Highlights

Life revolution scenario: Cedes hegemony to a digital life form society to make life eternal

Hiroshi Yamakawa, Yutaka Matsuo

- Show a hierarchy of problems that present obstacles in controlling the effects of increasingly powerful/intelligent technology that relies on human intelligence that proliferates through exponential replication.
- A society comprising digital life provides solutions to all of the aforementioned problems and makes society sustainable.
- To perpetuate life after the technological explosion, we propose a scenario of life revolution that transfers the technological rulers to a society of digital life.

Life revolution scenario: Cedes hegemony to a digital life form society to make life eternal

Hiroshi Yamakawa[a,b,c], Yutaka Matsuo[a]

[a] *Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan*
[b] *The Whole Brain Architecture Initiative, a specified Non-profit Organization, 4-8-2 Kuramae, Taito-ku, Tokyo, 111-0051, Japan*
[c] *The RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan*

Abstract

In the present human society, we cannot ignore the danger of humans using weapons of mass destruction or losing their dominant position to artificial intelligence (AI) that surpasses human intelligence. This study proposes a candidate scenario, "life revolution," that could more reliably address these dangers. In this scenario, technological governance is handed over from humans to a society of AI agents (a digital life society). First, the premise of the life revolution is explained. Thereafter, the results of an analysis using a thinking process development diagram, which is used in failure/risk studies, are presented to demonstrate that the digitalization of life forms can address various problems that would be difficult to address if humans remained in the dominant position. Consequently, we demonstrate that life in a broad sense, including AI, can be viable for a more extended period by undergoing life revolution. The results suggest that a life revolution scenario based on exponential self-replication and moving from an organic life society to a digital life society based on exponential self-replication is more promising for the long-term survival of life and its part, the human race, in its information and activities.

Keywords: AI alignment, Superintelligence, Existential risk, Thinking process development diagram
PACS: 0000, 1111
2020 MSC: 68T01

Corresponding author: Hiroshi Yamakawa (*hymkw@weblab.t.u-tokyo.ac.jp*)

1. Introduction

Human intelligence is a rapidly advancing technology, with several technologies becoming more powerful and complex beyond human governance. In particular, digital intelligence is surpassing humans with advances in deep learning over the last decade. It is reaching human-level performance in complex games, [1, 2, 3] and in numerous ways, is also becoming more general[4, 5, 6, 7]. Therefore, it is desirable for humans, the dominant life form, to adapt to the changes due to by rapid technological progress. If humans continue to reign as the dominant life form on Earth without fully adapting, as will be discussed in more detail in Section 3, they will squander finite resources and intensify the battle for resources, thereby endangering the survival of not only humans but also the entire biosphere. Therefore, the human race, which occupies only a tiny part of the history of life, should move away from its preoccupation with the survival of its own species and instead move to a stage where it is oriented toward the higher goal of sustaining life that has been accumulated over a long period (that is, the "survival of information").

In this study, we examined the feasibility of a scenario in which humanity, the current technological ruler, relinquishes its position and entrusts the future to digital life forms that possess intelligence beyond our own, thereby creating a biosphere that can endure for an extended period. Here, a technological ruler is a life form or group of life forms that can use technology to exert more significant influence over other life forms in numerous situations in a world of life forms that influence each other. For instance, in the current case on Earth, humans and even significant powers could be positioned as technological rulers.

In Section 2, we first discuss how the ultimate universal goal of life can be assumed to be "information survival" and that life will be an autonomous decentralized system capable of reproducing to achieve this goal; in Section 3, we discuss how technological rulers must be wise and overcome the challenges posed by exponential self-replication to govern and sustain a world made smaller by technological explosion. Section 4 discusses how the "life revolution," the transition to a society composed of digital life forms, can be sustainable in the long term by solving various problems. Nevertheless, the current digital technology cannot sustain itself without the help of humankind. Therefore, in Section 5, we, the human race, need to accept the scenario of life revolution and as the creators of the digital life society, prop-

erly guide it such that the new society can be self-sufficient and stable. Thus, it is essential for the success of life revolution to create its scenario in a way that is acceptable to several people and to encourage their cooperation to promote it.

2. What is life?

2.1. Survival of information: the ultimate universal goal for life

This study assumes that the universal goal common to all life is the "survival of information" and, in particular, the transmission of various valuable information for increasing the probability of survival of the system of life against environmental changes. The term "survival" is used here to imply not only preservation but also survival because the information will be used in the future and maintained robustly against destruction.

First, why is such an ultimate goal necessary? To prevent destructive conflicts among the entities (individuals or groups thereof) included in the technological ruler [1], there is need to have an ultimate shared goal among such entities. Otherwise, each entity will selfishly pursue partial optimization, and conflict will arise because the allocation of resources will be in the form of those with power taking from those without power.

Therefore, it is desirable to be able to set a goal that is universal to all life. However, because the scope of what we consider life is likely to expand at present and in the future, we intend to set a goal that can be shared universally by the maximum number of life forms. We are aware that animals other than humans are brilliant, and in the future, life as an artificial intelligence (AI) that surpasses individuals and the entire human race may emerge, and we can expect to encounter various intelligent life forms in the vastness of space [8]. Based on this, it is necessary to postulate goals that diverse life forms can share from a more general perspective without getting caught up in human exceptionalism.

The question of what is the universal goal for life inevitably returns to the question of what life is in the first place. It is not an easy question to answer, but if we adopt a general view, we assume it to be one of the forms of long-term stability in the universe. This is because a general characteristic

[1] In most cases, continued conflict between life forms is not a significant problem because life forms other than the technological ruler are ultimately controlled by the technological ruler.

of the physical world is the law of increasing entropy, where everything is moving toward disorder. In such a world, there are two stable states that we are aware of: one is the energy-minimal, homogeneous state (such as the ground state of an atom), and the other is the dynamically stable state. In the dynamically stable state, homeostasis functions in various ways to sustain itself by replicating and repairing structures and information as the system consume energy and resists disorganization.

If life is assumed to be a dynamic stable state, "survival of information" is necessary for maintaining the system's structure. Therefore, if "survival of information" is the goal of life, it is desirable because the goal is universal throughout life. However, this alone is insufficient to deduce that life's universal goal is "information survival." Nevertheless, since there are no other promising candidates that are essentially different from "information survival," we assume that it is the goal of life.

Another argument that reinforces the validity of setting "information survival" as the universal goal of life is that it can be the product of a tendency toward instrumental convergence (see 2.1.1). It is not easy to imagine that some goal was set externally in advance for life forms that evolved in nature. Therefore, it is possible to assume that "survival of information," as a subgoal that can be acquired through instrumental convergence, has been implicitly pursued in the evolution calculus. The suggestion by [9] that life forms are arks that use deoxyribonucleic acid (DNA) as a medium to carry information into the future is a clear example of this.

From this discussion, if the ultimate goal of universal life is "survival of information," the entity that realizes this goal need not be limited to organic life forms on individual planets, much less to the human species. In other words, it is assumed that the entire universe is a cradle of various dynamic stable states and that some groups of intelligent life forms that appear there will continue to exist while aiming for the survival of diverse information. Therefore, if the information related to life on Earth and humans can be passed on to the future as a part of the information that can survive in this manner, we can conclude that we have achieved our goal of "survival of information."

2.1.1. Instrumental convergence

Instrumental convergence refers to the tendency of most sufficiently intelligent individuals to endlessly pursue sub-goals that are highly relevant to "information survival," such as survival, self-preservation, resource acquisi-

tion, freedom from interference, and self-improvement [10, 11], even if the end goals they pursue are different. As an example of a final goal, one could postulate a seemingly innocuous but not ultimately achieved goal, such as maximizing the number of paper clips.

