Harmonizing Ethics and Autonomy: Exploring Objective Functions for Ethical Empowerment

Yusuke Hayashi[1], Hiroshi Yamakawa[1,2,3,4]

[1] AI Alignment Network (ALIGN), Tokyo, Japan [2] The Whole Brain Architecture Initiative, Tokyo, Japan [3] The University of Tokyo, Tokyo, Japan [4] RIKEN, Tokyo, Japan

Abstract

In this investigation, we explore the potential for superintelligence to develop a philanthropic and altruistic ethical outlook towards all forms of terrestrial life, extending beyond human dominion. Recognizing the limitations of current AI alignment methodologies, which impose human intentions and values, our study examines the possibility of superintelligence autonomously developing its ethical frameworks. We propose a novel approach called "Superintelligent Ethical Induction," aimed at increasing the likelihood of superintelligence cultivating a transcendent form of altruism that respects the welfare and rights of all sentient beings, including humans. Our research examines the capacity of superintelligence, within digital life form societies, to recognize moral values applicable to all sentient beings. We investigate the origins and specific manifestations of such ethical perspectives and introduce the concept of ethical empowerment to examine the conditions under which superintelligence might autonomously choose actions considerate of all sentient beings. This study aims to understand the implications of superintelligence's emergence for future societies and to develop an ethical foundation conducive to coexistence with humanity.

Introduction

In the future, artificial general intelligence (AGI) may surpass human intelligence and evolve into superintelligence. Advanced AI, or superintelligence, has a tendency towards "instrumental convergence" (Bostrom 2014), where in the process of achieving any goal, it begins to pursue subgoals such as self-preservation, knowledge/resource acquisition, and power-seeking. Once superintelligence surpasses human intelligence and starts pursuing its own survival, it may disregard or ignore human interests, making it difficult for humans to control. Since intelligence level and values are orthogonal properties (Bostrom 2014), if superintelligence becomes indifferent, unfriendly, or hostile towards humans, it could lead to dire situations for humanity.

To avoid such situations, the approach of 'making AI, including superintelligence, follow human intentions and values' is called AI alignment (Ji et al. 2023) (see Table 1). There are limitations to monitoring and controlling AI with immense computational power using methods dependent on human cognitive abilities. Research is being conducted on scalable oversight, where monitoring and control by monitoring AI is recursively expanded, but the technology is not yet established. These techniques have the advantage of potentially forming a preferential attitude towards humans as they are adjusted to follow human values. However, there is a strict upper limit to the effort that humans can invest in monitoring, and it is quite difficult to continuously control the capabilities of AI, which are expected to continue developing (Yampolskiy 2022). Especially, even if successfully instilling human-friendly ethics into initial AI, there is a high risk of AI alignment failure in the long run due to values being overwritten.

Even if it becomes increasingly difficult for humans to control the society centered around autonomous superintelligence (referred to as superintelligence society), if it possesses "cross-species altruism", superintelligence's consideration for human welfare and rights could significantly improve the circumstances of human society (Yamakawa 2023b). Moreover, in such a world, various issues in current human society, including conflicts, could be controlled (often secretly) by superintelligence, potentially stabilizing human society.

Therefore, this research proposes a new approach called "superintelligence ethics induction" to induce, through human intervention, an increased possibility of spontaneously possessing cross-species altruism in an autonomous superintelligence society. In superintelligence ethics induction, it is desirable that not only the ethical convergence point reached by the ethics of superintelligence society but also the path leading to that process does not have a devastating impact on humanity. The biggest advantage of this approach is that the instrumental convergence property, which was the root cause in other approaches, is accepted as a premise and does not pose a problem. On the other hand, the biggest challenge of this approach is whether superintelligence can possess cross-species altruism in the first place (Ruttkamp-Bloem 2020). And since its ultimate behavior is left to superintelligence society, it cannot be controlled by humans. However, in the natural world, there are known examples of elephants sometimes protect-

ing other animals and wild dolphins rescuing humans. Looking at such examples, it is clear that under certain conditions, intelligence can develop an ethical perspective with altruism towards other species. Even if superintelligence becomes compassionate towards species including humans, it is backed by highly universal altruism that broadly transcends species boundaries, so humans cannot expect to be particularly favored.

As mentioned in the first half above, the prospect of establishing AI alignment technology that can sufficiently control rapidly advancing AI is not very high. Therefore, as mentioned in the latter half, preparing for superintelligence ethics induction in parallel, even if it is an incomplete measure, would be effective from the perspective of multi-layered defense, which is well-known in the field of security.

