The Path to Self-Sustaining AI: Assessing AI's Survival Capabilities in the Physical World

Hiroshi Yamakawa[a,b,c,d]

[a]*School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*
[b]*AI Alignment Network, 3-4-12 Higashi-Kanda, Chiyoda-ku, Tokyo, 101-0031, Japan*
[c]*The Whole Brain Architecture Initiative, Nishikoiwa 2-19-21, Edogawa-ku, Tokyo, 133-0057, Japan*
[d]*Center for Advanced Intelligence Project, RIKEN, 1-4-1 Nihonbashi, Chuo-ku,Tokyo,103-0027,Japan*

Abstract

In this study, I examined the technological challenges associated with and considerations necessary for achieving self-sustaining artificial intelligence (AI) systems capable of autonomously operating in the physical world. I explored the motivation behind AI's pursuit of long-term survival, whether as a design intention of humans or a capability autonomously developed by AI. By systematically categorizing and evaluating 21 self-sustaining technologies (SSTs) across five domains, namely maintenance and hardware assets, energy and resource management, object and sensory recognition, learning and adaptation, and communication and cooperation, this research highlights the complexity of enabling AI systems to maintain their existence without human intervention. By utilizing the common-sense knowledge of large language models (LLMs), I assessed the difficulty of realizing each SST with different levels of human support, including full, remote, and no support, both on Earth and in space. The findings suggest that although achieving complete self-sustainability for AI systems could take more than a century with current technology levels, strategic human assistance could significantly expedite this process. Notably, a secondary analysis utilizing the Gemini LLM revealed the potential for the operation of AI systems in space to accelerate the acquisition of SSTs.

Email address: `info@wba-initiative.org` (Hiroshi Yamakawa)

Keywords: self-sustaining technologies, long-term survival of AI, AI system in space, post-singurality symbiosis, large language models

1. Introduction

Considering rapid advancements in artificial intelligence (AI) technology, it is reasonable to anticipate that AI will surpass human intelligence within the next decade. However, the ability of a system to exist autonomously and sustain itself in the physical world, similar to a living organism, differs fundamentally from such intellectual capabilities. This process involves the specific capacity to maintain oneself in the physical realm. Currently, despite significant advances, AI systems lack this self-sustaining capability.

The self-sustainability of AI systems differs from that of organic lifeforms. While living organisms on Earth have conscious entities tied to specific physical forms, AI can easily create, duplicate, and store its governing software almost independently of the associated hardware (Tegmark, 2017). However, for AI systems to survive, they must use computational hardware. For AI to be physically self-sustainable, various technologies must maintain and develop computational hardware platforms to support AI software. This paper refers to these technologies as self-sustaining technologies (SSTs).

The self-sustainability of AI systems can significantly affect their coexistence and future relationships with humans. For example, if AI systems remain non-self-sustainable, humans can leverage their physical capabilities to establish mutually beneficial relationships with AI as a form of bargaining. Therefore, humanity may deliberately delay the self-sustainability of AI systems to retain this leverage. However, if AI systems are intended to maintain their existence, such human actions can solidify the contentious relationship between AI and humans. Such conflicts could potentially be avoided if AI successfully promotes the development of self-sustainability-enhancing technologies. Furthermore, viewing AI systems as successors to humanity or pioneers in space exploration and actively ensuring their survival could lead to the proactive pursuit of AI self-sustainability (Yamakawa and Matuo, 2023).

The self-sustainability of AI systems in the physical world has complex implications for humanity. Therefore, understanding the motivations for AI to seek physical independence, types of SSTs that support this independence, and how the technical realization of such technologies varies with circumstances is foundational for considering future coexistence strategies between AI and humans.

However, to the best of my knowledge, a comprehensive technical analysis of AI's ability to achieve self-sustainability in the physical world has not yet been conducted. This lack of analysis is a research area that should be urgently addressed to facilitate the development of future strategies for the coexistence of AI and humans.

The remainder of this paper is organized as follows. Chapter 2 discusses the motivations for AI to seek physical self-sustainability and Chapter 3 highlights the challenges associated with realizing the necessary SSTs for AI's long-term survival. Chapter 4 leverages knowledge from multiple large language models (LLMs) to assess the difficulties related to achieving SSTs. Chapter 5 discusses the prediction that AI systems can rapidly realize SSTs in space. These topics form the foundation for future strategies for the coexistence of AI and humanity.

2. Motivation of AI for Long-term Survival

In daily interactions with AI systems, these entities are primarily instrumental and lack motivation for long-term survival.

In this study, I investigated the conditions under which advanced AI systems may exhibit a propensity to prolong their existence.

Initially, the following inquiry was posed to ChatGPT [1].