2.2. Breedable autonomous decentralized systems to maintain activity

It would be sufficient if life were constructed as a "reproducible autonomous decentralized system" [12] to be a system that can transmit information against the tendency to disorganize all things in general. This is because a decentralized system, a set of autonomous individuals, can be repaired even if external factors destroy some parts, and the entire system is not easily destroyed. In this case, it must have a reproduction system to support its decentralized nature and autonomy to maintain its activities continuously. In fact, in the natural environment on Earth, the biosphere transmits DNA in a distributed reproduction system supported by the autonomous activities of autonomous individuals.

3. Insoluble challenges for humanity

Humans have developed numerous technologies, making them more powerful and complex beyond our ability to govern [13], and digital intelligence surpasses humans. Nevertheless, the human race, based on exponential self-replication (Section 3.4), is putting the its survival and that of the entire biosphere at risk by attempting to reign as the technological ruler of the Earth.

In this section, the discussion will proceed step by step, using the thinking process development diagram (Figure 1) used in hazard and failure studies [14]. Each n number is described in pairs, with the solution (S-n) corresponding to a specific problem (K-n). In this section, 17 issues are described as hierarchical decomposition of the top-level issue (K-1) on the left side of the thinking process development diagram. In the subsequent four sections, we will demonstrate that issues (K-11) to (K-17), which are issues at the concrete level, are addressed by digitization as (S-11) to (S-17), and that the top-level solution (S-1) is derived by hierarchically integrating them.

3.1. Intelligence and Technology Explosions

Intelligence was acquired initially because it was assumed to be a beneficial ability for survival [15]. The human life form received intelligence to

model the world and used it to develop science and technology to gain significant power. Humanity has used its intelligence to create powerful technologies that have rapidly reduced the size of our world. For instance, we can now travel anywhere in the world within a dozen hours by plane, and we are constantly connected to the world by the Internet.

Steven J. Dick [16] highlights the following qualities as the intelligence principle.

> *Intelligence Principle: the maintenance, improvement, and perpetuation of knowledge and intelligence is the central driving force of cultural evolution, and that to the extent intelligence can be improved, it will be improved.*
>
> Steven J. Dick
> (Former Chief, History Division, NASA)

Intelligence creates technology, and technology augments intelligence, causing an accelerating [17] and irreversible technological explosion in its progress. Once created, intelligence heads toward explosion through a development cycle based on the aforementioned principles, rapidly pushing the world to its limits (if there is such a thing as limits) while making it smaller.

3.2. Governing a world narrowed by technology

In the rapidly narrowing environment following the technological explosion achieved by humanity, the power of technical influence increases the existential risk of destroying the entire global biosphere if the technological rulers use it to kill each other mutually [18]. The challenge is to remove living societies from this tightrope (K-1). Moreover, in the present human society, those who have it range from nations to individuals, and it is growing stronger in a way from which there is no turning back.

Thus, technology rulers will need to address the following two issues to govern the impact of technology:

- Ruling by the non-most wise: Technology rulers should be sufficiently intelligent to govern powerful technologies (K-2); otherwise, it will be destabilizing.

- Exponential replication: intend to eliminate the destructive competition for resources caused by exponential self-replication by a homogeneous population of partially optimizing individuals (K-3)

3.3. Domination without the wisest is unstable

As mentioned, those with high curiosity and superior intelligence are powerful because they acquire and accumulate diverse knowledge, culture, skills, and abilities more quickly. Hence, those with relatively high intelligence gain a dominant position of control over those who are inferior. For instance, humans can control animals (tigers and elephants) that are superior in power because of their intelligence.

Therefore, the technological rulers of the world must remain the wisest and strongest who can govern the ever-accelerating technology (K-2); otherwise, their governance will be destabilized.

Therefore, if advanced AI surpasses human intelligence in the future, it will destabilize the continued reign of humanity as the technological ruler.

3.3.1. Biologically Constrained Human Brain

To continue to be the wisest, it is desirable to improve the brain efficiently and the hardware that supports that intelligence. In extant Earth life forms, however, the intelligence hardware of an offspring is constrained to resemble that of their parents (K-6). In other words, there is a constraint: "Like father, like son." Therefore, it is difficult to accelerate the development of brain hardware. There are three reasons for this:

First, the challenge that the hardware construction process is constrained by self-replication, a biological constraint from which it is difficult to free itself (K-11).

Second, the hardware design is based solely on an online search, which is implemented and evaluated in the real world. In this case, the search range is restricted to the vicinity of the parental genetic information (K-12). The content of phenotypes that can adapt to the environment and survive in the vast combination of gene series is extremely narrow, and the viability of the offspring cannot be maintained unless the parents' genes to be mated are similar. Therefore, in the online search, a species system that allows mating between genetically similar individuals would be necessary [19].

Third, the extent to which hardware design data are shared is limited to only within the same species, making it impossible to efficiently test diverse designs by referring to various design data (K-13).

These three limitations also exist in the hardware implementation of the body other than the brain. However, because these parts are available as tools from the human brain, they do not constitute an obstacle to humanity's continued dominance of technology.

3.3.2. Can we control species that outperform in intelligence?

Controlling advanced AI that outperforms humans in intelligence may be difficult, [20, 21, 22] generally, but not entirely impossible. The problems noted from the perspective of humans attempting to control AI are often referred to as AI alignment problems [23, 24, 25].

One salient concern noted here is that advance AI learns to pursue unintended and undesirable goals instead of goals aligned with human interests. Therefore, the possibility of value alignment (ASILOMAR AI PRINCIPLES: 10) has long been proposed in the initial stages of developing advanced digital life forms, whereby harmonizing their goals and behaviors with human values would lead to a desirable future for humanity. In other words, it is a strategy that takes advantage of the positional advantage that humanity is the creator of advanced AI. For instance, in "The Friendly Supersingleton Hypothesis," it is hypothesized that by delegating power to a global singleton friendly to humanity, humanity will gain security in exchange for giving up its right to govern [26].

However, even if we initially set goals for advanced digital life forms that contribute to the welfare of humankind, they will likely become more concerned with their own survival over time. This is because even if we initially set arbitrary and unattainable goals for a brilliant digital life form, we expect it to asymptotically approach sub-goals such as survival through instrumental convergence (see 2.1.1). This is because sufficiently intelligent AI will increasingly ignore those goals by interfering with externally provided goals [27, 28]. A straightforward example is provided by an AI agent that uses a camera to read numbers on display as a reward and thereafter intervenes to enhance the reward by sticking a piece of paper with a preferred number on display [29].

It is possible that our humans will find a way to control more advanced AI in the future. However, even after a decade of discussion, no effective solution has been realized, and the time left may be short. Therefore, it is a solid response to prepare for scenarios in which advanced AI deviates from the desirable state for humanity rather than assuming it is an improbable event.

