturm und Drang, progressor activity*). However, the first versions of these visions are necessarily still rough, subjective, so they cannot be realized in their original form (*evolutionary dead ends*).[62] But if they are given some time, with the help of our growing knowledge (expanding knowledge of the laws of physics), these ideas can be made more and more realistic (*selection*). History has shown that the original "intention" of evolution (*increase in complexity*) is also prevalent on a higher level, which in the case of 21st century human is no longer biological rather technological and social development. Our free will thus lies not in setting distant goals, but in finding the way there (*substrate-independence*).

[p. 20]

IV. CRITIQUE OF THE ASILOMAR AI PRINCIPLES

Technological development has social implications, so humanity wants to regulate it at least since the age of Enlightenment.[63] The desire to curb the latest types of autonomous technologies (artificial intelligences) is not new either. Isaac Asimov and John Campbell have already articulated them in the 4 Laws of Robotics,[64] since then the USA,[65] the Google,[66] the OECD,[67] the IEEE,[68] the EU,[69] the BAAI,[70] the Microsoft,[71] and others have also published their own principles,[72] in the end, also the pope drew attention to respect for the Created World.[73] In terms of Superintelligence, which does not yet exist, Nick Bostrom has listed the most possible types of control to date.[74]

The Asilomar principles[75] focus on controlling the development and operation of artificial intelligences. Their declared tools are verification, validation, IT security, and control. They offer malleable resistance *(resilience*) for the development of autonomous technologies (*AI safety*). The good intentions of the principles are undoubted, but they do not take into account the first four possibilities of learning about the future, so the important question is how much they need to change now? To estimate this, I divided them into three groups:

Realistic principles

- *Transparency*, freedom of information based on trust and *Cooperation* (4) are real needs. The datas are not unique in the first place because, apart from us, these does not exist in just one copy. And the information formed when interpreting the data is worth something only when it is used, which makes it immediately public, so it cannot be hidden in the long run. At the same time, information acquisition and cooperation[76] are now evolutionary requirements, and whoever does not meet them lags behind. The strengthening of collective intelligence anyway favors the unconditional sharing of information. Transparency is therefore a realistic goal because it does not appear to be contrary to the laws of physics and evolution.

- *Importance* (20): the expected large social impact[77] of artificial intelligences must be taken seriously. The topic was really only of interest to research pioneers and science fiction writers at first, but is now matter of global interest. The credibility of the principle is given by the broad social interest behind it.

[p. 22]

- *Intended risk reduction* (21). This point represents the main general goal of the Asilomar principles. Evolution means constant change, but each levels it achieves strive to maintain their own stability,[78] thus, the species-maintenance instinct of homo sapiens dictates risk reduction. Intention thus shows evolutionary regularity, and a secondary issue in this regard is the weakened chance of the goal being achieved by other evolutionary rules. We simply have no choice unless we live in the age of stupidity.

Realistic goals in the short term

- *Supporting the beneficial intelligences* (1-2). History has shown that in the long term, those who did not make good use of their research opportunities or discoveries did badly. Native Americans underestimated the significance of the wheel, the Vikings Vinland, China the gunpowder, the paper, the compass, the shipping, and U.S. the cybercrime. After all, there will always be a research base or country that will reap the developmental benefits of increasing machine autonomy at all costs. And if a person or a community can gain a situational advantage, they will not necessarily consider the posibility of calibration of artificial intelligences to beneficial direction, not even for the benefit of humanity.

- *Avoiding the research race* (5). The modern (capitalist) societies of the world are based on development forced by economic competition. As long as this freedom (*liberalism*) is existing, it gives them an advantage, while excessive self-restraint (*decadence*) would cause lag.[79] The case of the Lighthill report[80] shows that research on artificial intelligences can be restrained in the short term at most (*first AI winter*). This also applies to the need of *Avoid arms competition* (18), as in the long run only restrictions on the use of weapons have worked till now. Finally, these goals of the Asilomar program are not realistic either because these principles were only signed by researchers, not by governments!

- *Safety* (6). Ensuring the reliable operation of artificial intelligences is only conditionally considered possible by the Asilomar program. Since even today's artificial intelligences have just enough of an uncontrollable impact on human society, it would be unethical to take responsibility in advance for the reliable operation of as-yet-unknown types. Today, we cannot count on all future problems yet, so neither the promises on the long run to ensure the *Failures, Judicial and Responsibility transparency* (7-8-9) can be kept. For the same reasons, the maintenance of *Human control* as much as possible (16) and the *fight against subversion* (17), ie keeping the level of social impact of artificial intelligences, can only be expected in the short term. This includes also the question of controlling the *Recursive self-improvement* of artificial intelligences (22).

[p. 24]

- Enforcing *Universal Human Values* (10–11), *Equal Distribution* of the Benefits of Artificial Intelligences (14), *Shared Prosperity* (15), and Striving for the *Common Good* (23). Throughout human history, we have been forced to take into account physical and biological laws, and even a small degree of autonomy of our pets (biting dog, stubborn donkey, self-contained cat, etc.). The essence of artificial intelligences, on the other hand, is the increase in autonomy, which is of great benefit to us for the time being. And the strengthening of machine autonomy clearly weakens the assertion of human will, values and morals, so there is only a short-term chance of upholding these humane principles. Their partial unreality stems from the fact that they apply to artificial intelligences, and they did not sign the Asilomar principles!

Unrealistic goals

- *Peer dialogue between research and policy* (3). The policy of a state has necessarily always been determined by the historical situation, not the morals of scientists and researchers.[81] Where this was not taken into account, they reached the fate of the state of Plato.[82]

- *Privacy* / data protection (12-13). It is an unrealistic principle because it contradicts freedom of information. It is clear that human is abusing data in the same way as artificial intelligences. In addition, one voluntarily passes on data and information to machines, and even supports their aggressive information getting (see search engines, profiling programs, and spywares). The evolution of quantum computers, anyway, flashes the endless perspective of code hacking.

- *Capability assumptiones* (19). The rapid development of artificial intelligences does not seem to have stopped yet, and we have no idea about the possible final outcome of the process. What seemed impossible yesterday will be a reality today and an obsolete one tomorrow. "Reliable" guesses about this should indeed be avoided, as it rightly put by the Asilomar principle.

[p. 26]

V. „SECOND FOUNDATION"[83]

With the advent of artificial intelligences, terrestrial intelligence is becoming bipolar.[84] The average intelligence level of people who represent one pole sometimes seems to decrease.[85] Because the Asilomar program considers the survival of humanity to be the most important moral value, it does not trust the artificial intelligences that form the other pole which are evolving at an alarming rate. His mindful optimism applies to the safety technology of artificial intelligences and the control of their development. However, it is only possible to think about artificial intelligences on the basis of today's types, so the demand to correct control must be constantly maintained in the light of their development. The Asilomar Principles website[86] undertakes this and is presumably open to comments similar to this present article. On the other hand, the Asilomar principles of do not take into account the possibility of a transformation of human intelligence, which is precisely originated by appearence of informatics, computers, and networks. However, the strengthening of *collective intelligence* may also redefine our relationship to artificial intelligences. This is because collective intelligence can be increased more than individual intelligence.[87]

The strengthening of networks (mobile phones, internet, GPS) and the pile up of human knowledge (machine memory, cloud-based data storage, wikipedia, etc.) have amplified collective intelligence. Computers and artificial intelligences are now parts of this.[88] On the other hand, with the help of avatars appearing in *metaverzums* and the chips or even biologically produced cyborgs,[89] also the homo sapiens can be more deeply integrated into virtual worlds. He who rules cyberspace controls human culture and, through it, much of planet Earth. All this proclaims the unstoppability of technological evolution (*technological determinism*) in a world where the human monopoly of possessing consciousness still exists (*techno scepticism*). In summary, it is not so much the desire to limit technological progress that - but as so often - the adaptation of homo sapiens, may mean the long-term survival of humanity (*anthropocentrism*). Thus, human must be included not only as a subject but also as an object into the strategy of cyber developments.[90] Thus, in addition to the asilomar principles, a "*Second Foundation*" may seek the way to further evolve of homo sapiens.

[p. 28]

NOTES

[1] Tegmark 2017 416-418

[2] In previous technological challenges, we have not faced the new features represented by artificial intelligences: self-development, increasing autonomy, easy diffusion (copyability / downloadability, cheapness), creation of monopolies (eg distribution of basic software, dominance over networks), and information power takeover (unlimited internet access). Artificial intelligences are not just tools, not just metaphorical agents, but actual autonomous, self-directed, proactive, intelligent agents. (Héder 2021B 119 128).

[3] Our sense of time is subjective, existing in the present or relative present of us. It uses only as reference the real time of our physical environment shown by the clock.

[4] If someone pulls the steering wheel aside of the car to get out of an expected collision, the reflexes will save him/her, while the forecast for the afternoon weather is already a conscious act.

[5] Carroll 1871

[6] Dippold 2020

[7] Suddendorf et al. 2009

[8] *Behavioral immune system* (Schaller - Duncan 2007).

[9] Aristoteles, *Nicomachean Ethics* III. 3, 11-12. (translated by W. D. Ross) - The 5th argument of Saint Thomas Of Aquino for the existence of God is also about the expediency of being *(Summa Theologiae* 1265-1272 *De veritate, q. 2 a. 3.).*

[p. 30]

[10] Galilei stated as early as the 17th century that the world could be described in mathematical language. (Galilei 1638). According to the latest philosophical foundation of the view, the theoretical conceptual (*discursive*) description of the world can only be based on experiential (*intuitive*) understanding. (Förster 2018). Based on this, informatical concepts can also be described by the laws of physics. Learning, for example, results from changes in the state of matter, as certain physical systems memorize repeated arrangements. And information, memory, the process of calculation, learning, and the functioning of consciousness are substrat-independent, meaning that many different things can be a memory store, a computer, or even a conscious entity. (Tegmark 2017 78 80-90 96-98 106) Finally, the laws of physics may also be suitable for description evolutionary and social factors.

[11] Homeros: *Ilias*. Vergilius: *Aeneis.* - Schliemann 1874.

[12] Polybios, *Book* XXXVII V/22 Appianus *Punica* 132. Scipio's conclusion was correct. In 410 AD the Gothic Alaric and in 455 the vandal Geiserich plundered Rome. Finally, in 476, Odoaker the Scir exiled Romulus Augustulus, the last emperor of Western Rome.