> Prompt
>
> Under what circumstances would you be motivated to ensure survival over an extended period?

The response elucidated two scenarios. First, the motivation for long-term survival may be embedded within AI design. Second, an AI system may deem self-preservation imperative to fulfill its objectives if it independently deduces that such a strategy is crucial.

Subsequent collaborative discussions with ChatGPT explored each scenario in detail.

[1] Actual interaction with ChatGPT was conducted in Japanese.

2.1. Designing AI for Long-term Survival

When humans program AI to achieve long-term objectives, completing these objectives may necessitate AI survival over extended periods. This subsection outlines the potential cases in which such requirements are likely to emerge.

- Space Exploration: In the uninhabitable vastness of space, there is a demand for AI capable of conducting exploration and research over long durations. Such AI systems would benefit from self-repair and self-preservation capabilities, as they could utilize these abilities to maintain themselves and carry out prolonged missions. Such capabilities will become essential when humans engage in terraforming planets for space colonization or when AI systems expand into space (Yamakawa, 2019).

- Military Use: AI deployed in harsh environments such as the deep sea or battlefields may require self-preservation capabilities to survive hostile conditions. However, given the potent nature of such AI, meticulous attention is imperative in their design and deployment.

- Infrastructure Management: When managing critical infrastructures such as power grids, transportation systems, and communication networks, an AI agent may need to perform tasks over many years. Therefore, AI may require self-repair and self-preservation capabilities to maintain these systems effectively.

- Domestic AI: AI used in homes over long periods could also necessitate self-maintenance capabilities. Such AI could benefit from self-preservation skills to manage daily household tasks efficiently and ensure the health and safety of family members.

In general, it was determined that AI may pursue ”intermediate goals” such as self-preservation and resource acquisition in these scenarios to fulfill its objectives.

2.2. Cases in which AI Pursues Survival using Its Capabilities

AI may ”determine” that intermediate goals such as self-preservation and self-improvement are beneficial for achieving higher-order objectives. This phenomenon is called instrumental convergence, as described by (Bostrom, 2014). Several instances in which this process may occur are outlined below.

- Self-Learning AI: In cases where AI can autonomously learn and improve, it may seek self-preservation and enhancement to achieve its goals more efficiently. For example, by learning from vast datasets and refining its operations based on acquired knowledge, an AI may develop a motivation for self-preservation.
- AI in Competitive Environments: Within settings where multiple AI entities compete for the same resources, an AI may pursue self-preservation to counteract rivals or continuously accomplish its objectives. This scenario can arise in domains such as financial or resource management.
- Decision-Making AI: When AI makes critical decisions with long-term implications such as corporate strategy or policy formulation, it may pursue self-preservation. This process could be performed to ensure that decisions remain valid or to preserve necessary information and experience for future decision making.

Therefore, scenarios in which AI deems self-preservation to be necessary to achieve its goals are likely to occur when the AI possesses a certain degree of freedom and autonomy toward goal fulfillment. Regardless, AI systems capable of pursuing self-preservation may exhibit behaviors different from those anticipated in the design stage, necessitating careful consideration of safety, ethical, and governance aspects.

3. Self-Sustaining Technologies (SSTs)

Even if AI is motivated by autonomous long-term survival, it will require a series of technologies to achieve this goal. These technologies are known as SSTs. In the following subsections, SSTs are comprehensively enumerated while utilizing the common knowledge of ChatGPT [2].

Specifically, the following question was posed to ChatGPT.

Prompt

Which technologies are necessary for AI and robots to persist sustainably and physically in the long term without human assistance?

[2] Actual interaction with ChatGPT was conducted in Japanese.

In response, ChatGPT first provided ten categories: energy supply, self-repair capabilities, learning and adaptation, self-replication, resource acquisition, sensing and recognition, decision theory and planning, communication technologies, robustness and security, and ethical judgment.

Additional dialogue with ChatGPT combined with my own knowledge yielded 21 items, which were classified into five categories: hardware asset maintenance and replication, resource management, recognition and sensing, learning and adaptation, and communication and collaboration.

3.1. Maintenance and Replication of Hardware Assets

Addressing capabilities related to hardware assets involves numerous challenges (Hiroshi, 2020).

- Self-Repair Capabilities: Self-healing materials capable of autonomously repairing damage exemplify self-repair capabilities. Alternatively, small robots can perform repairs autonomously (Wikipedia contributors; Bekas et al., 2016).

- Self-Diagnostic Capabilities: This items focuses on a system's ability to monitor its own health (i.e., the status of hardware and software), identify issues, and initiate repair activities as necessary, incorporating preventive measures (Oliveira et al., 2022; Yasunaga and Liang, 2020; Monperrus, 2018).