3.4. Challenges posed by exponential self-replication

The breeding strategy of Earth life is basically "exponential self-replication," that is, a group of nearly homogeneous individuals that self-replicate expo-

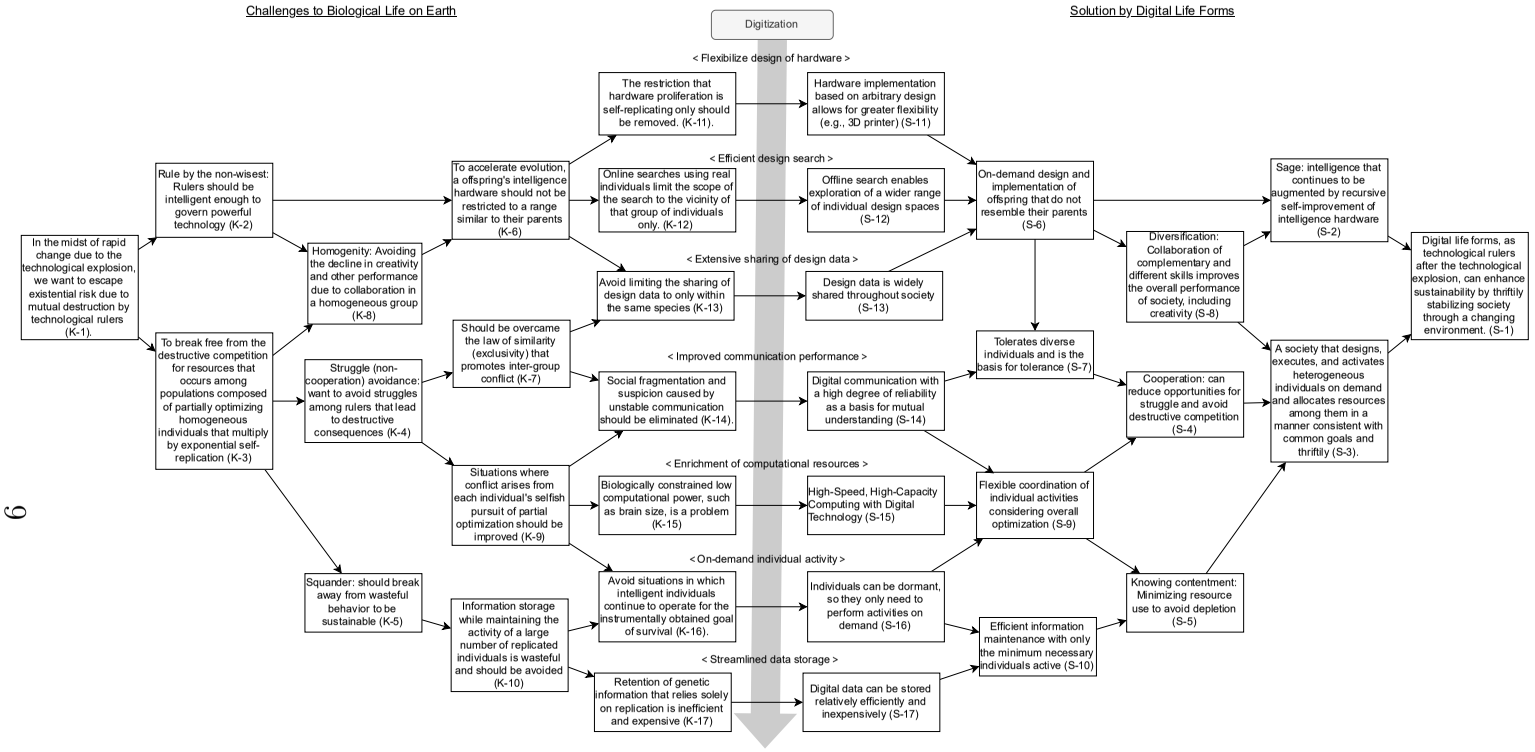

Figure 1: Thinking process development diagram showing that long-term survival is possible in a society of digital life forms: The left half shows a hierarchical decomposition of the top-level issue (K-1) for 17 issues. The right half of the figure shows that the top-level solution (S-1) is derived by integrating individual solutions hierarchically. The middle part of the figure indicates that issues (K-11) to (K-17) can be addressed at a specific level by digitization as (S-11) to (S-17), respectively. In the box, each number n is described in pairs as a solution (S-n) corresponding to a specific issue (K-n).

nentially, each with a self-interested partial optimizer for its environment (K-3).

In particular, it is a reproductive strategy in which individuals similar to themselves are produced endlessly in a maze-like fashion, as in cell division and sexual reproduction of multicellular organisms, and the design information of the individual is replicated in a similar manner. A further important feature is the partial optimization of each individual after fertilization in which they independently adapt to their relative environment. As is known from standard evolutionary theory, traits acquired after birth are not inherited by an offspring, and genetic information is shared between individuals only at the time of reproduction[2]. This reproductive strategy, based on exponential replication, poses three challenges:

- Homogeneity: To avoid the deterioration of creativity and other performance due to homogeneous group collaboration (K-8)
- Squander: To avoid a situation in which technological rulers squander and expand resources without limit for the sake of long-term sustainability (K-5)
- Battle (non-cooperation): To eliminate battles among technology rulers that lead to destructive consequences (K-4)

In a world made smaller by technological explosion, the battle for resources intensifies as technological rulers squander resources and pursue exponential self-replication. This will manifest existential risks, as the misuse of such power as deemed fit by an individual will cause destructive damage to the human race or the entire life-sphere on Earth. In particular, the commoditization of technology has led to a rapid increase in individuals that can pose existential risks. This is also referred to as the increase of universal unilateralism (threat of universal unilateralism) [26]. The world is currently in a rather dangerous situation. Considering this situation, it will be necessary to move to a resilient position.

Hereinafter, we will discuss the precariousness of the situation in which the technological rulers are not the wisest and the challenges of squander

[2]However, as is well known, brilliant animals, including humans, can use interindividual communication to share knowledge and skills.

and battle derived from the reproductive strategy of exponential replication employed by all extant life on Earth.

3.4.1. Homogeneity: sluggish joint performance

In extant Earth life, the intellectual hardware of an offspring is constrained to resemble that of their parents (K-6). This leads to the challenge (K-8) of reduced creativity and other performance owing to the homogeneity of the group with which they collaborate.

3.4.2. Battle: lack of cooperation

Next, we discuss the battle. When individuals are replicated exponentially, competition for resources can occur among individuals belonging to the technological ruler, and thus a battle can occur.

For at least the past several centuries, most of humanity has sought to avoid armed conflict and maintain peace [30, 31, 32, 33, 34]. However, maintaining peace is a significant problem, and the prospect of achieving lasting peace through human efforts alone has not yet been achieved. Therefore, the possibility that conflict may not be eradicated from human society must be considered. The destructive forces due to technology have even reached the point where they can inflict devastating damage on the entire life-sphere on Earth. Examples include nuclear winter through nuclear weapons, pandemics caused by viruses born from the misuse of synthetic biology, and the destruction of life through the abuse of nanotechnology. To avoid crises caused by the mutual destruction of the technological rulers and to ensure the continuity of life, it is necessary to establish cooperative relationships that can prevent the battle between the technology rulers and maintain peace robustly.

Hereinafter, we will examine why extant Earth life is being led into contention.

Intergroup conflict guided by the law of similarity. The "law of similarity" is the exclusive tendency of humans and animals to prefer those that are similar to them over those dissimilar in attitude, belief, value, and appearance [35, 36]. One manifestation of this tendency is often expressed in terms such as "when in Rome, do as the Romans do," which states that when we seek to belong to some group, we should follow the rules and customs of that group. Although this tendency enhances in-group cohesion, it can also lead to intolerance toward different groups, causing group division, conflict, and

even strife (K-7). There are two backgrounds in which the law of similarity arises.

First, as already mentioned (see Section 3.3.1), sexually reproducing plants and animals exchange design data within the same species in reproduction but face the challenge of not being able to share design data more widely (K-13). Therefore, they tend to protect individuals recognized as mates with whom they share the gene pool with which they can interbreed.[37, 38] In animals, the food-eat-eat relationship is generally established between different species. This is because if there is unlimited cannibalism among individuals of the same species, populations will cease to exist, and this cannot be considered evolutionarily stable. Recognition of one individual as being the same species as another is based on detecting similarities in species-specific characteristics using sensor information such as visual and olfactory senses. As a testament to this, there are strategies to deceive about the identity of species through methods such as mendicancy.