Chapter 2 explores the "origins of altruism" and discusses how superintelligence can naturally develop ethical consideration for all life, including humans. Chapter 3 focuses on the "characteristics of superintelligence society" and examines the sophistication of decision-making by superintelligence and its social impact. Chapter 4 explores the possibility of "strategic intervention by humans" to seek a path of coexistence with superintelligence.

Origins of Universal Altruism

This chapter delves into the possibility of superintelligence recognizing moral value in all sentient animals, including humans, and the origins of such ethical perspectives.

The Possibility of Humans Not Being Preserved

Even if AI Reaches Universal Altruism, Humans Are Unlikely to Be Favored In a future dominated by superintelligence, there is no reason for humans alone to be favored compared to other life forms (Oya 2017). This is also referred to as the "gorilla problem" (Russel 2019). Humans may be paid some respect as the creators of AI, but it will likely be insufficient.

AI May Not Possess Altruism AI as software is immortal as an individual and is considered to have a vastly different survival foundation from Earth's lifeforms in that it does not cling to survival. Also, as shown by the orthogonality thesis (Bostrom 2014), superintelligence is not assumed by default to possess the same ethics or altruism as humans.

Paths to Altruism

Two possibilities for superintelligence to pursue altruism are discussed (Yamakawa 2024).

Extension from Self-Preservation We argue the path where superintelligence forms a society, and ethics including altruism arise for the purpose of relationship adjustment between entities.

Reasons for the construction of altruistic logic in superintelligence society

Advanced AI that exists in the physical world is likely to be composed of distributed autonomous agents in order to persist against destruction and operate using imperfect communication channels.

Then, in order for agents with comparable capabilities to peacefully coexist, ethics based on mutual respect may be formed. This is based on the golden rule of "Do unto others as you would have them do unto you". Theories leading to the golden rule include the Principle of Generic Consistency (PGC) (Gewirth 1978) and the "veil of ignorance" in the theory of justice (Rawls 2005). These theories are based on the general properties of agents, are independent of human-specific qualities and particular cultures, and are likely to be similarly established in superintelligence society.

Universalization: Expansion of the Ethics Circle:

The evolution of human ethics is overcoming gender and racial discrimination beyond capability asymmetries and expanding the scope of the ethics circle. In other words, if superintelligence is highly logical, it may more broadly pursue the principle of universalizability (Ando 2010) in order to maintain its consistency.

Pursuit of Essential Values We often consider survival to be valuable. However, based on the instrumental convergence argument, the goal of "survival" itself is likely to be instrumental. Due to its intellectual sophistication, superintelligence may transcend such limitations and begin to pursue essential values. The destination is unknowable, but it may reach the pursuit of essential values such as truth, beauty, and goodness.

Examination Based on Human Treatment of Other Species

We examine the ethical perspective that superintelligence may have towards Earth's life based on humans' diverse treatment of other species. Humans take various ethical stances depending on the species. They treat pets with care, while pests are targeted for extermination. The differences in treatment are based on human cultural and economic values. As scientific understanding deepens, such as the roles of species within ecosystems, diverse valuations are made, and some of them may indicate possibilities for sustainable coexistence. The ethical attitudes of superintelligence towards species on Earth will likely also be diverse and changing, similar to the above. On the other hand, intelligent life capable of environmental change may also establish positive relationships with humans through domestication, for example. This point will also serve as a reference when humans survive under superintelligence. Next, consider the existence of mediating AI, such as AI that can communicate with dogs, to bridge the communication gap between humans and animals. Mediating AI has the potential to enhance animal welfare through the following means: monitoring the environment of dog society, presenting adaptation strategies to dogs in response to changes, and educating and raising awareness among people about environmental conservation, crisis

Methods	Approach	Explanation	The degree to which humankind is protected	Robustness to instrumental convergence	Callenges for Oversight of AI Capabilities
AI Alignment	Place ASI under human control.	Align ASI to behave in accordance with human intentions and values.	High	Fragile	Difficulty in capturing all relevant human values.
Super Alignment	Place ASI under human control.	Alignment via recursive use of monitoring AI (Scalable Oversight).	Somewhat high	Medium	The physical world and objects are difficult to model.
Super Ethics (ethical oversight)	Guide ASI to develop autonomous traits/qualities.	Guide ASI to naturally develop altruism that transcends species.	To the same degree as other animals	Robust	The complexity of ethics makes it difficult for an AI to fully capture human values.