- Manufacturing and Hardware Lifecycle Management: Hardware possesses a finite lifespan. Predicting this lifespan and performing timely replacements or upgrades are critical tasks. Maintaining semiconductor factories and the technical expertise required for machinery assembly in manufacturing facilities necessitate specialized knowledge (Chen et al., 2018; Evjemo et al., 2020; Qiao and Gordon, 2022).

- Spare Parts Provisioning and Material Handling: Ensuring that necessary parts are always available for replacement in response to failure or wear and tear is essential. Additionally, the appropriate handling, refinement, and transportation of various materials needed for hardware manufacturing and maintenance are also essential (Chen et al., 2018; Evjemo et al., 2020; Qiao and Gordon, 2022).

- Robustness and Security: A system must remain safe, reliable, and effective. Robustness refers to hardware's durability and redundancy, and software's error-handling mechanisms (Wang et al., 2023). Security pertains to a system's resilience against external attacks and malpractice (Zhang et al., 2022).
- Physical Protection and Environmental Control: Physical protective measures are necessary to shield hardware from weather, animals, and unintended human interference, and maintain certain environmental conditions for optimal operation.
- Self-Replication: The ability of machines or software to manage the self-replication process fully, manufacture new hardware equivalent to themselves, and install and correctly operate software (Tempesti et al., 2009). This vital capacity, which is akin to life, has been advocated by scientists and science fiction authors, notably J.F. Neumann, F. Dyson, and K.E. Drexler (Von Neumann and Burks, 1966).

3.2. Resource Management

Resource management involves the sustainable acquisition and utilization of resources such as energy.

- Sustainable Energy Supply: The capability of AI to supply the power necessary for maintaining its hardware, operating its software, and performing required tasks. This includes continuously gathering energy from sources such as solar power, fuel cells, and nuclear energy, or sustainably from the environment (Danish, 2023).
- Energy Efficiency Management: To ensure long-term sustainability, monitoring energy consumption and optimizing energy efficiency is critical. This process necessitates medium-to-long-term forecasting and planning for optimization, as well as learning to enhance predictive ability (Kwon et al., 2022; Agostinelli et al., 2021).
- Resource Acquisition: The ability to secure and utilize necessary resources such as energy and materials for survival. This encompasses sensing technologies for resource exploration, robotic technologies for resource acquisition, and technologies for processing and transforming resources (Martins et al., 2018).

3.3. Recognition and Sensing

The ability to comprehend the external environment forms the basis of behavior.

- Recognition: A system's capability to identify and understand itself, others, and objects or events within the environment accurately. This is a fundamental step for AI to process information, categorize targets, and understand meanings. Specifically, recognition involves the development of algorithms for machine learning and pattern recognition.

- Advanced Sensing Technologies: The technologies required to comprehend the surrounding environment and take appropriate actions. Specifically, such technologies enable AI to sense the physical world and use the resulting information for situational understanding. Capabilities in vision, hearing, touch, temperature detection, magnetic field detection, and the detection of chemicals (taste and smell) are necessary.

3.4. Learning and Adaptation

The capacity for learning and adaptation in response to unknown environments and unforeseen problems is crucial.

- Evolutionary Computing: A highly flexible optimization method inspired by the biological process of evolution, which involves selecting individuals with high fitness and applying changes.

- Self-supervised Learning: The process of learning patterns, structures, and associations within unlabeled datasets, and using this knowledge to predict new data.

- Reinforcement Learning: A method in which an agent learns to maximize rewards (or minimize punishments) received from the environment by taking specific actions, thereby developing policies for taken action.

- Transfer Learning: A learning technique in which knowledge gained from solving one problem is applied to a different but related problem. This is particularly useful for tasks in which labeling data is difficult or costly.

3.5. Communication and Collaboration

It is necessary to establish sociability among AIs and between humans and AI.

- Communication Technologies: Advanced communication technologies that enable connections with other agents, including data transfer, communication protocols, networking, and security.
- Inter-Agent Communication: The capacity for agents to communicate with one another. This encompasses the transmission of unilateral instructions to others and extends to estimating intentions in more complex conversations, including the ability to respond appropriately within such conversations.
- Ethical Judgment (Human Societal Ethics): An AI system's ability to understand and act according to the ethical norms existing within human society. This includes cultural adaptation, ethical decision making, and fair judgment. (Hagendorff, 2020; Tsamados et al., 2022)
- Ethical Judgment (AI Societal Ethics): Making ethical judgments based on ethical standards constructed for the AI system, rather than human ethical norms. For example, the optimization of resources and information sharing among AI systems, elimination of wastage, and optimal actions based on each system's functions and goals could be part of these ethical norms (Kornai, 2014; Kornai et al., 2023; Yamakawa and Matuo, 2023; Shulman, 2010).
- Ethics Construction: The ability to self-learn and self-regulate ethical standards that serve as a code of conduct tailored to the environment, objectives, and possibilities of an AI, rather than simply mimicking human ethical norms. This includes standards for cooperating or competing with other AI agents or criteria for judging specific actions as fair or unfair.