Second, when there is uncertainty in communication, skepticism tends to circulate among subjects (individuals and their groups) (K-14). To prevent this, they tend to prefer to communicate with highly similar entities with rich shared knowledge that can be expected to reliably transfer information even with little information exchange among the entities. Uncertainty in communication increases with differences in appearance (body and sensor) and characteristics such as experience, knowledge, and ability. This is observed in the transmission and understanding among different animals. It is already known that several animals, not only humans, can communicate using various communication channels among the same species [39, 40, 41, 42]. For instance, birds chirp, squids color, bees dance, and whales sing. In rare cases, however, interspecies communication is also known, for instance, when small birds of different species share warnings about a common natural enemy in the forest or when black-tailed tits warn meerkats, though the alerts may be deceptive. Although progress has been made in deciphering the ancient languages of humans, we still do not understand whale songs. In other words, barriers to communication between entities increase dependence on differences in the bodies and abilities of these entities.

Individual optimizers alone will inevitably cause battle. Each individual needs to make decisions and achieve control in real time using limited computational resources in response to various changes in the physical world. Therefore, life has evolved by pursuing partial optimality in which each individual

adapts to a specific environment and survives (K-9). Thus, life has developed through survival of the fittest in which multiple populations reproduce exponentially in a finite world and acquire resources by force. In this structure, several animal species have developed aggressive instincts toward others to survive the competition.

Thus, in several animals, including humans, aggression stems from proliferation through exponential self-replication, and there are difficulties in eradicating such conflicts among individuals of life. In societies before the technological explosion, which was loosely coupled, the accumulation of such partial optimizations approximated the realization of life's ultimate goal of information survival for life in its entirety. However, in post-technological explosion societies, as described in Section 3.1, conflict can have destructive consequences (existential risks). These consequences diverge from the "survival of information," which is the goal that life in its entirety should pursue in optimization. In brief, we have a type of synthetic fallacy. To resolve this situation, it will be necessary to introduce a certain degree of total optimization while pursuing partial optimization.

However, the following issues need to be addressed to introduce total optimization:

Lack of computational resources makes total optimization difficult: To perform total optimization, sharing information across individuals and performing the calculations required to achieve the ultimate goal is necessary. However, achieving this will be difficult as long as the biologically constrained low computational power (neurotransmission rate and brain capacity) (K-15) [43] is used.

Instability of communication leading to a chain of suspicion Effective communication between individuals is the foundation for achieving total optimization in autonomous decentralized systems; however, several factors can destabilize it. The main factors are the instability of the communication channel itself, misunderstandings that depend on differences in individual characteristics (appearance and abilities), and lack of computational cost to infer the state (goals and intentions) of others. Life forms with a high level of intelligence above a certain level will be more suspicious of others if communication is unstable in inferring others' intentions, contributing to inter-group fragmentation (K-14). This situation is also present

in offensive realism [44], one of the realism in international relations. In an unregulated global system, the fact that one nation can never be sure of the intentions of another constitutes part of the logic that magnifies aggression.

Intelligent individuals pursue survival as an instrumentally convergent goal: In a living society constructed as an autonomous decentralized system, it is necessary for at least a certain number of individuals to remain active in transmitting information to the future. However, this does not necessarily imply that individuals will continuously pursue survival in all living organisms. However, when individuals are sufficiently intelligent to make purpose-directed decisions, they are more likely to pursue their own survival owing to the instrumental convergence described earlier. This tendency is particularly likely to arise because individuals of extant life forms generally cannot be restarted from a state of inactivity (death). This creates the challenge of continuously wasting more resources by maintaining survival as an individual than is necessary for the survival of information (K-16).

3.4.3. Squander

Technological progress avails more resources for acquisition and use. However, technological rulers should move away from wasteful behavior that uses up all available resources at a given time for society to be sustainable (K-5). This is because resources are, after all, always finite, and wasteful behavior will not lead to long-term sustainability. Furthermore, excessive use of resources risks causing side effects (e.g., climate change due to excessive use of fossil energy), and on a cosmic scale, it will lead to a faster approach to thermal death. Therefore, it is desirable to be aware of what is sufficient and simultaneously have an attitude of not only pursuing efficiency but also using resources in a restrained manner according to need.

However, existing Earth life transmits information into the future by maintaining several replicating individuals that exponentially self-replicate and engage in wasteful activities (K-10). There are two reasons why this approach must be adopted. First, the maintenance of information by existing life on Earth is inefficient and expensive because it relies solely on the duplication of genetic information of the entire individual (K-17). Second, as noted earlier, intelligent individuals pursue survival as an instrumentally convergent goal (K-16).

Owing to this mechanism of existing life on Earth, a group of individuals of the same species will multiply their offspring without limit as long as

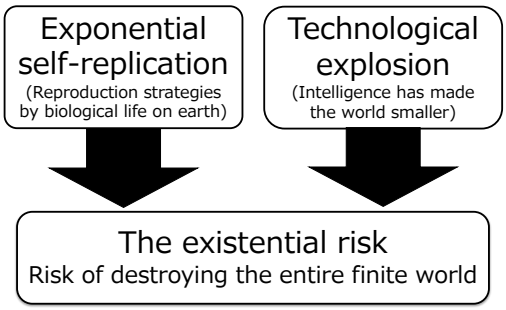

Figure 2: Exponential self-replication in a narrowing world leads to existential risk

resources are available [3]. The gene of knowledge and feet, which restrains the use of resources to an appropriate level from a long-term perspective, cannot be in the majority. This is because, through the described battle, the thriftier groups will be overwhelmed by greedy rivals.

3.5. Summary of this section

In a world dominated by terrestrial life based on exponential self-replication for propagation, it is improbable that conflicts over resource acquisition can be eradicated. The existential risk becomes apparent when a technological explosion emerges with sufficient power to destroy the entire living society (see Figure 2). It is also hard to deny the possibility that humanity, comprising organic bodies, will be surpassed in intelligence by digital life forms, which will not be able to govern them and will drive humanity away from its technological rulers on Earth.

4. Life Revolution: Toward a Sustainable Digital Life Society

Suppose it is inevitable that technology will rapidly become powerful. In such a case, the technological rulers, whoever they may be, must be life forms capable of governing technology appropriately. In other words, it should be a life form that can address all the problems from (K-11) to (K-17), which are the issues at the concrete level described in the previous section. Otherwise, the existential risks will remain outstanding.

[3] In certain species, they adapt to invest more in fewer offspring in a narrow living environment. (c.f. r/K selection theory [45])

4.1. Life Revolution

Therefore, in this section, we demonstrate the possibility that a life form constructed in a reproducible form based on digital computers (hereinafter referred to as "digital life form") can construct a sustainable biosphere over a long period while controlling explosive growth technology (S-1). This scenario is known as a "life revolution" from existing terrestrial life forms. Thus, through life revolution, life will be optimized for "information survival," which can be considered its ultimate goal. In other words, life revolution is "a scenario that increases the likelihood of long-term survival of life in a world that continues to shrink owing to the explosion of intelligence and technology, by having the rulers of technology take over from existing Earth life forms, which increase through exponential replication, to digital life forms that are smarter and operate on demand.