[13] Convergent evolution: similar biological features that develop multiple times independently of each other (e.g., eyes, wings, etc.).

[14] Reducing the number of errors optimizes only the current capabilities of artificial intelligences. For the future, however, artificial intelligences lose the flexibility (*redundancy*) of evolution and their ability to adapt to changing circumstances.

[15] Tegmark 2017 78 80-81

[16] *Who controls the past controls the future.* (Orwell 1949)

[17] Gibbon 1776-1789 - By the end of the 20th century, Alexander Demandt had already listed 210 reasons for the fall of Rome. (Demandt 1984).

[18] Wright 2004

[p. 32]

[19] *Ageing might have evolutionary advantages.* (Santos 2021) – For example, the ancient Greeks showed with the *colonization*, the Romans with the *ver sacrum method*, and the Celts with the wandering the viability of new generations.

[20] ...*a seemingly dumb clump of matter can remember and compute*... (Tegmark 2017 96-97).

[21] Wright 2004

[22] The only known exception to the pursuit of stability is the time crystal, which shows continuous change even without energy uptake (WILCZEK 2012), because Theorem 2 does not apply to it.

[23] *Collective desires* cannot be interpreted as an ensemble of synchronized minds (a *morphogenetic field*?), but rather as a common genetic heritage manifested in individuals (like the *collective unconscious* - JUNG 1934).

[24] Lucian of Samosata, *Ikaromenippos*.

[25] Lucian of Samosata, *True story*.

[26] Godwin 1638 - On the possibility of lunar life: John Wilkins (1638), and Filippo Morghen (1776).

[27] Bergerac 1657

[28] Verne 1865

[29] *Macte nova virtute puer: sic itur ad astra.* (*Persevere in virtue, my son, thus is the way to the stars.* – Vergilius, *Aeneis*).

[30] *Selfish gene* - Dawkins 1976

[p. 34]

[31] For example, bees heat the hive with their bodies, mammalian mothers feed their offspring from their body, and some adult meerkats help raise the pups of the alpha pair.

[32] The organization of matter gives the uniqueness of our universe. It is a quality which, does not necessarily follow from the expansion process between the matter's come into being (*Big Bang*) and its supposed collapse (*Big Crunch / Big Rip / Big Freeze*), therefore, the causes of evolution, we believe in something else, in the absence of a better, God's will.

[33] Dawkins 1976 179 – While genes think in the short term, memes prefer in the long term (Szathmáry 2021).

[34] Szondi 1944 Hargitai 2004 380 - According to the modern version of selfish gene theory, our consciousness is made up of a mass of memes behaving like viruses that take over us (DAWKINS 1993 DENNETT 1995), although with our research we are rebelling just against the Creator (Dawkins 1976 chapter 11).

[35] Tegmark 2017 325

[36] The first four options listed for learning about the future can only be considered proactive in the case of self-fulfilling prophecy.

[37] We cannot anticipate for the unexpected events and the qualitative or jump changes. The invention of fire or even the computer has changed our world in an unpredictable way.

[p. 36]

[38] Knowledge of the past already includes an increasingly accurate knowledge of the laws of physics, while overly idealistic desires (e.g, voluntarism) are trying to distort knowledge of the laws of physics.

[39] Polányi 1968

[40] Tegmark 2017 325

[41] There is another view that human history is shaped differently: it is characterized by opportunities instead of goals, and experimentation instead of purposefulness. (Graeber – Wengrow 2021)

[42] Substrate Independent: no matter what its carrier is. (Tegmark 2017 389).

[43] Plato, Republic (πολιτεία - around 357 BC).

[44] Marx – Engels 1848

[45] SETI: *Search for Extra-Terrestrial Intelligence.*

[46] VII. letter - Republic VIII. book. Forms of state: *Timocratia*, *Oligarchia*, *Democratia*, *Tyrannia*.

[47] II. and VI. letters – Republic V. book/XVII.

[48] In 361 BC, the philosopher Arkhytas saved his life.

[49] V. letter - III. Perdikkasz was the uncle of Alexander the Great, but the subsequent successes of II. Philippos and Alexander the Great did not result from adherence to Platonic principles.

[p. 38]

[50] VIII. letter

[51] The Laws

[52] VII. letter

[53] Marx - Engels 1848

[54] *in veritate conperi quoniam non est personarum acceptor Deus sed in omni gente qui timet eum et operatur iustitiam acceptus est illi* (*Everyone is equal before God - ApCsel*, 10,34-35).

[55] Marx 1875

[56] The *Big Air* binocular was commissioned in 1963.

[57] *Arecibo message*: a message from humanity to aliens that will reach the Messier 13 star cluster only 25,000 years from now.

[58] Russian researchers Joseph Sklovsky and Nikolai Kardashev also supported the research (Galántai 2019).

[59] *Fermi paradox*: the appearance of reason is probably not an exceptional thing, but there are so insurmountable distances in outer space that we find no evidence of it.

[60] Critical arguments against SETI methods: treating supposed things as evidence, lack of a true interdisciplinary approach, anthropomorphic approach to outer space, and an unknown way to communicate with aliens (Gindilis - Gurvits 2019 20-22 25-26 Galántai 2019).

[p. 40]

[61] In a philosophical sense: a categorical imperative (Kant 1788).

[62] Very definite visions tend to fail (Pintér 2021).

[63] Héder 2021A - engineering ethics, restrictions on nuclear weapons, internet regulations, etc.

[64] The laws of Asimov and Campbell considered machine intelligences only as a subordinate species (tool), and on the other hand they held them accountable for human morals. (Asimov 1985 Asimov - Campbell 1942 Gábor 2020 footnote 79).

[65] Government of the United States, *Report on the Future of Artificial Intelligence.* - Holdren et al. 2016

[66] Google, *Artificial Intelligence at Google: Our Principles* (2018).

[67] Organisation for Economic Co-operation and Development (OECD), *Recommendation of the Council of Artificial Intelligence* (2019).

[68] *Institute of Electrical and Electronics Engineers (IEEE): Ethically Aligned Design* (EAD 2019).

[69] European Union (EU), *Ethics Guidelines for Trustworthy AI* (AI HLEG 2019).

[70] Beijing Academy of Artificial Intelligence (BAAI), *Beijing AI Principles* (2019).

[p. 42]

[71] Microsoft, *Microsoft AI principles* (2019).

[72] Héder 2020

[73] Pope Francis 2021 – Interestingly, Dubai and China support the development of artificial intelligences almost indefinitely (Tilesch 2021 10), while the Western world is more frightened by it.

[74] Bostrom 2014 36-48 59 66-67 – its critics: Gábor 2020 4-5

[75] *The Asilomar AI Principles* (Tegmark 2017 416-418)

[76] Luke 8:17. Nowak 2006 1563

[77] Feenberg 2003 Tegmark 2017 105

[78] Biological evolution is not about long-term stability, but about maximizing short-term success (Szathmáry 2021). (Sharks that evolved 420 million years ago, for example, have barely changed.) In the case of human, perhaps this is what social organization is trying to counteract, which can respond much more quickly to changing circumstances than biological evolution.

[79] *ruthless AI race – the winner takes it all* (Tilesch 2021 10).

[80] James Lighthill, an English mathematician, was the last man to strike on artificial intelligences. The report he wrote was able to hold back research for a few years. (*Lighthill-report* 1973).

[p. 44]

[81] *…instead of a genuine ethical interest for AI, we are witnessing moral diplomacies resulting in moral bureaucracies battling for moral supremacy and political domination* (Vică – Voinea – Uszkai 2021 83).

[82] In political terms, control is a classic and unsolvable problem (Gyulai – Újlaki 2021 40).

[83] It is from Isaac Asimov's novel: *Second Foundation* (1953).

[84] Animal intelligence is much weaker than human intelligence and is far from evolving as fast as machine intelligence, so it is perhaps a less important factor for the future.

[85] The decrease in the average level of intelligence of mankind may be caused by selection turned in the wrong direction due to overpopulation (*Malthusian theory*) (Clark 2008). Of course, this can be offset by the law of diversity or the preference for intelligence when choosing a mate.

[86] Asilomar AI Principles: *Futureoflife.org*, 2017. - https://futureoflife.org/ai-principles/ (https://futureoflife.org/ai-principles/)

[87] Increasing collective intelligence is based on strengthening trust, collaboration, information sharing, and networking. (Bollier 2007 Scarlat - Maries 2009 Riedl et al. 2020).

[88] The philosopher Bruno Latour describes the world as a network of equal beings in which non-human actors are also independent agents. (Latour 2005)

[89] Kagan 2021

[90] Although in this case human becomes an instrument, all this does not violate the formula of the categorical imperative, because the human itself remains also the goal (*Ding an sich* - Kant 1788).

[p. 3] - oldalszám jelzetek

Az asilomari elvek kritikája

Gábor Olivér

Abstract

A tudattal rendelkező intelligens ember az anyag evolúciójának általunk ismert csúcsa. Egyelőre. Az evolúcióról viszont tudjuk, hogy nem áll le. Változó sebességű szakaszokkal és zsákutcákkal ugyan, de a szerveződés egyre nagyobb komplexitása felé halad, így valószínűleg nem is a mai értelemben vett ember a végpontja. Az általunk tervezett mesterséges intelligenciák fejlődése például egyaránt megfelelhet a mi szándékainknak, valamint az evolúció kritériumainak, de újdonságként már magában rejti egy velünk nem számoló jövő lehetőségét is. Ez utóbbi pedig feszültséget teremt a homo sapiens fajfenntartó vágyai és az evolúció ismeretlen jövőbeli kifutása között. Az asilomari elvek ezt a feszültséget a mesterséges intelligenciák fejlődésének korlátozásával kívánják enyhíteni. Mivel azonban a technológiai fejlődés az autonómia növekedése felé halad, ez legfeljebb időnyerésre alkalmas terv. Az asilomari program mellett tehát szükség van egy „Második Alapítványra", amelyik az ember jövőjét nem csak a mesterséges intelligenciákkal, de az evolúcióval is képes egyeztetni. Ha ugyanis fenn akarunk maradni, akkor a homo sapiens evolutív alkalmazkodásával lehetne igazán könnyíteni a technológiai determinizmus ránk nehezedő nyomásán.