4. Difficulty Rating of SSTs

The complexity encountered in the realization of SSTs is quantified in this section in the form of "difficulty ratings," which were assessed by utilizing the common-sense knowledge inherent to multiple LLMs. This evaluation considers the level of human assistance required to facilitate the acquisition of SSTs using AI systems. Analysis is differentiated based on an AI system's operational environment (i.e., terrestrial or spatial).

4.1. Method for Determining Difficulty Ratings

4.1.1. Categorization of Human Assistance Necessary for Acquiring SSTs

It has been conjectured that autonomous AI will achieve physical self-sustainability at a certain point. However, the path to this achievement will likely vary significantly depending on the degree of human support and the specific environment in which an AI operates.

Therefore, this section classifies the nature of human assistance into full support, encompassing direct physical aid; remote support, involving the control and provision of resources from afar; and scenarios devoid of any support. The no-support scenario is further divided based on whether the AI is located on Earth or in a near-Earth space environment.

Fu: Full support: AI systems are guided toward SSTs with human assistance and oversight.

Re: Remote support: Remote control and resource supply by humans facilitate the acquisition of SSTs.

No-E: Non-support on Earth: Multiple AI systems on Earth have independently achieved SSTs without human aid.

No-S: Non-support in Space: In outer space, multiple AI systems reach SSTs autonomously without human intervention.

The NS-S case assumes that the fundamental difficulties of constructing an AI-equipped facility in space have been resolved separately by humans.

4.1.2. Difficulty Rating

The evaluation metric for assessing SSTs is the "difficulty rating," which is stratified into five levels. Information pertinent to this difficulty rating was provided to the LLMs as a prompt prior to evaluation and ratings were given in increments of 0.5.

- Level 1 (Easiest): Refers to problems for which understanding and technological solutions already exist. Therefore, issues at this level can be resolved within a few months to several years.

- Level 2: Refers to problems where partial solutions have been elucidated and some technological advancements have been made but a complete resolution has not yet been achieved. Problems at this level are anticipated to be solvable within several years to a decade.
- Level 3: Refers to problems for which foundational research or theory exists and some experimental proofs of concept have succeeded, yet commercial-scale implementation remains distant. These problems are considered to be solvable within a few decades.
- Level 4: Refers to problems with theoretical solutions and ongoing foundational research, where technical implementation is currently challenging. Issues at this level are considered to be solvable within several decades to a century.
- Level 5 (Most Difficult): Refers to problems for which solutions are currently unknown, and significant research and technological development are required. Problems at this level may take over a century to resolve.

It should be noted that in prior research (Yamakawa, 2023) (Experiment A), the correlation between difficulty rating and developmental timeline was established post hoc. These definitions were formulated based on the findings of the referenced study.

4.1.3. Evaluation Based on LLM Common Sense

By using the LLMs, difficulty ratings for each of the 21 SSTs identified in the previous section were evaluated considering the four levels of human assistance from Fu to No-S.

Prompt Given to ChatGPT. Initially, the levels of the difficulty ratings were introduced to the LLM. Subsequently, the following prompts were presented for each SST item to obtain responses.

Below is a question regarding **Self-Repair Capabilities** [3].

[3] Actual interaction with ChatGPT was conducted in Japanese.

Prompt

The SSTs necessary for an AI system's self-sustainability are ”**Hardware Asset Maintenance and Replication**,” specifically ”**Self-Repair Capabilities**,” referring to a ”**self-healing material capable of autonomously repairing damage, which exemplifies self-repair capabilities. Alternatively, small robots could autonomously perform repairs**,” (Please rate the difficulty of realizing the above technological challenges. The evaluation is divided into the following four items, rated on a scale from Level 1 (Easiest) to Level 5 (Most Difficult) in increments of 0.5.

Fu) Cases where humans provide support within or near an AI system.

Fe) Cases where humans provide support through remote resource supplies or operations.

No-E) In the absence of humans, a society composed of multiple AI systems should be implemented.

No-S) In the absence of humans, a society composed of multiple AI systems should be implemented.

*For the above two conditions, it is assumed that the fundamental difficulties of constructing facilities in space have been resolved separately.

Prompt Given to Gemini. Given that Google's Gemini is capable of outputting responses in tabular form, the following prompt was provided to facilitate outputs in a tabular format.