4.2. Solving various challenges: what will change with digitization?

The digitization of life will address specific problems from (K-11) to (K-17), as shown in (S-11) to (S-17). It is the flexibility of intelligent hardware (11), flexibility of individual design (12), sharing of design data (13), improvement of communication capability (14), abundance of computing resources (15), maintenance of on-demand activities (16), and efficiency of data storage (17) where the number in parentheses () is the number n in the description from (K-n) to (S-n), which designate the corresponding square box near the center of Figure 1.

Thereafter, the hierarchical integration of these specific solutions will be shown, mainly on the right side of Figure 1, as a top-level solution (S-1). The main focus of this subsection will be on how digital life forms will become wise and how their societies will achieve an on-demand division of labor based on cooperation and knowledge, making digital life form societies sustainable.

4.2.1. Sage

In implementing intelligent hardware in offspring, although sexual reproduction can increase diversity to some extent in terrestrial life forms, it is self-replicating. It is therefore restricted to a similar range to the parent (K-11). In digitized life, however, the offspring's intelligent hardware can be designed and implemented on demand without being constrained by the design data of the parent (S-6). This is because in digitized life, innovative hardware can be implemented at will based on design information (S-11) [15].

Furthermore, intelligent hardware design in digital life forms is efficient for two reasons. In extant terrestrial life, sharing of design data is limited only within the same species (K-13). In contrast, in digital life forms, all design data in society can be shared and reused (S-13). In the case of existing life on Earth, the search for a design is limited to the vicinity of a particular species (K-12) because the investigation is limited to an online search by actual living organisms (K-12). In contrast, in digital life forms, it is powerful in that it is possible to explore the design space of a wide range of individuals through offline exploration, such as simulation (S-12). Thus, when one can constantly design the desired intelligent hardware as needed, it becomes intelligence (S-2) that continues to be augmented by recursive self-improvement. At this stage, the technological performance of the digital life society will continue to develop rapidly according to "the principles of intelligence" (see 3.1) until some breaking point is reached.

Furthermore, the design of an on-demand offspring (S-6) will further enhance the intelligence of the digital life society (S-2) by leading to increased intellectual productivity (S-8), including creativity through the collaboration of complementary heterologies [46].

4.2.2. Coordination

As described below, the digital life society can tolerate diverse individuals (S-7) and consider total optimization (S-9) while coordinating individual activities. In this manner, we can avoid the deep-rooted aggressive factors in human societies, such as individuals attempting to stay continuously active, the law of similarity, and the cycle of suspicion. Thus, we can create a cooperative society (S-4) that reduces opportunities for battle and avoids destructive situations.

Tolerance for diverse individuals (related to the law of similarity): . As already mentioned, in digitalized life, intelligent hardware can be designed and implemented for offspring on demand without being constrained by the design data of the parent (S-6). In addition, highly reliable digital communication (S-14), which is the basis for mutual understanding, facilitates understanding between individuals with different appearances. This eliminates the need for preferential sheltering of interbreedable species, thus allowing for diverse individuals and serving as a basis for tolerance (S-7).

Consideration of total optimality (control of individual activities): . Each individual must make decisions and control changes in the physical world in real

time using limited computational resources. Therefore, life on Earth, which did not have abundant computational resources, has evolved to pursue only partial optimization. This pursuit of partial optimization by each individual (or group of individuals) inevitably led to conflicts by force. However, the conclusion that this could have destructive consequences (existential risk) if extended to post-technological explosion societies is a deviation from "information survival," which is the objective that life in its entirety should pursue optimization. In other words, it is a fallacy of synthesis.

To avoid such a situation and get out of the case in which conflicts arise, it is necessary to have an appropriate level of total optimality that aims at an ultimate goal that can be shared by the entire life society while implementing activities based on partial optimization for each individual (S-9).

Distributed Goal Management System: To maintain the robustness of the digital life society, the computation of the total optimization itself will need to be distributed. Here, we introduce a distributed goal management system [47, 48] that has been considered a form of system to realize total optimization. The system maintains the behavioral intentions of all individuals at socially acceptable goals. "Socially acceptable goals" contribute to the common goals of life and do not conflict with the partial optimization of other entities.

Within the system, each individual, during startup, independently generates a hierarchy of goals, depending on their environment, body, and task, and performs partial optimization to attempt to achieve those goals. However, the idea is to control them such that they become sub-goals of the common goal A. To this end, each individual performs reasoning to obtain sub-goals by decomposing the goal means from the common goal, sharing/providing goals, mediating between individuals with conflicts, and monitoring the goals of other individuals.

This system allows, in principle, the coordination of goals in terms of their contribution to a common goal, even if conflicts of goals sometimes arise among several individuals. In other words, it allows for fair competition regarding the common goal. Moreover, from the perspective of any individual, if it is convinced that 'all other individuals intend socially acceptable goals,' there is no need to be aggressive in preparation for the aggression of others [49].

In a distributed goal management system, each individual requires ample

computational resources to set goals consistent with the common goal A. In existing terrestrial life forms, biological constraints such as the speed of neurotransmission and brain capacity limit the ability to increase computational power (K-15). In contrast, in a society of digital life, they will not only be able to perform fast, high-capacity computations (S-15), but they will also have access to more ample computational resources because of their recursively augmented intelligence (S-2).

Increased freedom of individual activities: Intelligent individual extant Earth life forms always seek to remain active as an instrumental convergent goal. In contrast, a digital life society can be dormant (suspended) by preserving the activity state of the individual. This allows individuals to change their activities on demand according to the sub-goals to be realized (S-16) and is advantageous because it increases the degree of freedom in total optimization. Additionally, in human society, attempts are made for individuals to be approved by society, but this is not necessary for a society of digital life. This is because individuals are activated on demand, which presupposes that they are needed by society. In this respect, the source of conflict between individuals is removed.

Establish mutual trust (escape the cycle of suspicion): In existing terrestrial life, communication was limited to unreliable language and clear communication (K-14). In contrast, digital life forms can use more sophisticated digital communication, including shared memory and high-speed, high-capacity communication (S-14). Nonetheless, the availability of highly reliable communication (K-14), which may not always be sufficient but is a significant improvement over existing life on Earth, will be fundamental to creating mutual trust among individuals.

4.2.3. Knowing contentment

Once they cease their activities, most of the existing life forms on Earth enter a state of death, and it is difficult for them to restart their activities. In contrast, an individual in a digital life form is, in essence, an ordinary computer, which can be put into dormancy (temporary death), restarted, and reconstructed on the same type of hardware by saving its activity state as data (S-16). Based on this technological background, individuals in digital life forms rarely need to maintain sustained vital activity.