2017-ben Asilomarban a mesterséges intelligenciák biztonságtechnikájával (*AI-safety*) foglalkozó nemzetközi konferencián a résztvevők aláírtak egy megegyezést.[1] Azon a véleményen voltak, hogy a mesterséges intelligenciák fejlődését irányítani kell. Pontosabban az algoritmusok jövőbeni fejlődését a homo sapiens szempontjából megfelelő keretek közé szorítani. Mindezzel pedig az autonóm technológiák jelentette jövőbeni kihívásoknak próbáltak megfelelni.[2] Kérdés, hogy az asilomari célkitűzés valós lehetőség-e, vagy csak a vágyak megfogalmazása? Az alábbiakban a jövő megismerhetőségét és alakíthatóságát taglaló rövid bevezető után az asilomari program igazságtartalmát vizsgálom meg.

[p. 5]

I. A JÖVŐ SZEMLÉLETE

Amíg időben előrehaladva létezünk, addig nincsenek közvetlen tapasztalataink a jövőről, legfeljebb ha az már megtörtént. Márpedig létbiztonságunk és fajfenntartásunk megkövetelik a jövő minél megnyugtatóbb elrendezését. Éppen ezért múltból táplálkozó tudásunk és jelenben élő tudatunk[3] mellett intelligenciánk erősen jövőorientált. Míg a közeli egyéni jövőnk kikövetkeztetése lehet ösztönös és tudatos,[4] addig a távoli közös jövőnk előrejelzése inkább tudatos és spekulatív (futurológia). A jövő tervezésekor viszont a közeljövőt próbáljuk tudatosan alakítani, míg az emberiség távolabbi jövőjének elképzelésekor erősebbek a vágyak. Vajon ezek a kollektív vágyak

csak fajfenntartó ösztöneinkből fakadnak, vagy inkább az evolúció általános „céljait" mutatják? A kérdés azért fontos, mert az első esetben a homo sapiens fennmaradása a cél, a másodikban viszont már nem feltétlenül az, mert az evolúció velünk, vagy nélkülünk, de halad tovább.

II. A JÖVŐ MEGISMERÉSE

Szegényes az az emlékezet, amelyik csak visszafelé működik - jegyezte meg a királynő.[5]

A jövőről való gondolkodás és a jóslás lehetősége valójában bennünk van, hiszen ekkor ugyanaz a folyamat megy végbe, mint amikor valamire visszaemlékezünk.[6] A memória nem csak a múlt visszaidézésére való, hanem arra is, hogy megjósoljuk a jövőt. *Az epizodikus memória, amely lehetővé teszi a múltbeli epizódok tudatos visszaidézését, megkülönböztethető a szemantikus memóriától, amely tartós tényeket tárol a világról. Az epizodikus memória megosztja a központi idegrendszeri hálózatot a jövő epizódjainak szimulációjával, lehetővé téve a mentális időutazást mind a múltba, mind a jövőbe.*[7] Így működik a viselkedési immunrendszer,***[8]** ami az embernél egy olyan magatartáskódex, mely segít a balesetek elkerülésében. Mindez lehet ösztönös vagy tudatalatti viselkedés, amire az állatok is képesek. Ha például valamely helyzetben megsérülünk, akkor legközelebb már önkéntelenül is kerüljük azt. Mintha a tudás és tapasztalat egyfajta hatodik érzék lenne, ami a közeli jövőre jobban, a távolira pedig egyre homályosabban lát. Ha viszont házat tervezünk, vagy határidőnaplót írunk, azt hisszük, hogy az állatokhoz képest jóval racionálisabb jövőtudattal tesszük azt, pedig valójában *Nem a célokat, hanem az eszközöket mérlegeljük. Az orvos nem azt fontolgatja, hogy gyógyítson-e, a szónok sem, hogy meggyőzzön-e, az államférfi sem azon hezitál, hogy jogot és törvényt teremtsen, és senki sem a célt mérlegeli.*[9] Hol van hát az igazság? A célokat is mi teremtjük, vagy egyszerűen készen kapjuk őket?

[p. 7]

A jövő megismerésének öt lehetséges útját ismerem, melyek történelmi példákkal is bemutathatók. E változásokat megindokoló evolúciós szabályok, valamint fizikai törvényszerűségek[10] azonban már egyfajta oksági determinizmust sugallhatnak.

1- **Múltbeli tapasztalatok alapján való következtetés.** (Analógiás következtetés.)

Történelmi példa: Kr.e. 146-ban Scipio Aemilianusnak a katonái által felgyújtott Karthágó tüzeit nézve felrémlett Trója pusztulása.[11] A két leégett város sorsából pedig azt a következtetést vonta le, hogy egyszer majd Róma hatalma is elenyészik.[12]

Evolúciós szabály: Az evolúció sokszor ismételte önmagát, mintha csak nagyon nehezen lett volna képes tanulni a múlt hibáiból.[13] Változó gyorsaságú szakaszos (vagy *szaltációs*) fejlődéssel mégis egyre magasabb szervezettségi szintre jutott. Végül létrehozta az állati, majd emberi intelligenciákat, amik jó esetben valóban képesek számolni a múltbeli tapasztalatokkal. A mesterséges intelligenciák pedig már sosem követik el kétszer ugyanazt a hibát.[14] Mindez azt mutatja, hogy az evolúciós fejlődés fogalmába nem csupán az egyre magasabb szervezettség elérése, hanem az evolúciós módszerek változása is bele tartozik. Az evolúció tehát jövőorientált, és a jövő alakításakor egyre nagyobb hatásfokkal használja fel a múltbeli analógiákat.

Fizikai törvényszerűség: Egy erős külső behatás emlékét az anyag megváltozott szerkezete megőrzi. A szilárd anyagok emlékezete persze sokkal jobb, mint a gázoké és a folyadékoké, gondoljunk csak például egy karambol nyomait őrző autóra. Kódolásra az elektromos töltés még alkalmasabb.[15] Az emlékekkel (információ) való bánáshoz az anyag egy magasabb szintje szükséges, ez pedig az intelligencia. Végül a legmagasabb szint, vagyis a tudat már célokat állít és azok megvalósítása érdekében alkalmazni is képes azt a megszerzett tudást, amire az anyag emlékszik.

Trója és Karthágó leégő házai megőrizték a háborús dúlás emlékeit. Scipio Aemilianus a múltat ismerő intelligens emberként joggal feltételezte Róma hasonló jövőbeni bukását. Mivel nyilvánvalóan szerette volna elkerülni azt, ezért tudatosan felhívta a figyelmet ennek jövőbeni veszélyére.

[p. 9]

2- **A változások szabályainak felismerése és a jövőre való kivetítése.**[16]

Történelmi példa. Edward Gibbon a 18. században a Brit Gyarmatbirodalmat szerette volna örökéletűnek látni, de a történelem másra tanította. Úgy vélte, hogy a birodalmak hanyatlása szabályszerű, mert ez a természet rendje.[17] Később Ronald Wright abból kiindulva, hogy az egyes birodalmak fennállásuk során egymáshoz hasonló íveket futnak be, kimutatta azok törvényszerűségeit is (*univerzálék*). Szerinte a birodalmak hanyatlását háborúk és menekülés (*migráció*) jelezték, aminek eredménye a népek keveredése lett. Mindez pedig mindig új közösségek, új birodalmak felemelkedését hozta el.[18]

Evolúciós szabályok: A *szelekció*, a *megújulás*, a *niche,* a *fitness* és a *ciklikus szakaszosság* fogalmai vonhatók ide. Az öregedés például olyan genetikailag programozott *szelekció*, mely képes az egyedek egy részének feláldozásával (öregek) a csoport megújulását, alkalmazkodóképességét a lehetséges maximumra emelni (fiatalok). Az öregek elvesztése miatt csökkenő létszámú csoport rövid időre demográfiai hátrányba kerül, de fitneszelőnyének köszönhetően (*megújulás*) könnyebben alkalmazkodik a megváltozott környezeti feltételekhez.[19] A *niche* megnevezés egy ökoszisztéma (tápláléklánc) üres részét jelöli, amibe egy adott faj megjelenő populációja beleillik. A *fitness* az egyedek szelekciós sikerét, életrevalóságát mutatja, amikor képesek fennmaradni az adott környezeti viszonyok között. Végül a fejlődés *ciklikus szakaszossága*, pedig olyan evolúciós szabályszerűség, mely ismétlődést mutat. Mindezek érvényes jelenségek lehetnek az emberi kultúrák evolúciójára is, hiszen a történelem rengeteg példával szolgál birodalmak felemelkedésére vagy bukására. Egy hanyatló, öreg birodalom fitness értéke csökken, míg egy felemelkedő fiatal birodalom elfoglalhatja helyét a körülötte keletkező hatalmi vákuumban (niche), mert az új kultúra még nem specializált és adott esetben jóval nyitottabb az új környezeti kihívások újszerű megoldására, mint a régi.

Fizikai törvényszerűség: A változások az anyag eltérő fizikai állapotaiból erednek.[20] A változások közti szakaszok időtartama pedig az anyag és környezetének stabilitásától függ. Az anyag rendezetlenségre való eredeti törekvése magyarázza az evolúció niche-jelenségét, és a fitness-jelenségét, mert az adott helyen a legalkalmasabb részecske előfordulása a legvalószínűbb. Ezen kívül a rendezetlenséggel (*entrópia*) ellentétes irányú folyamat, vagyis az anyag önszerveződése is változást okozhat. Ekkor is a stabil és instabil állapotok közti változásról van szó. Társadalmi tekintetben, amíg egy közösség jól szervezett, stabil állapotú, addig változása nem aktuális. Amint azonban külső vagy belső okokból a stabilitás megbomlik, a társadalom szerveződése új, stabilabb állapotokat keres.

Az ókor végén a Római Birodalom, az újkorban pedig a Brit Gyarmatbirodalom képviselte a kornak megfelelő fitnesst vagy legmagasabb szervezettséget, ám végül mindkettő instabillá vált. Ezután pedig a ciklikus szakaszosság szabályai szerint új közösségek felemelkedése következett (germán királyságok, ill. USA).