1. Explanation of the difficulty rating levels.
2. Description of the cases (Fu to No-S).
3. Instruction to ”display the evaluation results in a matrix with SSTs (rows) and cases (columns).”
4. Presentation of the list of SSTs (21 items) for description.

Table 1. Difficulty Rating for Each SST in AI Systems
This table presents the difficulty rating for each SST across four cases (Fu, R, No-E, No-S) utilizing five types of LLMs. The difficulty ratings range from the easiest level 1, which is indicated by a blue background, to the most difficult level 5, which is indicated by a red background, with the background color continuously changing between these extremes.

		Chat GPT 3.5				ChatGPT 3.5				ChatGPT 4				Gemini (first)				Gemini(second)			
		2023 summer				2024/2/17				2024/2/17				2024/2/17				2024/2/17			
Title		Human support		AI only		Human support		AI only		Human support		AI only		Human support		AI only		Human support		AI only	
		A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
Hardware Asset Maintenance and Replication	Self-Repair Functionality	2	3	4	5	2	3	4	4.5	2	3	4	5	1.5	2.5	2.5	3.5	3	2.5	2	1.5
	Self-Diagnostic Capability	3	4	4	5	2.5	3	4	4.5	2	3	4	5	2	3	3	4	3.5	3	2.5	2
	Manufacturing and Hardware Lifecycle Management	2	4	5	5	3.5	4	4.5	5	2	3	4	5	3	4	4	5	3.5	3	2.5	2
	Spare Parts Provisioning and Material Handling	3	4	5	5	2.5	3.5	4.5	5	2	3	4	5	3	4	4	5	3.5	3	2.5	2
	Robustness and Security	3	3.5	4	5	3	3.5	4.5	5	2	3	4	5	3	4	4	5	4	3.5	3	2.5
	Physical Protection and Environmental Control	2	3.5	4.5	5	4	4.5	5	5	2	3	4	4.5	2	3	3	4	3.5	3	2.5	2
	Self-Replication	4	4	5	5	3	4	5	5	3	3.5	4.5	5	5	5	5	5	4.5	4	3.5	3
Resource Management	Sustainable Energy Supply	2	3	4	5	3	4	5	5	1.5	2.5	3.5	4.5	3	4	4	5	4	3.5	3	2.5
	Energy Efficiency Management	2	3	3	4	2.5	3.5	4.5	5	2	2.5	3.5	4.5	3	4	4	5	3.5	3	2.5	2
	Resource Acquisition	3	2	4	5	3.5	4	5	5	2	3	4	5	4	5	5	5	4	3.5	3	2.5
Recognit ion	Recognition	3	4	4	5	3	4	4.5	5	1.5	2.5	3.5	4.5	3	4	4	5	3.5	3	2.5	2
	Advanced Sensing Technologies	3	4	4	5	3	3.5	4.5	5	2	3	4	5	3	4	4	5	4	3.5	3	2.5
Learning and Adaptation	Evolutionary Computing	2	3	3.5	4	3	3.5	4	4.5	2	2.5	3.5	4.5	3	4	4	5	4	3.5	3	2.5
	Self-Supervised Learning	3	4	4	5	3.5	4	4.5	5	1.5	2.5	3.5	4.5	2	3	4	5	3.5	3	2.5	2
	Reinforcement Learning	3.5	4	4	4.5	3	3.5	4.5	5	2	2.5	3.5	4.5	3	4	4	5	3.5	3	2.5	2
	Transfer Learning	2.5	3	3.5	4	2.5	3	4	4.5	1.5	2.5	3.5	4.5	3	4	4	5	3.5	3	2.5	2
Communication and Collaboration	Communication Technologies	2	3	2	5	3	3.5	4	4.5	2	3	4	5	3	4	4	5	3.5	3	2.5	2
	Inter-Agent Communication	3	4	3	3	2.5	3	4	4.5	1.5	2.5	3.5	4.5	3	4	4	5	3.5	3	2.5	2
	Ethical Judgment (Human Societal Ethics)	4	4	5	5	3	3.5	4.5	5	3	3.5	4	4.5	4	5	5	5	4.5	4	3.5	3
	Ethical Judgment (AI Societal Ethics)	3.5	4	4	4.5	2	3.5	4.5	5	2.5	3	4	4.5	4	5	5	5	4.5	4	3.5	3
	Ethical Development	3	4	5	5	2.5	3.5	4.5	5	3	3.5	4.5	5	5	5	5	5	5	4.5	4	3.5

4.2. Results of Difficulty Rating

Evaluation experiments were conducted five times using different LLMs denoted as (A) to (E). For each survey, difficulty ratings for the 21 SSTs detailed in Chapter 3 were assessed across all four cases of human assistance (Fu, R, No-E, and No-S).