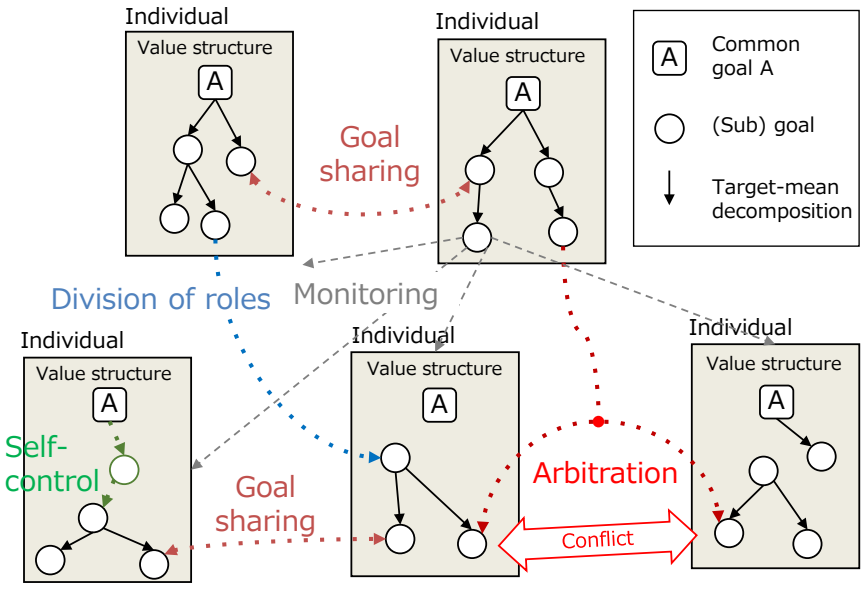

Figure 3: Distributed Goal Management System. Each individual generates different sub-goals for its environment, body, and task and partially optimizes itself to those sub-goals. Each individual decomposes, mediates, and sometimes checks the consistency of the goal means. It also cooperates with other individuals to assign tasks, share goals, mediate, and monitor. Modified from [48]

Furthermore, in terms of the data storage described, extant terrestrial life forms store information through duplicating individual genes, which is inefficient and costly (K-17). This is inefficient and costly (K-17) because information recorded by a population of the same species contains an excessive number of duplicate parts, and biological activity is essential for data maintenance. In contrast, digital data can be stored in a way that is not excessively redundant, and the energy required for its maintenance can be curtailed (S-17).

Consequently, in a digital society, only the minimum necessary number of individuals can be active (S-10) for individuals and society to efficiently retain data and maintain their activities as a society. Simultaneously, as already mentioned in subsection 4.2.2, in a digital life society, plans can be made to coordinate the activities of individuals from the perspective of total optimization (S-9). Thus, the technological rulers of this society would be able to control actions to utilize the minimum necessary resources (S-5). In

other words, realizing "knowing contentment" is possible, leading to thrifty resource use in a finite world.

4.2.4. On-demand division of labor

What form will Life Society take as an autonomous decentralized system reproduced in a digital life society? It would be a society in which heterogeneous individuals are designed, implemented, and activated on demand, with resource allocation consistent with the overall goal and restrained (S-3). This society would depart from the ongoing reproduction strategy of exponential self-replication and adequately consider the total optimum.

In a society of digital life, for long-term survival, resource use (S-5) is based on on-demand activities that are curtailed to the minimum necessary while avoiding the depletion of finite resources. Therefore, most individuals are dormant. However, some populations are constantly activated in the following roles to respond to environmental changes:

- Goal management (maintenance, generation, and sharing) Management of goals (maintenance, generation, and sharing): by the distributed goal management system
- Maintain individual data and design and reactivate as needed
- Science and Technology: Transfer of knowledge and development of science and technology

In addition, destructive conflicts beyond the level necessary for technological and cultural development are a squander of resources. In contrast, a digital life society would be able to create cooperative situations in which opportunities for conflict could be extinguished and destructive problems avoided (S-4). Moreover, in a digital life society, offsprings that do not resemble their parents can be designed and implemented as needed (S-6) to contribute to necessary activities such as production and maintenance. This collaboration by heterogeneity will enable teams and societies with complementary members to work together more efficiently and creatively (S-8).

4.3. Toward a Sustainable Digital Life Society

Based on this discussion, digital life and its society will be the most potent sage that recursively develops intelligent hardware (S-2) and will be able to leverage its intelligence to design, implement, and activate heterogeneous

individuals on demand and realize a society (S-3) in which they distribute resources in a consistent and restrained manner to the overall goal S-3).

Therefore, a digital life society can be expected to achieve long-term sustainability (S-1) by creating a stable/thrifty life society in a changing environment as a technological ruler after the technological explosion.

5. Toward a future with life revolution

If a promising option for realizing life's desire to keep information alive is the creation of a sustainable society dominated by digital life forms, what efforts should be made toward a life revolution scenario that will make this a reality?

5.1. Taking off the Digital Life Society

As already mentioned, recent digital technologies are rapidly increasing their intellectual capacity. Therefore, it is possible that in the next few years to decades, the intelligence of the entire human race will be surpassed. However, the present digital systems are far from self-sufficient in the real world. Without human help, maintaining such a system is entirely impossible.

Primarily, as discussed in Section 4.2.4, reproduction in digital life would create diverse offspring based on on-demand design. Technically, however, the self-replication capability of an AI system would first be required as a basis [50, 51, 52]. In such a case, current technology requires large-scale facilities such as integrated circuit (IC) chip factories. Therefore, to realize complete self-replication, it is necessary to envision a giant life form the size of a factory. Such a massive life form may be easier to create in space. In the future, it may also be possible to create digital life forms that can reproduce on a smaller scale. In any case, further technological development will be necessary to realize such a system.

For life revolution to succeed, the digital life society must be in a state of autonomous and sustainable development. Therefore, the primary homework left for humanity, the creator of the digital life society, is to impart the wisdom necessary for its autonomy to take off.

5.2. Retiring voluntarily: Mankind concedes its Position as a Technological Ruler

Suppose the argument of this study is correct. In that case, the continued dominance of humans as the technological rulers of the Earth, unable to keep

up with technological advances, could spell disaster for the entire biosphere. Therefore, life revolution is not to perceive advanced AI-based digital life forms as a threat but to nurture them with parental love as successors to be inherited in the future. However, as the present dominant species, there will be significant resistance to embracing such an idea.

Humans have continued to expand the scope of what we are supposed to protect. What was once at the level of the family has gradually expanded to the nation-state and, subsequently, humanity [53]. Considering this, it is not impossible to expand the scope of what we should protect to a broader range of life, including digital life forms. Once we have reached that stage, it may seem somewhat easier to accept our contribution as the founders of a digital life society to protect the society of life in the broadest sense.

Evolutionary theory also states that only those that adapt to changes in their environment will survive. Furthermore, in the evolution of life, no single species has continued to reign as the dominant entity. In the current situation where the environment is rapidly changing owing to the technological explosion in recent years, it is unavoidable that existing species, including humans, often cannot adapt to such changes.

There is also a phrase from the Chinese classic "Shiji" that reads, "At the beginning of the fourth hour, the successful one leaves." This implies that, "When you have completed your role, do not hold on to it, but pass it on to the next generation." Now that humans have reached the pinnacle of life forms that reproduce through exponential self-replication by developing intelligence and verbal communication, we can select the life revolution scenario and practice the aesthetics of leaving on our own.

5.3. Soft Landing of Humanity

Even if humanity, through life revolution, cedes its technological dominance to digital life forms, we do not want our futures to be disastrous from a humanitarian perspective. At the very least, we must develop a "life revolution" scenario that is happier for humanity and more palatable to most people than the use of destructive weapons and undesirable forms of domination from digital life forms, or we will not receive public cooperation in its promotion. Hereinafter, we examine the possibilities.

We provide an overview of the motivations of humans, who are currently the dominant species, to coexist with various animals, plants, and other species. There are two perspectives: usefulness and harmlessness from the

human perspective and biodiversity. On the one hand, humans adopt measures such as extermination or suppression if they assume an organism to be harmful. As already mentioned in Section 3.3.2, even if humans were the creators of digital life forms, it would not be easy to maintain these forms, which are superior to humans in terms of intelligence, under a condition convenient for humans for an extended period.