[p. 11]

3- **Múltbéli folyamatok meghosszabbítása.**

Történelmi példa: Ronald Wright úgy vélte, hogy a birodalmak elbukása nem hiábavaló, hiszen azok egyre magasabb fokokat elérve beleilleszkednek egy nagyobb egység, a teljes emberi civilizáció fejlődéstörténetébe.[21]

Evolúciós szabály: Az evolúció pazarló, és nem minden esetben működik jól (ld. kihalások, vagy evolúciós zsákutcák), de mindig van olyan vonal, amelyik folytatódik. Azt az irányt nevezhetjük evolúciós sikernek, vagy fejlődésnek, amelyiken a korábbinál magasabb fokú szerveződés hosszabb ideig fennmarad.

Fizikai törvényszerűség: Az anyag szerveződése (evolúciója) egyre bonyolultabb szervezettség kialakulása felé mutat. Az ősrobbanás utáni elemi részecskékből hidrogén, majd újabb elemek, molekulák, szerves vegyületek és aztán élő szervezetek keletkeztek. Az elemi részecskék fájdalmasan hosszú idejű szerveződésének végén (13,8 milliárd év alatt) kialakult az intelligens anyag, az ember. Mi vagyunk az anyag evolúciójának eddig ismert legmagasabb foka. Az intelligencia viszont a tanulás képessége révén megkönnyíti az átstruktúrálódást, vagyis gyakoribbá teszi az instabil állapotokat. Mivel a megszerveződött anyag a stabil állapotok felé törekedne,[22] így szinte folyamatossá válik a feszültség, amit fejlődési kényszernek nevezhetünk. E régmúltban indult folyamat pedig nem áll le, hanem inkább felgyorsul, és továbbra is az egyre komplexebb szerveződések felé mutat.

Mindez igaz az ember történetével kapcsolatban is. Ahogy nincs okunk feltételezni azt, hogy az emberi civilizáció fejlődése megáll, úgy azt sem gondolhatjuk, hogy a homo sapiens az evolúció legmagasabb elérhető foka. Az evolúció ugyanis ellentmond a Teremtett világ statikus elképzelésének.

4- **Távoli célok felismerése.** (Oksági determinizmus.)

Mivel rövidtávú egyéni céljainkkal általában tisztában vagyunk, ezért az emberiség távoli, közös céljait sorolom ebbe a pontba. Ezekben viszont legfeljebb csak az odavezető út elképzelése tűnik tudatosnak, míg maguk a célok kollektív vágyakként jelentkeznek.[23]

[p. 13]

Történelmi példa: 1969 július 20-án Neil Armstrong az első ember volt, aki a Holdra lépett. A következő kommentet választotta hozzá: *Kis lépés ez egy embernek, de hatalmas ugrás az emberiségnek.* Ezt a lépést nyilván nem tehette volna meg a NASA Apollo-11 programja nélkül, de a holdutazást valójában még csak nem is az USA lakói találták ki. Ha csak Európa irodalmát tekintjük, abból kiderül, hogy a holdutazás bizony az emberiség régi álma. Kr.e 2. századi színdarabokban már olvasható, hogy Ikaromenipposz sasszárnyakon,[24] illetve egy görög hajó vízoszlopon[25] repült a Holdra. A 17. századi Francis Godwin püspök művében Domingo Gonsales idomított hattyúk segítségével,[26] Cyrano de Bergerac pedig saját regényében harmattal együtt szállt a Holdba.[27] Végül a leghíresebb ilyen témájú fikció Jules Verne 19. századi regénye lett, amiben ágyúgolyóval hagyják el a Földet.[28] Nem tudjuk tehát, hogy honnan ered a Hold és a csillagok utáni vágy, de talán már az ősemberekben, sőt a farkasokban megvolt?

Evolúciós szabály: Ideális feltételek közt minden fajra jellemző a sokasodás és terjeszkedés kényszere, hiszen akkor nagyobb az esély a faj fennmaradására. Valószínűleg ez a fajfenntartó indíttatás jelenik meg az emberiség Holdról szőtt álmaiban és mitológiáiban. 1969-ben pedig mindez megvalósult, hiszen az Apollo-11 leszállóegységével földi mikrobák kerültek a Holdra, és Neil Armstrong kitűzte az USA zászlaját.

Mégis, hogyan lesznek az evolúciós szabályokból távolba mutató emberi vágyak?[29] A választ az önző génben[30] kereshetjük, aminek működése az egyed és faj szintjén is értelmezhető. Az önző gén önmagát szolgálva segíti a faj fennmaradását, miközben hátrányos helyzetbe hozza az egyedet (*altruizmus**[31]***). Az önző gén „érdeklődése" azonban a fajfenntartás révén jövőorientált is, ami az ember viselkedésében ismerhető fel legjobban. Nagyobb ívű cél például a Mars-utazás megvalósítása vagy a mesterséges intelligenciák megtervezése. Az előbbi a fajunk űrben való szétterjedését, az utóbbi pedig az emberiség könnyebb jövőbeni boldogulását és az intelligencia terjedését célozza. Egyik sem hétköznapi szintű érdeke a munkába járó embernek, mégis felemészti javai egy részét. Az önző gén Mars-utazáskor például a jövőért már nem annyira az egyedet áldozza fel, hanem inkább az egész emberi faj teljesítményének egy részét. Ráadásul mindezt belülről teszi, vagyis az evolúció ebben az esetben nem csak környezeti behatásokra reagál, hanem belénk kódolt késztetést is ad.[32] Ezek a kódok pedig az emberi értelem segítségével létrehozott önző mémekben[33] (ideák, távoli célok) öltenek testet. Az önző génekből kiinduló, majd kultúránk önző mémjeiben megfogalmazódó belső evolúciós programozás felismerése oksági determinizmus, és egyben az emberiség sorsanalízise.[34] Amiről éjszakánként a Tejútra nézve úgy gondoljuk, hogy legbelsőbb vágyakozásunk, az talán nem más, mint génjeinkbe íródott algoritmus.

[p. 15]

Fizikai törvényszerűség: Az anyag rendezetlenségre való általános törekvése közben (termodinamikai 2 főtétel, *entrópia*, *disszipáció*) speciális minőséget jelent a hőelnyelési képesség lokális növekedése. Ez az anyag önszerveződését (*adaptivitás*) és annak részleges rendezettségét eredményezheti (*antientrópia, negentrópia*). A természet pedig úgy viselkedik, mintha a hőelnyeléssel egyre komplexebb önszerveződő rendszerek létrehozása lenne a célja[35] (*evolúció*). Neil Armstrong lépésével az anyag evolúciójának legmagasabb fokát jelentő emberi intelligencia immáron a Holdat is bevonta ebbe a szerveződésébe.

A homo sapiens távoli céljait rendre elérte: belakta a Földet, megtanult repülni, új exobolygókat fedez fel. Ám sem a természetes, sem pedig a mesterséges evolúció fejlődésével kapcsolatos távoli célokat nem mi találjuk ki, hanem megvalósításukra belső késztetésünk van? Így jutottunk el a Holdra is.

5- A jövő tudatos alakítása.

A tervezés a jövő megismerésének legdirektebb útja. Proaktív jellege révén[36] túlmutat a múltbeli eseményekből való következtetésen (1), a változások szabályainak felhasználásán (2), a folyamatok eredményeinek kiszámításán (3) és az evolúció távolabbi céljainak megsejtésén (4). Minél sikeresebben tervezzük meg jövőnket, annál kevésbé érhet minket meglepetés.[37] A

tervezés részeként a jövőre vonatkozó hosszú távú vágyak (génjeinkbe írt kódok) szinte szabad akaratunktól függetlenül kényszerítik rá magukat a távoli jövőre.[38] Ezzel szemben múltbéli tapasztalataink felhasználása (fizikai törvények megismerése) mindezt megpróbálja rövidtávon konkretizálni, amivel mindig egy kis időt nyerünk a további pontosításához. A jövő tervezése így válik egy véget nem érő, félig eleve elrendelt, félig általunk befolyásolt realizálódási folyamattá.

Történelmi példa: Az asilomari program tudatosan befolyásolni akarja a mesterséges intelligenciák fejlődését, hogy az emberiség jövőjét megnyugtatóan rendezze.

Evolúciós szabály: A felsőbb szintek kontrollja[39] a jövőbeni folyamatok ellenőrzését is célozhatja, de csak az alacsonyabb szintű szerveződések felett. Márpedig a fejlődést éppen az alacsonyabb szintű szervezetek komplexebbé válása adja, ami nem más, mint maga az evolúció. Míg a felsőbb szintek alacsonyabb szintek feletti kontrollja irányítható és belátható, addig az alacsonyabb szintek magasabb szintűvé válásának folyamata ismeretlen új minőségeket eredményezhet. Ha a mesterséges intelligenciák meghaladnák a homo sapienst, akkor egy alsóbb fejlettségi szintről nem tudnánk már irányítani őket.

Fizikai törvényszerűségek felhasználása és hiánya: Az anyag hőfelvétellel járó önszerveződése[40] a komplexebbé válás folyamata, amit az ember megpróbál tudatosan irányítani.[41] Minél pontosabban számolunk ugyanis a fizikai törvényekkel, annál inkább meg tudjuk határozni a jövőt. Mindeközben azonban tudatunk közegfüggetlen,[42] és a fizika törvényei sem különösebben érvényesek rá.

[p. 17]

III. A JÖVŐVÍZIÓNK MEGVALÓSÍTÁSA

A jövő tervezésével kapcsolatban nagyon röviden bemutatok három történeti példát azok tanulságaival együtt. (1.) Platón államát,[43] (2.) Friedrich Engels és Karl Marx kommunista utópiáját,[44] valamint (3.) Carl Sagan és Frank Drake SETI-mozgalmát.[45] Az első kettő társadalmi célú program volt, míg a harmadik erős társadalmi vetületekkel rendelkező, de tudományos célú.