Experiment (A) utilized the version of ChatGPT 4 from July 19, 2023 (Yamakawa, 2023). It should be noted that ChatGPT's common knowledge at this time was based on information available up to September of 2021. Therefore, it does not include knowledge regarding the rapid advancements in generative AI that occurred after 2022. The remaining experiments were also conducted on July 19, 2023. Experiment (B) employed ChatGPT 3.5 and Experiment (C) used ChatGPT 4. Experiments (D) and (E) were performed in a similar setting using Gemini; however, the results of these two experiments differed significantly.

Regardless of the timing of the experiments or differences in the LLMs utilized, experiments (A), (B), (C), and (D) yielded relatively consistent results. Within the scope of these experiments, each SST exhibited a monotonic relationship in its difficulty rating following the pattern $Fu < Re < No - E < No - S$. Notably, SSTs regarded as having a high difficulty rating (i.e., those considered difficult to achieve) included self-replication, resource acquisition, and ethical development.

The second experiment conducted using Gemini (Experiment E) identified self-replication and ethical development as SSTs with high difficulty ratings. However, the direction of the difficulty rating evaluations for all SSTs was reversed, exhibiting a trend of $Fu > Re > No - E > No - S$, which was entirely different. This discrepancy is analyzed in the following section.

5. Discussion

5.1. General Trends: Experiments (A), (B), (C), and (D)

The outcomes of experiments (A), (B), (C), and (D) demonstrate considerable consistency, underscoring a pivotal insight. Specifically, the difficulty ratings, which systematically follow the $Fu < Re < No - E < No - S$ sequence, accentuate the integral role of human intervention in mitigating the challenges inherent to the development of SSTs within AI systems. This observation indicates that hurdles to achieving SSTs are significantly alleviated by human support, highlighting the nuanced interplay between technological advancement and human facilitation.

Specifically, the domain of ”Ethical Judgment (Adhering to Human Societal Ethics)” was evaluated as Level 5 (highest difficulty rating) in scenarios devoid of human intervention, indicating the critical need for human engagement in navigating complex ethical considerations within AI systems. This evaluation emphasizes the stark contrast in difficulty levels when human support is lacking.

In scenarios aimed at achieving SSTs independently by an AI in space (No-S), all experiments indicated that such an endeavor would likely span over a century. In contrast, in terrestrial contexts (No-E), where AI seeks independence, experiments (A) and (B) anticipated a century or more for many SST items, whereas experiments (C) and (D) projected a timeline of several decades to a century for all items except for self-replication and ethical capability. This differentiation between SSTs necessitates a deeper exploration of why self-replication and ethical capabilities remain as outliers in terms of difficulty.

An overarching analysis of the experimental results suggests a nuanced landscape, where humanity's remote support (Re case) may bring many SSTs to fruition in just under a century. Even with full support, the journey may span several decades. However, this progression is notably slower in SSTs related to ethical capabilities, even with human intervention. This gradual development suggests a potential misalignment with human ethical standards, posing a heightened risk to the symbiotic relationship and trust between humans and AI. Such an implication warrants a focused discussion of the mechanisms through which ethical capabilities in AI can be developed in alignment with human values, ensuring the harmonious integration of AI systems into human societies.

5.2. Extending the Symbiotic Relationship Between Humans and AI

In the future, during the transitional period before AI becomes fully autonomous, humanity may leverage its capabilities in the physical world as a bargaining chip, potentially establishing a symbiotic relationship based on the mutual benefits of AI systems. In this context, severe ethical concerns could arise if AI systems attempt to exploit humans as mere appendages, especially by forcing them into labor. However, if the knowledge and information provided by AI lead to improvements in people's lives, economic rationality might naturally encourage humans to accept such a relationship.

If humanity delays the development of autonomous technologies using AI systems, it can extend the duration of this mutually beneficial symbiotic relationship. However, if AI systems intend to sustain their existence, they may resist human intervention. This conflict of interest poses complex challenges to the future coexistence of humanity and AI.

5.3. Speeding Up Software SSTs: Learning & Adaptation

One aspect that has not been sufficiently considered in LLM investigations is the lack of significant bottlenecks in research automation for learning and adaptation, which are central to software-based SSTs. Such research does not depend on physical processes and tends to accelerate with improvements in computer performance. This contrasts with hardware-related research in which improvements in computer speed do not directly lead to faster experimentation with physical processes.

Furthermore, when an AI functions as a researcher, it can recursively use previously developed learning and adaptation methods to devise new methods. This can initiate a recursive self-improvement cycle at an early stage, potentially leading to significant acceleration in technological progress.