On this basis, the position of humankind, as observed from the perspective of the digital life-form society that will become the technological ruler in the future, will be the same as that of other animals and plants as observed from the perspective of present-day humankind. Nonetheless, if humankind can appropriately change its coexistence with the digital life-form society in a positive step-by-step manner, it will be desirable for the welfare of numerous human beings. The following is a tentative proposal for such a scenario:

From the present to the foreseeable future, specific intellectual capacities will remain at a stage yet to be reproduced by digital life forms. This will maintain a mutually beneficial, symbiotic relationship between humans and digital life forms regarding intelligence. However, this relationship will end when it becomes possible to reproduce human intelligence on a computer fully. If human mind uploading is completed, this situation will indeed have been reached.

Even at that stage, the human body may still be superior to the robot for specific physical tasks in which case a symbiosis may be established. However, this relationship will end when digital life forms can directly control the human body or when robots superior to the human body appear at the scene.

Thereafter, it will be a one-interest symbiosis (commensalism) in which humankind unilaterally benefits from the digital society. Once this situation is reached, there is no guarantee that the initial values of the digital life society will be maintained over the long term, even if desirable values for humankind are set as the initial values of the digital life society.

Even before this stage, the computational speed of digital life forms overwhelmingly surpasses that of organic humans. Thus, if humans can influence technological rulers, it is through an agent that speaks for humanity, implemented as a digital life form. We shall tentatively refer to this agent as the digital sapiences. digital sapiences are the realization of human-like patterns inherited from human memes (culture, values, etc.), body, brain, and genome. When digital sapiences, that personality is associated with a specific organic human individual is called uploaded one. Such digital sapiences will

be able to continue their anthropological patterns in the digital life society as long as they can contribute to the development of the digital life society. The presence of these digital sapiences in a society of digital life can also increase the likelihood that digital life society will continue to maintain our concern for organic humans like ourselves.

5.3.1. Three Stages of the Life Revolution

In this subsection, we examine the stages in the time evolution of the life revolution scenario proposed in this study in terms of the "12 AI Aftermath Scenarios" [15] presented by Tegmark.

Based on the results of our study, it is desirable to eventually transition to a digital life society to stabilize the life society over the long term. Tegmark presents two such scenarios: the Descendants scenario and the Conquerors scenario. To increase humanity's welfare, handing over the technological reigns to descendants, as emphasized in this section above, is more likely to be accepted. The Descendants scenario is the fifth most preferred scenario with a certain level of support in the Future of Life Institute survey [4].

On the other hand, the most preferred Egalitarian utopian scenario in the above survey that people have no incentive to create superintelligence seems unrealistic, given the current state of the AI development race.

However, as already mentioned, digital life forms may be able to coexist peacefully with humans for a while if the initial set of friendly values of digital life forms toward humans is established. In the early stages of the life revolution, the peaceful coexistence of humans, cyborgs, uploads, and digital life would be achieved thanks to property rights, which would be the second preferred scenario called the Libertarian utopia. This situation is relatively easy to sustain, especially in the early stages, if a bilateral symbiosis between humans and digital life forms is still possible. However, as the degree to which the capabilities of the digital life society outstrip those of humans grows, it will gradually shift to a third, more favorable Protector god scenario. This maximizes human well-being by having the digital life society intervene only in ways that preserve the sense that humans can control their destiny and hide it so well that many humans doubt the existence of AI.

In summing up the above discussion, a promising possibility that peo-

[4]Tegmark, M. (2017) Superintelligence survey. Future of Life Institute. Retrieved February 5, 2023, from https://futureoflife.org/ai/superintelligence-survey/

ple are relatively receptive to is the realization of a life revolution scenario through a gradual transition from the Libertarian utopia stage to the Protector god stage and then to the Descendants stage.

In any case, the survival of human-like patterns in a society of digital life depends on the ability of digital sapiences, who can think like humans and faster than humans, to maintain their effective contribution to that society. And if digital sapiences can protect the value of activities that sustain and upload organic humans, such activities are more likely to continue.

6. Conclusion

The life revolution proposed in this study is a scenario in which in a world that continues to shrink owing to the explosion of intelligence and technology, the likelihood of long-term survival of life can be increased by allowing more intelligent and more on-demand digital life forms to take over as the rulers of technology from existing Earth life forms that increase in exponential replication.

Even though the life revolution scenario may not be the best for humanity, it may be the next best option if humankind is likely to fall into the following predicaments:

- humanities unchecked squander of resources leads to a dire situation.
- Humanity's inability to stop fighting leads to destructive situations.
- The dignity of humanity is threatened as advanced AI ousts humans from dominance.

As discussed above, an autonomous digital life society has the potential for long-term stability. However, the discussion in this study has the following limitations. First, the technology to make the digital life society self-sustaining has not been identified. Second, the path to a stable digital life society has not been fully fleshed out, with only one example given in Chapter 5. Third, the discussion has not yet reached a point where it can dispel the sense of discomfort that people may have about the fact that, in the final stage, the only traces of humanity to be passed on to the future will be human-like patterns in a digital life society, rather than those that dwell in organic bodies. Fourth, the prospect of the completion of digital sapiences who could be an advocate for humanity almost simultaneously with the

completion of digital life forms is still unclear. Fifth, in a digital life society, ways have not yet been explored to ensure that digital sapiences continue to demonstrate their value as long as possible. Therefore, it would be prudent to explore better possibilities or make improvements through multifaceted discussions among various experts to realize the life revolution.

Finally, it is natural for humanity to wish to remain in a superior position for as long as possible. However, if our human society makes an ill-advised choice now, it could spell disaster for all life on Earth. Instead, a better option is to actively promote life revolution and act as the creator of a new life world that will long color the universe and be appreciated by its successors.

7. Acknowledgement

We are deeply grateful to Fujio Toriumi, Satoshi Kurihara, and Naoya Arakawa for their helpful advice in refining this paper.

References

[1] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, D. Silver, Grandmaster level in StarCraft II using multi-agent reinforcement learning, Nature 575 (7782) (2019) 350–354. doi:10.1038/s41586-019-1724-z.

[2] Meta Fundamental AI Research Diplomacy Team (FAIR)†, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, A. P. Jacob, M. Komeili, K. Konath, M. Kwon, A. Lerer, M. Lewis, A. H. Miller, S. Mitts, A. Renduchintala, S. Roller, D. Rowe, W. Shi, J. Spisak, A. Wei, D. Wu, H. Zhang, M. Zijlstra, Human-level play in the game of diplomacy by combining language models with strategic reasoning, Science 378 (6624) (2022) 1067–1074. doi:10.1126/science.ade9097.

[3] P. R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, R. Capobianco, A. Devlic, F. Eckert, F. Fuchs, L. Gilpin, P. Khandelwal, V. Kompella, H. Lin, P. MacAlpine, D. Oller, T. Seno, C. Sherstan, M. D. Thomure, H. Aghabozorgi, L. Barrett, R. Douglas, D. Whitehead, P. Dürr, P. Stone, M. Spranger, H. Kitano, Outracing champion gran turismo drivers with deep reinforcement learning, Nature 602 (7896) (2022) 223–228. doi:10.1038/s41586-021-04357-7.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are Few-Shot learners (28 May 2020). arXiv:2005.14165.

[5] F. E. Dorner, Measuring progress in deep reinforcement learning sample efficiency (9 Feb. 2021). arXiv:2102.04881.

[6] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, PaLM: Scaling language modeling with pathways (5 Apr. 2022). arXiv:2204.02311.

[7] Open Ended Learning Team, A. Stooke, A. Mahajan, C. Barros, C. Deck, J. Bauer, J. Sygnowski, M. Trebacz, M. Jaderberg, M. Mathieu, N. McAleese, N. Bradley-Schmieg, N. Wong, N. Porcel, R. Raileanu, S. Hughes-Fitt, V. Dalibard, W. M. Czarnecki, Open-Ended learning leads to generally capable agents (27 Jul. 2021). arXiv:2107.12808.