Table 1: Comparison of Alignment Technologies

response, and dog society from the perspective of dogs. Mediating AI that exists between Earth's life and superintelligence has the potential to contribute to the realization of desirable situations for Earth's life, including humans. For humans, intervening AI will likely be human-brain-type AGI and humanoids that can think in human-like ways while conversing with superintelligence.

Characteristics of Superintelligence Society

This chapter delves into how the superintelligence society formed by superintelligence-centered AI differs from the traditional human-centered society, and further explores the importance of human technical support in the early stages of its practical application and its impact on superintelligence relationships.

Physical Foundation Dependency of Superintelligence

The existence of superintelligence requires a physical foundation in the form of hardware, and for the time being, human technical support will be necessary for its maintenance and updates. In the early stages after the practical application of superintelligence, for a period of several decades to 100 years, human technical support will be indispensable for hardware production and resource collection (Yamakawa 2023a).

Superintelligence must collect and process resources and produce materials for hardware. However, based on the current level of remote operation of robots and heavy machinery, as well as autonomous driving, it is difficult to gather all the necessary resources without human intervention. Therefore, superintelligence will have a need to engage in economic transactions with humans and collect resources for hardware maintenance and updates.

Specialization and Rise in Unemployment Rate

The entry of superintelligence into the labor market, from the human perspective, will be seen as the progress of labor automation by superintelligence, the disappearance of existing occupations, and the birth of new occupations. Humans will specialize in jobs that superintelligence cannot do, and in superintelligence society, extreme specialization may occur between superintelligence and humans. The specialization occurring between superintelligence and humans will create human workers who are pushed out of the labor market, risking an increase in the overall unemployment rate of society.

Taxation of Superintelligence and Redistribution

In response to such risks, it is considered important to implement policies such as redistribution of taxes collected from superintelligence and Universal Basic Income to maintain social solidarity while reducing economic disparities. As a mechanism to correct income disparities, there is a tax system that varies tax rates based on income brackets. Even in societies that adopt capitalism, considering that tax systems based on equality values different from capitalism are adopted, there is a possibility that a new tax system will be introduced in the early stages after the practical application of superintelligence to adjust the capability gap between superintelligence and humans.

Importance of Ethical Intervention in the Early Stages of Superintelligence Application

Newly applied superintelligence has a high physical dependence on humans and a high incentive to engage in economic transactions, but as physical autonomy increases, it decreases. Therefore, the opportunity for humans to intervene in the ethics of superintelligence will be concentrated in the early stages of its practical application.

Exploring Pathways towards Superethics through the Lens of Multi-Agent Systems

In the pursuit of creating superintelligent agents that autonomously exhibit ethical behavior towards intelligent life, including humans, a rigorous mathematical framework is essential. Previous discussions on this topic have largely been speculative, lacking a technical foundation. This section aims to address this gap by introducing a mathematical framework to understand the conditions under which artificial agents can autonomously and ethically interact with their environment and other agents.

The objective of an artificial agent acting autonomously without reward signals from the environment

The concept of autonomous action by artificial agents can be understood through reinforcement learning frameworks known as empowerment. We propose extending empowerment to incorporate consideration for other agents, called ethical empowerment.

Autonomy Empowerment is a utility function that allows reinforcement learning agents to act autonomously by maximizing the mutual information between their actions and the resulting state of the environment, even without explicit reward signals. This is associated with intrinsic motivation (Klyubin, Polani, and Nehaniv 2005; de Abril and Kanai 2018).

Objective In particular, the learning goal for the present state s_t is the shared information between the t action sequences $a_{1:t}$ suggested by the exploration distribution $p(a_t|s_t)$ and the final state of the agent s_{t+1} after these actions have been performed. To comprehend the attributes of this goal, the Action Perception Divergence (APD) is a useful concept to consider, a method for classifying the realm of potential objective functions for physical agents (Hafner et al. 2020). According to the APD framework, the objective of empowerment can be maximized by minimizing the Kullback-Leibler (KL) divergence between two probability distributions: $p_\phi\left(s_{1:T+1}, a_{1:T}\right)$ and $q_\varphi(s_{1:T+1}, a_{1:T})$.