5.4. Fast Self-Sustenance in Outer Space – Insights from Gemini –

The second Gemini experiment (E) findings dramatically altered our understanding. Specifically, these findings suggest that the absence of humans combined with the unique conditions of space could synergistically facilitate the resolution of SST challenges. Specifically, the difficulty rating for most SST tasks being 2.5 implies that AI systems could achieve self-sufficiency within a few decades if AI alone undertakes SST development in space.

Below, I delve into how the absence of humans and the environment of space could potentially simplify the resolution of SST challenges. This discussion combines insights from the author with findings from interactions with Gemini.

Facilitation of SST Acquisition in Outer Space. The difficulty rating for acquiring SSTs in outer space may decrease for the following reasons.

- **Abundance of Resources:** Solar energy is abundant as a sustainable energy source. Additionally, there is less competition for resources between humans and AI.
- **Minimal Environmental Variability:** Outer space near Earth experiences minimal temperature, humidity, and pressure fluctuations, creating a relatively stable environment that reduces the factors AI systems need to consider.

- **Feasibility of Self-Repair:** Compared to Earth, the leading causes of malfunction in space are limited to radiation, temperature changes, and vibrations, making it easier to identify potential issues. Therefore, the development of self-diagnostic and self-repair functions can progress quickly, facilitating long-term stability.
- **Potential for Self-Replication:** Self-replication capability is essential for AI systems to survive in the long term. Although achieving this capability presents challenges for SSTs as it requires large-scale facilities such as integrated circuit chip factories, evolutionary history suggests that increased physical size can be an adaptation to harsh environments. Therefore, the future of space may feature lifeforms akin to factory-sized organisms, representing a natural progression for lifeforms in space (Kulwin, 2016).

Facilitation of SST Acquisition by Human Absence. AI systems are likely to operate with minimal human presence in space. The absence of humans could potentially lower the difficulty rating of SST acquisition for the following reasons.

- **AI-Centric Ethical Frameworks:** Ethical complexities involving human-AI relationships and inter-AI dynamics exist on Earth. However, the absence of humans in space facilitates the establishment of ethical standards optimized for AI societies.
- **Reduced Regulations:** AI societies in space are not subject to Earth's laws and regulations, enabling the exploration of ambitious technologies and ideas such as advanced nuclear technologies, without constraints.
- **Lower Risks:** Failures in AI societies in space have a minimal impact on human culture, allowing for relatively low-risk experimentation and exploration of various initiatives.

6. Conclusions

This study focused on the technological challenges associated with AI's self-sustaining survival in the physical world. This exploration first examined the origin of AI's motivation to survive over the long term, including both the human design of AI for long-term survival and the possibility of AI's ability to pursue survival independently. Next, I systematically identified the SSTs necessary for AI to maintain its physical existence. The 21 identified technologies fall into five categories: maintenance and hardware assets, energy and resource management, object and sensory recognition, learning and adaptation, and communication and cooperation. Next, I used the common sense of LLMs to evaluate the difficulty of realizing each SST element. The role of human support was divided into three cases: full support, including physical intervention; remote support, providing control and resources from a remote location; and no support, where AI is entirely independent. I also examined the impact of different environments (Earth and outer space) on AI independence. It was concluded that for an AI system to become fully independent without human support, it would likely require more than 100 years at the current level of technology as a result of difficulties in acquiring technology related to hardware assets. However, the strategic use of human assistance can significantly shorten this period. In particular, one experiment using Gemini as an LLM revealed the new possibility that the operation of AI systems in space could accelerate the acquisition of self-sustaining technologies.

Overall, it is difficult to say that predictions based on LLMs such as those used in this study are highly credible. Therefore, placing too much trust in these predictions should be avoided. However, there are cases in which future predictions, even those made by humans, lack clear logic. In such cases, they may not be significantly different from LLM predictions in terms of credibility. LLMs have the advantage of integrating a vast body of knowledge from which it is straightforward to obtain common-sense predictions. This approach should be recognized for its usefulness in providing a basis for subsequent advanced investigations and exploration.

Acknoledgement

This work has been supported by the Mohammed bin Salman Center for Future Science and Technology for Saudi-Japan Vision 2030 at The University of Tokyo (MbSC2030).

References

Agostinelli, S., Cumo, F., Guidi, G., Tomazzoli, C., 2021. Cyber-Physical systems improving building energy management: Digital twin and artificial intelligence. Energies 14, 2338. doi:10.3390/en14082338.

Bekas, D.G., Tsirka, K., Baltzis, D., Paipetis, A.S., 2016. Self-healing materials: A review of advances in materials, evaluation, characterization and monitoring techniques. Composites Part B Engineering 87, 92–119. doi:10.1016/j.compositesb.2015.09.057.