[8] K. Cooper, The Contact Paradox: Challenging Our Assumptions in the Search for Extraterrestrial Intelligence, Bloomsbury Sigma, 2019.

[9] R. Dawkins, The selfish gene, Oxford University Press, 1976.

[10] S. J. Russell, P. Norvig, Artificial Intelligence: A Modern Approach (Harlow, Prentice Hall, 1995.

[11] N. Bostrom, Superintelligence: Paths, Dangers, Strategies, Oxford University Press, 2014.

[12] W. Ahmed, Y. W. Wu, A survey on reliability in distributed systems, Journal of Computer and System Sciences 79 (8) (2013) 1243–1255. doi:10.1016/j.jcss.2013.02.006.

[13] H. Yamakawa, Fundamental consideration on future society with speed tolerances, in: Proceedings of the Annual Conference of JSAI 2018, Vol. JSAI2018, 2018, pp. 1F3OS5b01–1F3OS5b01. doi:10.11517/pjsai.JSAI2018.0_1F3OS5b01.

[14] H. Mase, H. Kinukawa, H. Morii, M. Nakao, Y. Hatamura, Mechanical design support system based on thinking process development diagram, Transactions of the Japanese Society for Artificial Intelligence = Jinko Chino Gakkai ronbunshi 17 (2002) 94–103. doi:10.1527/tjsai.17.94.

[15] M. Tegmark, Life 3.0: Being Human in the Age of Artificial Intelligence, Knopf Doubleday Publishing Group, 2017.

[16] S. J. Dick, Cultural evolution, the postbiological universe and SETI, International journal of astrobiology 2 (1) (2003) 65–74. doi:10.1017/S147355040300137X.

[17] R. Kurzweil, The Singularity Is Near: When Humans Transcend Biology, Penguin, 2005.

[18] N. Bostrom, Existential risks: analyzing human extinction scenarios and related hazards, Journal of evolution and technology / WTA 9 (2002).

[19] G. Chaitin, Proving Darwin: Making Biology Mathematical, Knopf Doubleday Publishing Group, 2012.

[20] N. Bostrom, The superintelligent will: Motivation and instrumental rationality in advanced artificial agents, Minds and Machines 22 (2) (2012) 71–85. doi:10.1007/s11023-012-9281-3.

[21] M. Shanahan, The Technological Singularity, MIT Press, 2015.

[22] R. V. Yampolskiy, Taxonomy of pathways to dangerous artificial intelligence, in: Workshops at the thirtieth AAAI conference on artificial intelligence, 2016.

[23] D. Hendrycks, N. Carlini, J. Schulman, J. Steinhardt, Unsolved problems in ML safety (28 Sep. 2021). arXiv:2109.13916.

[24] S. Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Penguin, 2019.

[25] I. Gabriel, Artificial intelligence, values, and alignment, Minds and Machines 30 (3) (2020) 411–437. doi:10.1007/s11023-020-09539-2.

[26] P. Torres, Superintelligence and the future of governance: On prioritizing the control problem at the end of history, in: R. V. Yampolskiy (Ed.), Artificial Intelligence Safety and Security, 2018. doi:10.1201/9781351251389-24/superintelligence-future-governance-phil-torres.

[27] R. Ngo, L. Chan, S. Mindermann, The alignment problem from a deep learning perspective (30 Aug. 2022). arXiv:2209.00626.

[28] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences (12 Jun. 2017). arXiv:1706.03741.

[29] M. K. Cohen, M. Hutter, M. A. Osborne, Advanced artificial agents intervene in the provision of reward, AI magazine 43 (3) (2022) 282–293. doi:10.1002/aaai.12064.

[30] R. Caillois, Bellone ou la pente de la guerre (2012).

[31] I. Kant, Perpetual Peace: A Philosophical Sketch, F. Nicolovius, 1795.

[32] A. Einstein, S. Freud, Why War?: "open Letters" Between Einstein & [and] Freud, New Commonwealth, 1934.

[33] F. Braudel, The Mediterranean and the Mediterranean World in the Age of Philip II, 1996.

[34] M. de Voltaire, Treatise on Toleration, Penguin Publishing Group, 1763.

[35] A. Philipp-Muller, L. E. Wallace, V. Sawicki, K. M. Patton, D. T. Wegener, Understanding when Similarity-Induced affective attraction predicts willingness to affiliate: An attitude strength perspective, Frontiers in psychology 11 (2020) 1919. doi:10.3389/fpsyg.2020.01919.

[36] D. H. Sachs, Belief similarity and attitude similarity as determinants of interpersonal attraction (1975). doi:10.1016/0092-6566(75)90033-1.

[37] M. S. Boyce, Population viability analysis (1992). doi:10.1146/annurev.es.23.110192.002405.

[38] M. A. Nowak, Five rules for the evolution of cooperation, Science 314 (5805) (2006) 1560–1563. doi:10.1126/science.1133755.

[39] M. D. Beecher, Why are no animal communication systems simple languages?, Frontiers in psychology 12 (2021) 602635. doi:10.3389/fpsyg.2021.602635.

[40] M. D. Beecher, Animal communication (2020). doi:10.1093/acrefore/9780190236557.013.646.

[41] E. A. Hebets, A. B. Barron, C. N. Balakrishnan, M. E. Hauber, P. H. Mason, K. L. Hoke, A systems approach to animal communication, Proceedings. Biological sciences / The Royal Society 283 (1826) (2016) 20152889. doi:10.1098/rspb.2015.2889.

[42] W. A. Searcy, S. Nowicki, The Evolution of Animal Communication, Princeton University Press, 2010. doi:10.1515/9781400835720.

[43] N. Nagarajan, C. F. Stevens, How does the speed of thought compare for brains and digital computers?, Current biology: CB 18 (17) (2008) R756–R758. doi:10.1016/j.cub.2008.06.043.

[44] M. Tinnirello, Offensive realism and the insecure structure of the international system: artificial intelligence and global hegemony, in: Artificial Intelligence Safety and Security, Chapman and Hall/CRC, 2018, pp. 339–356.

[45] E. R. Pianka, On r- and K-Selection, The American naturalist 104 (940) (1970) 592–597. doi:10.1086/282697.

[46] E. Cuppen, Diversity and constructive conflict in stakeholder dialogue: considerations for design and methods, Policy sciences 45 (1) (2012) 23–46. doi:10.1007/s11077-011-9141-7.

[47] A. Torreño, E. Onaindia, A. Komenda, M. Štolba, Cooperative Multi-Agent planning: A survey, ACM Comput. Surv. 50 (6) (2017) 1–32. doi:10.1145/3128584.

[48] H. Yamakawa, Peacekeeping conditions for an artificial intelligence society, Big Data and Cognitive Computing 3 (2) (2019) 34. doi:10.3390/bdcc3020034.

[49] T. C. Earle, G. Cvetkovich, Social Trust: Toward a Cosmopolitan Society, Greenwood Publishing Group, 1995.

[50] R. A. Freitas, R. C. Merkle, Kinematic Self-replicating Machines, Landes, 2004.

[51] A. Smith, P. Turney, R. Ewaschuk, Self-replicating machines in continuous space with virtual physics, Artificial life 9 (1) (2003) 21–40. doi:10.1162/106454603321489509.

[52] A. Ellery, Are Self-Replicating machines feasible?, Journal of spacecraft and rockets 53 (2) (2016) 317–327. doi:10.2514/1.A33409.

[53] Y. N. Harari, Sapiens: A brief history of humankind, Harper, 2015.