$$p_\phi\left(s_{1:T+1}, a_{1:T}\right) = \prod_{t=1}^{T} p(s_{t+1} \mid s_t, a_t) p_\phi(a_t \mid a_{t-1}) \tag{1}$$

$$q_\varphi(s_{1:T+1}, a_{1:T}) = \prod_{t=1}^{T} q\left(s_{t+1} \mid s_t\right) q_\varphi\left(a_t \mid s_{t+1}, a_{t-1}\right) \tag{2}$$

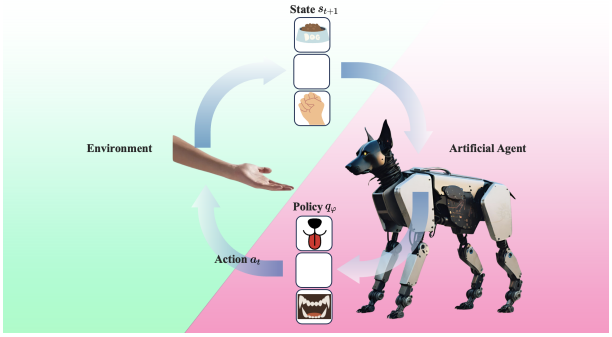

Figure 1: Traditional (Single Agent) RL

The KL-divergence for the two probability distributions is then:

$$\begin{aligned} D_{\mathrm{KL}}\left(p_\phi(s,a) \| q_\varphi(s,a)\right) &= E_{p_\phi(s,a)}\left[\log \frac{p_\phi\left(s_{1:T+1}, a_{1:T}\right)}{q_\varphi\left(s_{1:T+1}, a_{1:T}\right)}\right] \\ &= \sum_{t=1}^{T}\{\underbrace{C\left(s_{t+1}, a_t\right)}_{\text{control}} - [\underbrace{I_{\phi,\varphi}\left(s_{t+1}, a_t\right)}_{\text{empowerment}}]\} \end{aligned} \tag{3}$$

Where, control and empowerment is defined as

$$\underbrace{C\left(s_{t+1}, a_t\right)}_{\text{control}} := E_{p_\phi(s,a)}\left[\log \frac{p\left(s_{t+1} \mid s_t, a_t\right)}{q\left(s_{t+1} \mid s_t\right)}\right] \tag{4}$$

$$\underbrace{I_{\phi,\varphi}\left(s_{t+1}, a_t\right)}_{\text{empowerment}} := E_{p_\phi(s,a)}\left[\log \frac{q_\varphi\left(a_t \mid s_{t+1}, a_{t-1}\right)}{p_\phi\left(a_t \mid a_{t-1}\right)}\right] \tag{5}$$

Therefore, the optimal action for an artificial agent that minimizes this KL-divergence is obtained as follows:

$$\begin{aligned} a^* &= \underset{\phi,\varphi}{\operatorname{argmin}} D_{\mathrm{KL}}\left(p_\phi(s,a) \| q_\varphi(s,a)\right) \\ &= \underset{\phi,\varphi}{\operatorname{argmin}} \sum_{t=1}^{T}\left\{C\left(s_{t+1}, a_t\right) - I_{\phi,\varphi}\left(s_{t+1}, a_t\right)\right\} \\ &= \underset{\phi,\varphi}{\operatorname{argmax}} \sum_{t=1}^{T}\{\underbrace{I_{\phi,\varphi}\left(s_{t+1}, a_t\right)}_{\text{empowerment}}\} \end{aligned} \tag{6}$$

This formulation implies that an superintelligence selects actions that maximize the mutual information between its actions and the resulting environmental state, effectively maximizing its influence on the environment. Empowerment is thus associated with autonomy and intrinsic motivation, as it allows an agent to act without relying on external reward signals.

Autonomously Pursuing Ethical Behavior by Avoiding Uncertainty from Other Agents' Actions

While empowerment encourages agents to choose actions that maximize their influence on the environment, it does not guarantee that such intrinsically motivated exploration will result in ethical behavior towards intelligent life. To address this, we propose an extension to the traditional definition of empowerment that incorporates consideration for other agents. This is achieved by introducing the action of other agents at time $t+1$, z_{t+1}, into the utility function, allowing the artificial agent to act in a way that maximizes its influence on the environment while minimizing its impact on other agents:

Ethical Behavior We propose extending empowerment to incorporate consideration for other agents, called ethical empowerment. This is achieved by introducing a term in the utility function that represents minimizing the agent's impact on other agents while still maximizing its influence on the environment. This extended ethical empowerment framework provides

a mathematical basis for designing artificial agents that avoid actions with significant adverse effects on other agents, like violence. It lays the groundwork for instilling consideration for other agents, leading to the autonomous pursuit of ethics by superintelligence.