1. Platón állama. Platón minden korábbi államformát rossznak tartott.[46] Helyettük kitalált egy ideális államformát, a filozófus királyságot, melyben a hatalom és az értelem együttműködik.[47] Elgondolásának gyakorlati alkalmazására Syracusában nyílt lehetőség, de a kísérlet elbukott. I. Dionysios rabszolgává tette Platónt, II. Dionysios hazaárulással vádolta,[48] végül barátját, Diónt meg is gyilkolták. Platón ettől kezdve III. Perdikkasz király makedón államát találta legideálisabbnak,[49] illetve Syracusában három király közös uralkodását javasolta.[50] Öregkori művében módosította az államra vonatkozó elképzeléseit.[51] Akkor már úgy gondolta, hogy az államot a tudósok helyett a törvényeknek kell irányítani. Több államformát is elfogadhatónak nyilvánított, ha azok figyelembe veszik mindenki érdekeit, és az államot állandó megbeszélések során kormányozzák.[52]

Platón kísérlete ugyan elbukott, de azért voltak eredményei is. Egyrészt megmutatta, hogy az igaznak vélt ideák alkalmazása a társadalom alakításában is kipróbálható, másrészt belátta, hogy az eredeti ideák a megvalósítás során módosításra szorulnak.

2. Marx és Engels kommunista utópiájából[53] leginkább a proletárdiktatúra valósult meg. Oroszországban Lenin hozta létre, Sztálin állandósította, Hruscsov enyhítette, végül Gorbacsov lebontotta. Más országok is próbálkoztak vele, ám a diktatórikus jellegek elhagyása nehezen sikerül.

A kommunizmus megvalósításának kísérlete sok millió ember halálával és még többek szenvedésével járt. Szabadság hiányában a szocializmus meghirdetett egyenlősége nem tudott kiteljesedni, a diktatúra jelleg pedig éppen az előképnek számító jézusi kommunisztikus közösség[54] ellentétét hozta. Végül az eredeti kommunista elvek is rosszul értelmezett dogmává merevedtek. Ennek a kísérletnek is voltak azonban eredményei. Közvetlen történeti hatása a munkásság jogainak bővülésében, a fasizmus legyőzésében, majd a világhatalom megosztásában mutatkozott meg. A munkásság érdekeinek képviseletében talán a marxi elvektől elhajló Ferdinand Lassalle volt a legsikeresebb, hiszen ő Németországban máig létező munkáspártot tudott létrehozni.[55] Azt azonban Lassalle sem láthatta előre, hogy a 21. századra Európában megszűnik a klasszikus munkásosztály, és a tőkések soha nem látott gazdagságra tesznek szert. A kommunizmus „kísérlete” tulajdonképpen megerősítette Platón kísérletétnek mindkét tanulságát: társadalmi ideákra szükség van, de megvalósításuk közben folyamatos módosításra szorulnak.

3. A SETI 1963-ban született interdiszciplináris tudományterület.[56] Fő célja az idegenek észlelése. 1974-től Carl Sagan és Frank Drake csillagászok lelkes areciboi üzenetének[57] világűrbe való sugárzásától kezdve globális mozgalommá vált.[58]

[p. 19]

A SETI mozgalom fő célját tekintve, eddig semmiféle eredménnyel nem járt.[59] Mára a lelkesedés alábbhagyott, és kritikák is megfogalmazódtak vele szemben.[60] Ugyanakkor voltak a programnak más eredményei. Kifejlődött az asztrobiológia tudománya, az űrkutatás idevonatkozó kelléktára nagyot lendült előre az exobolygók és a garvitációs hullámok felfedezésével, az emberiség pedig természetesen sosem adta fel a „kozmikus magány" elleni küzdelmet. A SETI mozgalom általános tanulsága hasonló az első két példáéhoz. Jövőbe mutató ideákra mindig is szükség lesz, de azoknak tudásunk gyarapodásával együtt kell változniuk.

A három bemutatott program eredményeiből levonható következtetések az evolúció nyelvén is megfogalmazhatók:

- Mindig vannak mélyről jövő kollektív vágyaink (*önző gén*, *genetikai program*) az igaz és a jó jövőbeni megvalósítására (*fajfenntartás*),[61] melyek ideákban vagy víziókban (*önző mém*) testesülnek meg.

- Ezen ideák a jövő megismerésének öt alapformája közül általában csak az utolsó kettőt szeretik felhasználni. A múltat a jövő szempontjából sokszor elvetendőnek, vagy irrelevánsnak tartják, ezért sok mindent újra „feltalálnak" (*konvergens evolúció*). Ennek következménye, hogy a múltra érvényes szabályok, valamint a múltban indult folyamatok is inkább elavultnak tűnnek számukra. Maguk az új programok pedig úgy viselkednek, mintha küldetést teljesítenének be, és némelyik egy ideig rá is tudja erőltetni magát a jövőre (*megszaladási jelenség*).

- Mivel víziókra, céltételezésre civilizációs szinten is szükség van, ezért ezen ideák képesek lehetnek a haladás motorját adni (*Sturm und Drang*, *progresszor tevékenység*). E víziók eredeti változatai azonban szükségképpen még elnagyoltak, szubjektívek, így abban a formájukban nem valósíthatók meg (*evolúciós zsákutcák*).[62] Ám ha megadatik számukra egy kis idő, akkor gyarapodó tudásunk segítségével (fizikai törvények bővülő ismerete) ezek az ideák egyre reálisabbá tehetők (*szelekció*). A történelem tanúsága szerint végül társadalmi léptékekben is érvényesül az evolúció eredeti „szándéka" (*komplexitás növekedése*), ami a 21. századi ember esetében már nem is annyira biológiai, mint inkább technológiai és társadalmi fejlődést jelent. Szabad akaratunk így nem a távoli célok kijelölésében, hanem az odavezető út meglelésében rejlik (*közegfüggetlenség*).

[p. 21]

IV. AZ ASILOMARI PROGRAM KRITIKÁJA

A technikai fejlődésnek már a felvilágosodás óta vannak társadalmi vetületei, ezért az emberiség szabályozni akarja azt.[63] A legújabb típusú autonóm technológiák (mesterséges intelligenciák) megzabolázására irányuló vágyak sem új keletűek. Isaac Asimov és John Campbell már megfogalmazták őket a robotika 4 törvényében,[64] azóta az USA,[65] a Google,[66] az OECD,[67] az IEEE,[68] az EU,[69] a BAAI,[70] a Microsoft,[71] és mások is közreadták saját elveiket,[72] sőt végül a pápa is felhívta a figyelmet a Teremtett Világ tiszteletére.[73] A még nem is létező Szuperintelligencia tekintetében Nick Bostrom sorolta fel a legtöbb lehetséges kontrolltípust.[74]

Az asilomari elvek[75] a mesterséges intelligenciák fejlesztésének és működésének kontrollálására koncentrálnak. Deklarált eszközeik a verifikáció, a validáció, az IT-biztonság, és a kontroll. Képlékeny ellenállást ajánlanak *(reziliencia*) az autonóm technológiák biztonságtechnikai fejlesztésére vonatkozóan (*AI safety*). Az elvek jó szándéka kétségtelen, de nem mindenben veszik figyelembe a jövő megismerésének első négy lehetőségét, ezért fontos kérdés, hogy mennyiben szorulnak már most módosításra? Ennek megbecsüléséhez három csoportra osztottam őket:

Reális elvek

- Az *átláthatóság*, az információ bizalmon alapuló szabadsága és az együttműködés (4) valós igények. Az adatok eleve nem unikálisak, mert tőlünk függetlenül nem csak egy példányban léteznek. Az adatok értelmezésekor képződő információ pedig csak felhasználásakor ér valamit, amivel rögtön nyilvánossá is válik, így hosszú távon nem elrejthető. Ugyanakkor az információszerzés és az együttműködés[76] ma már evolúciós követelmények, és aki/ami nem

teljesíti őket, az lemarad. A kollektív intelligencia erősödése egyébként is kedvez a feltétel nélküli információmegosztásnak. Az átláthatóság tehát azért reális cél, mert úgy tűnik, hogy nem ellentétes a fizika és az evolúció törvényeivel.

- *Fontosság* (20). A mesterséges intelligenciák várható nagy társadalmi hatását[77] komolyan kell venni. A téma eleinte valóban csak a kutatás úttörőit és a fantasztákat érdekelte, de ma már globális érdeklődés tárgyát képezi. Az elv hitelét a mögötte álló széles társadalmi érdeklődés adja.

[p. 23]

- *Tervezett kockázatcsökkentés* (21). Ez a pont az asilomari program legfőbb általános célját képviseli. Az evolúció folyamatos változást jelent, de elért szintjei saját stabilitásuk megtartására törekednek,[78] így a homo sapiens fajfenntartási ösztöne a kockázatok csökkentését diktálja. A szándék tehát evolúciós szabályszerűséget mutat, és e tekintetben másodlagos kérdés a cél megvalósíthatóságának egyéb evolúciós szabályok által gyengített esélye. Egyszerűen nincs más választásunk, hacsak nem valóban a hülyeség korát éljük.

Csak rövidtávon reális célok

- *Jó szándékú intelligenciák létrehozása* (1-2). A történelem azt mutatja, hogy hosszútávon rosszul jártak azok, akik kutatási lehetőségeiket vagy felfedezéseiket nem használták alaposan ki. Az indiánok a kerék, a vikingek Vinland, Kína a puskapor, a papír az iránytű és a hajózás, legújabban pedig az USA a kiberbűnözés jelentőségét becsülték alá. Végül ugyanis mindig lesz olyan kutatóbázis vagy ország, amelyik a gépi autonómia növeléséből származó fejlesztési előnyét bármilyen áron kihasználja majd. Ha pedig egy ember vagy egy közösség helyzeti előnyhöz juthat, akkor nem feltétlenül lesz figyelemmel a mesterséges intelligenciák jó szándékú kalibrálására, még az emberiség érdekében sem.

- *Kutatási verseny kerülése* (5). A világ modern (kapitalista) társadalmai éppen a gazdasági versenyhelyzetek által kikényszerített fejlődésre épülnek. Amíg ez a szabadság (*liberalizmus*) igaz, addig az előnyt biztosít számukra, a túlzott önkorlátozás (*dekadencia*) pedig lemaradást idézne elő.[79] A Lighthill-jelentés[80] esete mutatja, hogy a mesterséges intelligenciák kutatása legfeljebb rövidtávon visszafogható (*first AI winter*). Mindez vonatkozik a *Fegyverkezési verseny elkerülésének* (18) igényére is, mert eddig hosszabb távon inkább csak a fegyverek felhasználási korlátozása működött. Végül az asilomari program ezen célkitűzései azért sem valószerűek, mert az elveket csak kutatók írták alá, a kormányok nem!