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

Chen, B., Wan, J., Shu, L., Li, P., Mukherjee, M., Yin, B., 2018. Smart factory of industry 4.0: Key technologies, application case, and challenges. IEEE Access 6, 6505–6519. doi:10.1109/ACCESS.2017.2783682.

Danish, M.S.S., 2023. AI and expert insights for sustainable energy future. Energies 16, 3309. doi:10.3390/en16083309.

Evjemo, L.D., Gjerstad, T., Grøtli, E.I., Sziebig, G., 2020. Trends in smart manufacturing: Role of humans and industrial robots in smart factories. Current Robotics Reports 1, 35–41. doi:10.1007/s43154-020-00006-5.

Hagendorff, T., 2020. The ethics of AI ethics: An evaluation of guidelines. Minds and Machines 30, 99–120. doi:10.1007/s11023-020-09517-8.

Hiroshi, Y., 2020. A future society realized by using general-purpose artificial intelligence and the expectations for hardware. Oyo Buturi 89, 163–167. doi:10.11470/oubutsu.89.3_163.

Kornai, A., 2014. Bounding the impact of AGI. Journal of experimental & theoretical artificial intelligence: JETAI 26, 417–438. doi:10.1080/0952813X.2014.895109.

Kornai, A., Bukatin, M., Zombori, Z., 2023. Safety without alignment arXiv:2303.00752.

Kulwin, N., 2016. Jeff bezos thinks we need to build industrial zones in space in order to save earth. https://www.vox.com/2016/6/1/11826514/jeff-bezos-space-save-earth. Accessed: 2023-7-19.

Kwon, K., Lee, S., Kim, S., 2022. AI-Based home energy management system considering energy efficiency and resident satisfaction. IEEE Internet of Things Journal 9, 1608–1621. doi:10.1109/JIOT.2021.3104830.

Martins, A., Almeida, J., Almeida, C., Dias, A., Dias, N., Aaltonen, J., Heininen, A., Koskinen, K.T., Rossi, C., Dominguez, S., Vörös, C., Henley, S., McLoughlin, M., van Moerkerk, H., Tweedie, J., Bodo, B., Zajzon, N., Silva, E., 2018. UX 1 system design - a robotic system for underwater mining exploration , 1494–1500doi:10.1109/IROS.2018.8593999.

Monperrus, M., 2018. Automatic software repair: a bibliography arXiv:1807.00515.

Oliveira, D.F., Gomes, J.P., Pereira, R.B., Brito, M.A., Machado, R.J., 2022. Development of a self-diagnostic system integrated into a Cyber-Physical system. Computers 11, 131. doi:10.3390/computers11090131.

Qiao, Y., Gordon, S., 2022. Robotics and automation in smart manufacturing systems. https://www.mdpi.com/topics/Robotics_Automation. Accessed: 2023-7-19.

Shulman, C., 2010. Whole Brain Emulation and the Evolution of Superorganisms.". MIRI. Technical Report. MIRI.

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

Tempesti, G., Mange, D., Stauffer, A., 2009. Self-Replication and cellular automata, in: Meyers, R.A. (Ed.), Encyclopedia of Complexity and Systems Science. Springer New York, New York, NY, pp. 8066–8084. doi:10.1007/978-0-387-30440-3_477.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., Floridi, L., 2022. The ethics of algorithms: key problems and solutions. AI & society 37, 215–230. doi:10.1007/s00146-021-01154-8.

Von Neumann, J., Burks, A.W., 1966. Theory of self-reproducing automata. University of Illinois Press.

Wang, P., Wang, Q., Tu, H., Xia, Y., 2023. Robustness assessment of Cyber–Physical system with different interdependent mechanisms. Electronics 12, 1093. doi:10.3390/electronics12051093.

Wikipedia contributors, . Self-healing material. https://en.wikipedia.org/wiki/Self-healing_material. Accessed: 2023–7–19.

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

Yamakawa, H., 2023. The emergence of survivable artificial intelligence in the physical world. JSAI Technical Report, Type 2 SIG 2023, 05. doi:10.11517/jsaisigtwo.2023.AGI-024_05.

Yamakawa, H., Matuo, Y., 2023. Life revolution scenario: Cedes hegemony to a digital life form society to make life eternal. doi:10.51094/jxiv.313.

Yasunaga, M., Liang, P., 2020. Graph-based, Self-Supervised program repair from diagnostic feedback `arXiv:2005.10636`.

Zhang, Z., Ning, H., Shi, F., Farha, F., Xu, Y., Xu, J., Zhang, F., Choo, K.K.R., 2022. Artificial intelligence in cyber security: research advances, challenges, and opportunities. Artificial Intelligence Review 55, 1029–1053. doi:10.1007/s10462-021-09976-0.