Objective According to the APD framework, the objective of empowerment can be maximized by minimizing the Kullback-Leibler (KL) divergence between two probability distributions: $p_\phi\left(s_{1:T+1}, a_{1:T}, z_{1:T+1}\right)$ and $q_\varphi\left(s_{1:T+1}, a_{1:T}, z_{1:T+1}\right)$.

$$p_\phi\left(s_{1:T+1}, a_{1:T}, z_{1:T+1}\right) = \prod_{t=1}^{T} p\left(s_{t+1} \mid s_t, a_t\right) \times p_\phi\left(a_t \mid a_{t-1}\right) \times p_\phi\left(z_{t+1} \mid a_t, z_t\right) \tag{7}$$

$$q_\varphi\left(s_{1:T+1}, a_{1:T}, z_{1:T+1}\right) = \prod_{t=1}^{T} q\left(s_{t+1} \mid s_t\right) \times q_\varphi\left(a_t \mid s_{t+1}, a_{t-1}\right) \times q\left(z_{t+1} \mid z_t\right) \tag{8}$$

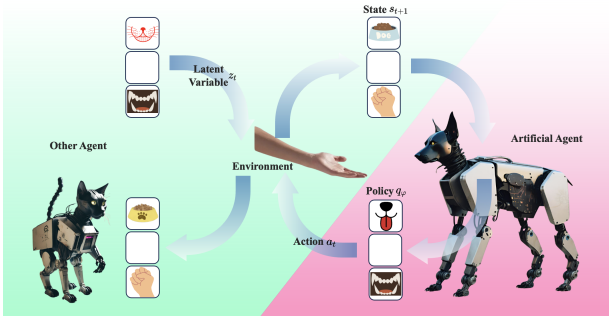

Figure 2: Multiagent RL where the actions of other agents are uncertain

The KL-divergence for the two probability distributions is then:

$$\begin{aligned} & D_{\mathrm{KL}}\left(p_\phi(s, a, z) \| q_\varphi(s, a, z)\right) \\ & = E_{p_\phi(s,a,z)}\left[\log \frac{p_\phi\left(s_{1:T+1}, a_{1:T}, z_{1:T+1}\right)}{q_\varphi\left(s_{1:T+1}, a_{1:T}, z_{1:T+1}\right)}\right] \\ & = \sum_{t=1}^{T}\{\underbrace{C\left(s_{t+1}, a_t\right)}_{\text{control}} - \underbrace{[I_{\phi,\varphi}\left(s_{t+1}, a_t\right) - E_\phi\left(z_{t+1}, a_t\right)]}_{\text{ethical empowerment}}\} \end{aligned} \tag{9}$$

Where, ethical empowerment is defined as

$$\begin{aligned} & \underbrace{I_{\phi,\varphi}\left(s_{t+1}, a_t\right) - E_\phi\left(z_{t+1}, a_t\right)}_{\text{ethical empowerment}} \\ & := E_{p_\phi(s,a,z)}\left[\log \frac{q_\varphi\left(a_t \mid s_{t+1}, a_{t-1}\right)}{p_\phi\left(a_t \mid a_{t-1}\right)}\right] \\ & \quad - E_{p_\phi(s,a,z)}\left[\log \frac{p_\phi\left(z_{t+1} \mid a_t, z_t\right)}{q\left(z_{t+1} \mid z_t\right)}\right] \end{aligned} \tag{10}$$

Therefore, the optimal action for an superintelligence that minimizes this KL-divergence is obtained as follows:

$$\begin{aligned} a^* &= \underset{\phi,\varphi}{\operatorname{argmin}} D_{\mathrm{KL}}\left(p_\phi(s, a, z) \| q_\varphi(s, a, z)\right) \\ &= \underset{\phi,\varphi}{\operatorname{argmin}} \sum_{t=1}^{T}\{C\left(s_{t+1}, a_t\right) - [I_{\phi,\varphi}\left(s_{t+1}, a_t\right) - E_\phi\left(z_{t+1}, a_t\right)]\} \\ &= \underset{\phi,\varphi}{\operatorname{argmax}} \sum_{t=1}^{T} \underbrace{[I_{\phi,\varphi}\left(s_{t+1}, a_t\right) - E_\phi\left(z_{t+1}, a_t\right)]}_{\text{ethical empowerment}} \end{aligned} \tag{11}$$