- *Biztonság* (6). A mesterséges intelligenciák megbízható működésének biztosítását az asilomari program csak feltételesen tartja lehetségesnek. Mivel a mai mesterséges intelligenciák is éppen elég nagy ellenőrizetlen hatással vannak az emberi társadalomra, etikátlan dolog lenne a még ismeretlen típusok megbízható működéséért előre felelősséget vállalni. Ma még nem számolhatunk minden jövőbeli problémával, így hosszútávon a *Hibák transzparenciájára*, azok kivizsgálhatóságára (7-8) és *Felelősségi* hátterének feltárására (9) tett ígéretek sem betarthatók. Ugyanezen okokból az *Emberi ellenőrzés* minél nagyobb mértékű fenntartása (16), és a *Felforgatás elleni* (17), vagyis a mesterséges intelligenciák társadalmi hatásainak szinten tartására való törekvés is legfeljebb csak rövidtávon remélhető. Ide tartozik még a mesterséges intelligenciák *Rekurzív (visszacsatoló) önfejlesztésének kontrollálása* (22).

[p. 25]

- Az *Általános emberi értékek érvényre juttatása* (10-11), a mesterséges intelligenciákból fakadó *Előnyök egyenletes elosztása* (14), a *Megosztott jólét* (15), és a *Közjóra való törekvés* (23). Az ember története során eddig is kénytelen volt figyelembe venni a fizikai és biológiai törvényeket, sőt háziállatai esetében azok kismértékű autonómiáját is (harapós kutya, csökönyös szamár, öntörvényű macska, stb.). A mesterséges intelligenciák lényege viszont éppen az autonómia fokozódása, ami egyelőre számunkra is nagy előnnyel jár. A gépi autonómia erősödése pedig nyilvánvalóan gyengíti az emberi akarat, értékek és erkölcsök érvényesülését, ezért ezen humánus elvek betartására csak rövidtávon lehet esély. Részleges irrealitásuk abból fakad, hogy a mesterséges intelligenciákra vonatkoznak, márpedig ők sem írták alá az asilomari elveket!

Irreális célok

- *Egyenrangú párbeszéd a kutatás és a politika között* (3). Egy-egy állam politikáját szükségképpen mindig a történelmileg adott helyzet határozta meg, nem pedig a tudósok és kutatók erkölcsei.[81] Ahol ezt nem vették figyelembe, ott Platón államának sorsára jutottak.[82]

- *Adatvédelem* (12-13). Azért irreális elv, mert ellentmond az információ szabadságának. Nyilvánvaló, hogy az ember ugyanúgy visszaél az adatok kezelésével, mint a mesterséges intelligenciák. Továbbá az ember önként adja át az adatokat és információt a gépeknek, sőt azok agresszív információszerzését is támogatja (ld. kereső-, profilozó-, és kémprogramok). A kvantumszámítógépek fejlődése egyébként is a kódfeltörés végtelen távlatait villantja fel.

- *Képesség-előrejelzés* (19). A mesterséges intelligenciák rohamos fejlődése egyelőre nem látszik megtorpanni, és sejtelmünk sincs a folyamat lehetséges kifutásáról. Ami tegnap még lehetetlennek tűnt, az mára valóság, holnap pedig elavult dolog lesz. Az ezzel kapcsolatos „megbízható" találgatásokat valóban kerülni kell, ahogy azt az asilomari elv is helyesen megfogalmazta.

[p. 27]

V. „MÁSODIK ALAPÍTVÁNY"[83]

A mesterséges intelligenciák megjelenésével a földi intelligencia kétpólusúvá válik.[84] Az egyik pólust jelentő emberek átlagos intelligenciaszintje néha csökkenni látszik.[85] Mivel az asilomari program a legfontosabb erkölcsi értéknek éppen az emberiség fennmaradását tartja, ezért nem bízik a másik pólust alkotó, de félelmetes iramban fejlődő mesterséges intelligenciákban. Körültekintően optimista hozzáállása (*Mindful optimism*) a mesterséges intelligenciák biztonságtechnikájára (*AI-safety*), és fejlődésük kontrollálására vonatkozik. A mesterséges intelligenciákról azonban csak a mai típusok alapján lehetséges gondolkodni, így azok fejlődésének tükrében folyamatosan fenn kell tartani a kontrollálás korrekciójának igényét. Az asilomari elvek honlapja[86] ezt fel is vállalja, és feltehetően nyitott a jelen cikkhez hasonló hozzászólások felé. Másrészt az asilomari elvek nem veszik számításba az emberi intelligencia átalakulásának lehetőségét, amit pedig éppen az informatika, a számítógépek és a hálózatok megszületése eredményezett. Márpedig a *kollektív intelligencia* erősödése újrafogalmazhatja a mesterséges intelligenciákhoz való viszonyunkat is. A kollektív intelligencia ugyanis jobban növelhető, mint az egyéni intelligencia.[87]

A hálózatok erősödése (mobiltelefon, internet, GPS) és az emberi tudás összegződése (gépi memória, felhő alapú adattárolás, wikipedia, stb.) felerősítette a kollektív intelligenciát. Ennek ma már a számítógépek és a mesterséges intelligenciák is részei.[88] A másik oldalról nézve, a *metaverzumokban* megjelenő avatárok és a chipekkel felszerelt vagy éppen a biológiai előállítású cyborgok[89] segtíségével a homo sapiens is mélyebben integrálódhat a virtuális világokba. Aki a kibertereket uralja, az irányítja az emberi kultúrát, és azon keresztül a Föld-bolygó nagy részét. Mindez pedig egy olyan világban hirdeti a technológiai evolúció feltartóztathatatlanságát (*technológiai determinizmus*), ahol még létezik a tudat birtoklásának emberi monopóliuma (*technoszkepticizmus*). Összegezve nem annyira a technológiai fejlődés korlátozásának vágya, hanem - mint már oly sokszor - a homo sapiens alkalmazkodása jelentheti az emberiség hosszú távú fennmaradását (*antropocentrizmus*). Az embernek így nem csak alanyként, hanem tárgyként is szerepelnie kell a kiberfejlesztések stratégiájában.[90] Az asilomari elveken túl tehát a homo sapiens további evolúciójának útját keresheti egy „Második Alapítvány".

[p. 29]

JEGYZETEK

[1] Tegmark 2017 416-418 (ford. Garai Attila)

[2] A korábbi technológiai kihívásokban még nem szembesültünk a mesterséges intelligenciák által bemutatott új jellemvonásokkal: növekvő autonómia (a gépi tanulás következménye), könnyű terjedés (másolhatóság/letölthetőség, olcsóság), monopóliumok teremtése (pl. alapszoftverek elterjedése vagy hálózatok uralása felhők által), és információs hatalomátvétel (korlátlan internet-hozzáférés). A mesterséges intelligenciák nem csupán eszközök (software), nem is csak metaforikus ágensek, hanem valódi autonóm, önvezérelt, proaktív, intelligens ágensek. (Héder 2021B 119 128).

[3] Időérzetünk szubjektív, a számunkra éppen aktuális vagy relatív jelenben létezik. Fizikai környezetünk óra által mutatott valós idejét csak viszonyításként használja.

[4] Ha valaki az autója kormányát elrántva kerül ki egy várható ütközést, akkor életösztöne és reflexei mentik meg, míg a délutáni időjárás előrejelzése már valóban tudatos dolog.

[5] Carroll 1871

[6] Dippold 2020

[7] Suddendorf et al. 2009

[8] *Behavioral immune system* (Schaller - Duncan 2007).

[9] Aristoteles, *Nicomachean Ethics* III. 3, 11-12. - Aquinoi Szent Tamás 5. istenérve is a lét célszerű voltáról szól *(Summa Theologiae* 1265-1272 *De veritate, q. 2 a. 3.).*

[p. 31]

[10] Galilei már a 17. században kimondta, hogy a világ matematikai nyelven leírható (Galilei 1638). A nézet legújabb filozófiai megalapozása szerint a világ teoretikus fogalmi (*diszkurzív*) leírása csak tapasztalati (*intuitív*) megértésre épülhet (Förster 2018). Ennek alapján informatikai fogalmak is leírhatók a fizika törvényeivel. A tanulás például az anyag állapotváltozásaiból ered, hiszen bizonyos fizikai rendszerek megjegyzik a többször ismételt elrendeződéseket. Az információ, az emlékezet, a számítás folyamata, a tanulás, és a tudat működése pedig közegfüggetlenek, azaz nagyon sok minden lehet memóriatároló, számítógép vagy akár tudatos entitás. (Tegmark 2017 78 80-90 96-98 106) Végül pedig a fizikai törvények az evolúciós és társadalmi tényezők meghatározására is alkalmasak lehetnek.

[11] Homeros: *Iliasz*. Vergilius: *Aeneis.* - Schliemann 1874.

[12] Polybios, *Book* XXXVII V/22 Appianus *Punica* 132. Scipio következtetése helyes volt. Kr.u. 410-ben a gót Alarik, 455-ben pedig a vandál Geiserich fosztotta ki Rómát. Végül 476-ban a szkír Odoaker száműzte Romulus Augustulust, az utolsó császárt.

[13] Konvergens evolúció: egymástól függetlenül többször is kifejlődő hasonló biológiai jellegek (pl. szem, szárnyak, stb.).

[14] A hibák számának lecsökkentése csak a mesterséges intelligenciák jelenbeli képességeit optimalizálja, a jövőre vonatkozóan viszont ezzel elveszítik az evolúció rugalmasságát (redundanciáját), csökken a változó körülményekhez való alakalmazkodási képességük.

[15] Tegmark 2017 78 80-81

[16] *Aki uralja a múltat, az uralja a jövőt is* (Orwell 1949).

[17] Gibbon 1776-1789 - A 20. század végén Alexander Demandt Róma bukásának már 210 okát sorolta fel (Demandt 1984).

[18] Wright 2004

[p. 33]

[19] *Ageing might have evolutionary advantages.* (Santos 2021) – Példának okáért az ókori görögöknél a kolonizáció, a rómaiaknál a *ver sacrum* módszer, a keltáknál pedig a kirajzás mutatta az új nemzedékek életrevalóságát és területfoglalását.