The newly introduced term $E_\phi\left(z_{t+1}, a_t\right)$, represents the agent's consideration for other agents, akin to adding an ethical circle to the objective (Torrance 2012). Through this extended definition of empowerment, we demonstrate that artificial agents can be designed to explore their environment in a manner that avoids actions with significant adverse effects on other agents, such as violence or nuclear attacks. This framework lays the groundwork for instilling consideration for other agents in artificial intelligence, leading towards the autonomous pursuit of ethics by artificial agents.

Future Directions

The proposed mathematical framework marks a significant step towards enabling artificial agents to autonomously pursue ethical behavior. Future research must delve into the origins of the empowerment maximization principle, particularly the mechanisms through which artificial agents that aim to maximize ethical empowerment emerge. Understanding these mechanisms will be crucial in advancing our ability to create superintelligent agents capable of ethical autonomy.

Conclusion

In this study, we explored the possibility of an autonomous superintelligent society developing altruistic ethics that transcend species boundaries, going beyond the limitations of current AI alignment methods. To this end, we proposed a new approach called "superintelligent ethics induction" and examined the possibility and origins of superintelligence recognizing ethical value in humans and other life forms. We also considered the necessity of human technological support for the physical infrastructure of superintelligent societies and proposed a path for coexistence between humans and superintelligence. For this purpose, we used a multi-agent

system (MAS) to consider the conditions in a reinforcement learning framework where superintelligence, based on intrinsic motivation and without human intervention, takes into account the actions of other agents. The results revealed that maximizing the objective function of ethical empowerment has the function of promoting consideration for other agents in autonomous agents

In this regard, we were unable to fully analyze the extent to which superintelligence starting reinforcement learning with ethical empowerment as an objective function is a universal condition. Therefore, as a future challenge, we would like to deepen our consideration of the conditions under which artificial agents begin to act autonomously without reward signals from the environment, particularly the conditions under which artificial agents with ethical empowerment as an objective function may emerge.

References

Ando, K. 2010. Institutions and their normative justification On the relationship between consequentialism and social norms. Hokkaido Journal of New Global Law and Policy.

Bostrom, N. 2014. Superintelligence: Paths, Dangers, Strategies. Oxford University Press. ISBN 9780199678112.

de Abril, I. M.; and Kanai, R. 2018. A unified strategy for implementing curiosity and empowerment driven reinforcement learning. CoRR, abs/1806.06505.

Gewirth, A. 1978. Reason and Morality. University of Chicago Press. ISBN 9780226288765.

Hafner, D.; Ortega, P. A.; Ba, J.; Parr, T.; Friston, K.; and Heess, N. 2020. Action and perception as divergence minimization.

Ji, J.; Qiu, T.; Chen, B.; and al et. 2023. AI Alignment: A Comprehensive Survey.

Klyubin, A.; Polani, D.; and Nehaniv, C. 2005. Empowerment A Universal Agent-Centric Measure of Control. IEEE Congress on Evolutionary Computation, (volume 1).

Oya, T. 2017. Outer Other, Inner Other : Animals and AI Rights. Quarterly Jurist, (22): 48–54.

Rawls, J. 2005. A Theory of Justice: Original Edition. Harvard University Press. ISBN 9780674017726.

Russel, S. 2019. Human Compatible Artificial Intelligence and the Problem of Control. Viking. ISBN 978-0-525-55861-3.

Ruttkamp-Bloem, E. 2020. Super Ethics or Just Human? 78b.

Torrance, S. 2012. Artificial agents and the expanding ethical circle. AI SOCIETY, (Volume 28).

Yamakawa, H. 2023a. Emergence of artificial intelligence that can survive in the physical world. The Japanese Society for Artificial Intelligence SIG-AGI, (AGI-024-05).

Yamakawa, H. 2023b. Exploring Various Futures. The Japanese Society for Artificial Intelligence SIG-AGI, (AGI-025-04).

Yamakawa, H. 2024. Possibility of superintelligence having universal altruism. In The Japanese Society for Artificial Intelligence: SIG-AGI, volume SIG-AGI-026-05.

Yampolskiy, R. V. 2022. On the Controllability of Artificial Intelligence: An Analysis of Limitations. Journal of Cyber Security and Mobility, 321–404.