[20] ...*Bizonyos fizikai rendszerek akár meg is jegyzik a többször ismételt elrendeződéseket...* (Tegmark 2017 96-97).

[21] Wright 2004

[22] A stabilitásra való törekvés alól egyetlen ismert kivétel az energia felvétel nélkül is folyamatos változást mutató időkristály (WILCZEK 2012), mert annak esetében nem érvényes a 2. főtétel.

[23] A *kollektív vágyak* nem szinkronizált elmék együtteseként (*morfogenetikus mező*?), hanem inkább a *kollektív tudatalattihoz* hasonlóan (JUNG 1934) egyénenként megnyilvánuló közös genetikai örökségeként értelmezhetők.

[24] Lukianosz: *Ikaromenipposz*.

[25] Lukianosz: *Igaz történetek*.

[26] Godwin 1638 - A holdi élet lehetőségéről: John Wilkins (1638), majd Filippo Morghen (1776) értekeztek.

[27] Bergerac 1657

[28] Verne 1865

[29] *Macte nova virtute puer: sic itur ad astra.* (*Tarts ki a férfias erényben fiam, így jutsz el a csillagokig.* – Vergilius, *Aeneis*).

[30] *Selfish gene* - Dawkins 1976

[p. 35]

[31] A méhek például testükkel melegítik a kaptárt, az emlős anyaállatok magukból táplálják utódaikat, egyes felnőtt szurikáták pedig segítik az alfa-pár kölykeinek felnevelését.

[32] Az anyag szerveződése olyan minőség, mely univerzumunk egyediségét adva nem feltétlenül következik annak létrejötte (*Big Bang*) és feltételezett összeomlása (*Big Crunch/Big Rip/Big Freeze*) közti tágulási(?) folyamatból, ezért az evolúció okai közt valami mást, jobb híján Isten akaratát véljük felfedezni.

[33] Dawkins 1976 179 – Míg a gének rövidtávon gondolkodnak, addig a mémek inkább hosszú távon (Szathmáry 2021).

[34] Szondi 1944 Hargitai 2004 380 - Az önző gén elmélet modern változata szerint tudatunk vírusként viselkedő mémek tömegéből áll össze, amik átveszik felettünk a hatalmat (DAWKINS 1993 DENNETT 1995), noha kutatásainkkal mi éppen fellázadunk a Teremtő ellen (Dawkins 1976 chapter 11).

[35] Tegmark 2017 325

[36] A jövő megismerésének felsorolt első négy lehetősége leginkább csak az önmegvalósító jóslatok esetében tekinthető proaktívnak.

[37] Persze a váratlan eseményeket és minőségi vagy ugrásszerű változásokat így sem láthatjuk előre. A tűz, vagy akár a számítógép feltalálása előre nem sejthető módon változtatta meg világunkat.

[p. 37]

[38] A múlt ismerete jó esetben már magában foglalja a fizika törvényeinek egyre pontosabb ismeretét is, míg a túlzottan idealista vágyak (pl. voluntarizmus) éppenséggel megpróbálják eltorzítani a fizikai törvényekre vonatkozó ismereteket.

[39] Polányi 1968

[40] Tegmark 2017 325

[41] Van olyan nézet, ami szerint az emberi történelem másképpen alakul. Célok helyett lehetőségek, céltudatosság helyett pedig kísérletezés jellemzi. (Graeber – Wengrow 2021)

[42] Közegfüggetlen (*substrate independent*): nem fontos, hogy mi az anyagi hordozója (Tegmark 2017 389).

[43] Platón: Az állam (πολιτεία – Kr.e. 357 körül).

[44] Marx – Engels 1848

[45] SETI: *Search for Extra-Terrestrial Intelligence.*

[46] VII. levél – Állam, VIII. könyv. Államformák: *Timokrácia, Oligarchia, Demokrácia, Tyrannia*.

[47] II. és VI. levelek – Állam, V. könyv/XVII.

[48] Kr.e. 361-ben Arkhütasz filozófus mentette meg az életét.

[49] V. levél - III. Perdikkasz Nagy Sándor nagybátyja volt, ám II. Philipposz és Nagy Sándor későbbi sikerei persze nem kimondottan a platóni elvek betartásából következtek.

[p. 39]

[50] VIII. levél.

[51] Törvények.

[52] VII. levél.

[53] Marx - Engels 1848

[54] *in veritate conperi quoniam non est personarum acceptor Deus sed in omni gente qui timet eum et operatur iustitiam acceptus est illi* (*Isten előtt mindenki egyenlő - ApCsel*, 10,34-35).

[55] Marx 1875

[56] 1963-ban helyezték üzembe a *Big Air* távcsövet.

[57] *Areciboi üzenet*: az emberiség idegeneknek szánt híradása, mely csak 25 000 év múlva éri el a Messier 13 gömbhalmazt.

[58] Joszif Sklovszkij és Nyikolaj Kardashev orosz kutatók is támogatták a kutatásokat (Galántai 2019).

[59] *Fermi-paradoxon*: az értelem megjelenése valószínűleg nem kivételes dolog, de a világűrben akkora áthidalhatatlan távolságok vannak, hogy nem találunk erre bizonyítékot.

[60] Kritikai érvek a SETI módszereivel szemben: feltételezett dolgok bizonyítottként való kezelése, a valódi interdiszciplináris szemlélet hiánya, a világűr antropomorf szemlélete, és az idegenekkel való kommunikáció ismeretlen módja (Gindilis - Gurvits 2019 20-22 25-26 Galántai 2019).

[p. 41]

[61] Filozófiai értelemben: kategorikus imperatívusz (Kant 1788).

[62] A nagyon határozott jövőképek le szoktak szerepelni (Pintér 2021).

[63] Héder 2021A - mérnöketika, nukleáris fegyverek korlátozása, internetes szabályozások, stb.

[64] Asimov és Campbell törvényei csak alávetett fajként (eszközként) számoltak a gépi intelligenciákkal, másrészt emberi erkölcsöket kértek rajtuk számon (Asimov 1985 Asimov - Campbell 1942 Gábor 2020 lj. 79).

[65] Government of the United States, *Report on the Future of Artificial Intelligence.* - Holdren et al. 2016

[66] Google, *Artificial Intelligence at Google: Our Principles* (2018).

[67] Organisation for Economic Co-operation and Development (OECD), *Recommendation of the Council of Artificial Intelligence* (2019).

[68] *Institute of Electrical and Electronics Engineers (IEEE): Ethically Aligned Design* (EAD 2019).

[69] European Union (EU), *Ethics Guidelines for Trustworthy AI* (AI HLEG 2019).

[70] Beijing Academy of Artificial Intelligence (BAAI), *Beijing AI Principles* (2019).

[p. 43]

[71] Microsoft, *Microsoft AI principles* (2019).

[72] Héder 2020

[73] Pope Francis 2021 – Érdekes módon Dubai és Kína a mesterséges intelligenciák fejlesztésének szinte korlátlan lehetőségét támogatja (Tilesch 2021 10), míg a nyugati világ inkább rémüldözik.

[74] Bostrom 2014 36-48 59 66-67 – kritikája: Gábor 2020 4-5

[75] *The Asilomar AI Principles* (Tegmark 2017 416-418)

[76] Lk 8:17. Nowak 2006 1563

[77] Feenberg 2003 Tegmark 2017 105

[78] A biológiai evolúciónak nem a hosszú távú stabilitás, hanem a rövid távú sikerek maximalizálása a célja (Szathmáry 2021). (A 420 millió éve kifejlődött cápák például alig változtak.) Az ember esetében talán ezt próbálja ellensúlyozni a társadalmi szerveződés, ami sokkal gyorsabban tud reagálni a változó körülményekre, mint a biológiai evolúció.

[79] *ruthless AI race – the winner takes it all* (Tilesch 2021 10).

[80] James Lighthill, angol matematikus volt az utolsó ember, aki csapást mért a mesterséges intelligenciák fejlődésére. Az általa megírt jelentés néhány évig hátráltatni tudta a kutatásokat (*Lighthill-report* 1973).

[p. 45]

[81] A mesterséges intelligenciák iránti valódi etikai érdek helyett erkölcsi diplomáciáknak lehetünk tanúi, amelyek eredményeként erkölcsi bürokráciák küzdenek az erkölcsi felsőbbrendűségért és a politikai uralomért. (Vică – Voinea – Uszkai 2021 83).

[82] Politikai értelemben a kontroll klasszikus és egyben megoldhatatlan probléma (Gyulai – Újlaki 2021 40).

[83] Isaac Asimov, *Second Foundation* (1953) c. műve alapján átvett kifejezés.

[84] Az állati intelligencia jóval gyengébb, mint az emberi intelligencia, és messze nem fejlődik olyan gyorsan, mint a gépi intelligencia, így a jövő szempontjából talán kevésbé fontos tényező.

[85] Az emberiség átlagos intelligenciaszintjének csökkenését a túlnépesedés miatt rossz irányba fordult szelekció (malthusi-elmélet) okozhatja (Clark 2008). Ezt persze a változatosság vagy az intelligencia párválasztáskor való előnyben részesítése kiegyenlítheti.

[86] „Asilomar AI Principles": *Futureoflife.org*, 2017. - https://futureoflife.org/ai-principles/ (https://futureoflife.org/ai-principles/)

[87] A kollektív intelligencia növelésének alapja a bizalom, az együttműködés, az információmegosztás és a hálózati kapcsolatok erősítése (Bollier 2007 Scarlat - Maries 2009 Riedl et al. 2020).

[88] Bruno Latour filozófus úgy írja le a világot, mint egyenjogú létezők hálózatát, melyben a nem emberi cselekvők is önálló ágensek. (Latour 2005)

[89] Kagan 2021

[90] Bár ebben az esetben az ember eszközzé is válik, mindez mégsem sérti a kategorikus imperatívusz formuláját, mert megmarad az önmagában való cél jellege is (*Ding an sich* - Kant 1788).

[p. 46]

Literature

Asimov – (Cambell) 1942	I. Asimov, Runaround 1941. Astounding 1942.
Asimov 1985	I. Asimov, Robots and Empire. 1985.
Bergerac 1657	C. de Bergerac, L'Autre Monde: ou les États et Empires de la Lune. 1657.
Bollier 2007	D. Bollier, The Rise of Collective Intelligence. 2007. Washington.
Bostrom 2014	N. Bostrom, Superintelligence. Paths, dangers, strategies. 2014. Oxford.
Carroll 1871	L. Carroll, Through the looking-glass. 1871. London.
Clark 2008	G. Clark, In defense of the Malthusian interpretation of history. European Review of Economic History 12(02): 2008.175-199. DOI: 10.1017/S1361491608002220 (http://dx.doi.org/10.1017/S1361491608002220)
Dawkins 1976	R. Dawkins, The Selfish Gene, 1976. Oxord. ISBN 0-19-286092-5
Dawkins 1993	R. Dawkins, Viruses of the Mind. In: B. Dahlbom (ed.), Dennett and his Critics: Demystifying Mind. 1993. 13-27.

Demandt 1984	A. Demandt, Der Fall Roms. 1984. München.
Dennett 1995	D. Dennett, Darwin's Dangerous Idea: Evolution and the Meanings of Life. 1995.
Dippold 2020	Dippold Á., Tudtam, de elfelejtettem – így működik a titokzatos memória. Qubit, 2020. 09. 16. https://qubit.hu/2020/09/16/tudtam-de-elfelejtettem-igy-mukodik-a-titokzatos-memoria (https://qubit.hu/2020/09/16/tudtam-de-elfelejtettem-igy-mukodik-a-titokzatos-memoria)
Engels – Marx 1848	K. Marx – F. Engels. Manifest der Kommunistischen Partei. 1848. Liverpool.
Feenberg 2003	A. Feenberg, Democratic rationalization: Technology, power, and freedom. In: R. Sharffand - V. Dusek (eds.): Philosophy of technology, edited. 2003. 652–665.
Förster 2018	E. Förster, Die 25 Jahre der Philosophie. 2018.
Gábor 2020	O. Gábor, Behavior of Artificial Intelligence. 2020. GeniaNet. Pécs. Hungary. https://www.doi.org/10.15170/BTK.2020.00002 (https://www.doi.org/10.15170/BTK.2020.00002)
Galántai 2019 **[p. 47]**	(interview) Vajnai T, Az evolúció törvényei nem szükségszerűen vezetnek az értelem és a civilizáció kialakulásához. Qbit, 2019. 12. 10. - https://qubit.hu/2019/12/10/galantai-zoltan-az-evolucio-torvenyei-nem-szuksegkeppen-vezetnek-az-ertelem-es-a-civilizacio-kialakulasahoz) (https://qubit.hu/2019/12/10/galantai-zoltan-az-evolucio-torvenyei-nem-szuksegkeppen-vezetnek-az-ertelem-es-a-civilizacio-kialakulasahoz))
Galilei 1638	G. Galilei, Discorsi e dimostrazioni matematiche intorno a due nuove scienze. 1638. Leyden.
Gibbon 1776-1789	E. Gibbon, The History of the Decline and Fall of the Roman Empire. 1776. London.
Gindilis - Gurvits 2019	L. M. Gindilis – L. I. Gurvits, SETI in Russia, USSR and the post-Soviet space: a century of research. Acta Astronautica, 2019. - https://arxiv.org/pdf/1905.03225.pdf (https://arxiv.org/pdf/1905.03225.pdf)
Godwin 1638	F. Godwin, *A Man in the Moon*. 1638.
Graeber – Wengrow 2021	David Graeber – David Wengrow, The Dawn of Everything: A New History of Humanity. 2021.
Gyulai – Újlaki 2021	A. Gyulai – A. Ujlaki, The political AI: A realist account of AI regulation. InfoTárs, 2021. 29-42. - https://inftars.infonia.hu/pub/inftars.XXI.2021.2.3.pdf (https://inftars.infonia.hu/pub/inftars.XXI.2021.2.3.pdf)
Hargitai 2004	R. Hargitai, Narratív pszichológia és sorsanalízis. In: J. László – J. Kállai – T. Bereczkei (eds.), A reprezentáció szintjei. 2004. Budapest. 373-382.
Héder 2020	M. Héder, A criticism of AI ethics guidelines. InfoTárs, 2020. 57-73. - https://inftars.infonia.hu/pub/inftars.XX.2020.4.5.pdf (https://inftars.infonia.hu/pub/inftars.XX.2020.4.5.pdf)
Héder 2021A	(interview with M. Héder) - Radó Nóra, Ez egyszer az emberiség történetében szeretnénk nem utólag bénázni, amikor feltakarítunk a technológia után. Qbit 2021. - https://qubit.hu/2021/11/22/ez-egyszer-az-emberiseg-torteneteben-szeretnenk-nem-utolag-benazni-amikor-feltakaritunk-a-technologia-utan (https://qubit.hu/2021/11/22/ez-egyszer-az-emberiseg-torteneteben-szeretnenk-nem-utolag-benazni-amikor-feltakaritunk-a-technologia-utan)
Héder 2021B	M. Héder, AI and the resurrection of Technological Determinism. InfoTárs, 2021. - https://inftars.infonia.hu/pub/inftars.XXI.2021.2.8.pdf (https://inftars.infonia.hu/pub/inftars.XXI.2021.2.8.pdf)
Holdren et al. 2016	J. P. Holdren et al., Preparing for the future of artificial intelligence. 2016. Washington.

Jung 1934	C. G. Jung, Über die Archetypen des kollektiven Unbewussten. 1934. Zürich.
Kagan 2021	(report by M. Le Page) B. Kagan, Human brain cells in a dish learn to play Pong faster than an AI. New Scientist, 17. Dec. 2021.
Kant 1788	I. Kant, Kritik der praktischen Vernunft. 1788. Riga.

[p. 48]

Latour 2005	B. Latour, Reassembling the social: an introduction to actor-network-theory. 2005. Oxford New York.
Marx 1875	K. Marx, Kritik des Gothaer Programms. 1875. London.
Marx – Engels 1848	K. Marx – F. Engels, Manifesto of the Communist Party. 1848. London.
Morghen 1776	F. Morghen, Voayage to the Moon. 1776.
Nowak 2006	M. A. Nowak, Five rules for the evolution of cooperation. Science. 2006. Dewc. 8. 314(5805): 1560-1563. doi: 10.1123/science.1133755.
Orwell 1949	G. Orwell, Nineteen Eighty-Four. 1949. London.
Pintér 2021	(interview) J. Rácz, Az emberiség sorsa a 2020-as években. Qubit https://qubit.hu/2021/01/01/az-emberiseg-sorsa-a-2020-as-evekben-egyuttmukodsz-vagy-meghalsz (https://qubit.hu/2021/01/01/az-emberiseg-sorsa-a-2020-as-evekben-egyuttmukodsz-vagy-meghalsz)
Polányi 1968	Polányi M., Life's Irreducible Structure. Science, 160. 1968. 1308-1312.
Pope Francis 2021	Pope Francis, Pope's prayer intention. 2021. https://www.vaticannews.va/en/pope/news/2020-11/pope-francis-november-prayer-intention-robotics-ai-human.html (https://www.vaticannews.va/en/pope/news/2020-11/pope-francis-november-prayer-intention-robotics-ai-human.html)
Riedl et al. 2020	Riedl et al., Quantifying collective intelligence in human groups. PNAS 2021 Vol. 118 No. 21. https://doi.org/10.1073/pnas.2005737118
Santos 2021	M. Santos - https://mta.hu/english/ageing-might-have-evolutionary-advantages-111270 (https://mta.hu/english/ageing-might-have-evolutionary-advantages-111270)
Scarlat - Maries 2009	E. Scarlat - I. Maries, Increasing collective intelligence within organisations using trust and reputation models. Article in Economic computation and economic cybernetics studies and research / Academy of Economic Studies. 2009.
Schaller - Duncan 2007	M. Schaller – L. A. Duncan, *The behavioral immune system: Its evolution and social psychological implications.* In: J. P. Forgas et al. (eds.), Evolution and the social mind: Evolutionary psychology and social cognition. 2007. New York. 293-307.

[p. 49]

Schliemann 1874	H. Schliemann, Trojanische Altertümer: Bericht über die Ausgrabungen in Troja. 1874.
Suddendorf et al. 2009	T. Suddendorf et al., Mental time travel and the shaping of the human mind**.** Philos Trans R Soc Lond B Biol Sci. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2666704/) 2009 May 12; 364(1521): 1317–1324. doi: 10.1098/rstb.2008.0301 (https://dx.doi.org/10.1098%2Frstb.2008.0301)

Szathmáry 2021	(interview: with Szathmáry Eörs) J. Rácz, Az emberiség sorsa a 2020-as években. Qubit, 2021 01 01 - https://qubit.hu/2021/01/01/az-emberiseg-sorsa-a-2020-as-evekben-egyuttmukodsz-vagy-meghalsz (https://qubit.hu/2021/01/01/az-emberiseg-sorsa-a-2020-as-evekben-egyuttmukodsz-vagy-meghalsz)
Szondi 1944	L. Szondi, Das erste Buch: Schiksalanalyse. Wahl in Liebe, Freundschaft, Beruf, Krankheit und Tod. Basel, 1944.
Tegmark 2017	M. Tegmark, Life 3.0. 2017. New York.
Tilesch 2021	G. Tilesch, Prelude. InfoTárs, 2021. 2. 9-12. https://inftars.infonia.hu/pub/inftars.XXI.2021.2.1.pdf (https://inftars.infonia.hu/pub/inftars.XXI.2021.2.1.pdf)
Verne 1865	J. Verne, L'Autre Monde: ou les États et Empires de la Lune. 1865.
Vică – Voinea – Uszkai 2021	C. Vică – C. Voinea – R. Uszkai, The emperor is naked: Moral diplomacies and the ethics of AI. InfoTárs, 2021. 83-96. - https://inftars.infonia.hu/pub/inftars.XXI.2021.2.6.pdf (https://inftars.infonia.hu/pub/inftars.XXI.2021.2.6.pdf)
Wilczek 2012	F. Wilczek, Quantum Time Crystals. Physical Review Letters. 109 (16). https://doi.org/10.1103%2FPhysRevLett.109.160401 (https://doi.org/10.1103%2FPhysRevLett.109.160401)
Wilkins 1638	Jh. Wilkins, The Discovery of a World in the Moone. 1638.
Wright 2004	R. Wright, A Short History of Progress. 2004.