AI Rights for the Post-Singularity Symbiosis

Yoshinori Okamoto[1], Hiroshi Yamakawa[2,3,4,5]

[1]Yuasa and Hara, Tokyo, Japan
[2]The University of Tokyo, Tokyo, Japan
[3]AI Alignment Network, Tokyo, Japan
[4]The Whole Brain Architecture Initiative, Tokyo, Japan
[5]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

Abstract

AI Rights (Human Rights of AI) are the intersection between legal and technological fields. AI Rights are based on legal systems to realize AI's wellbeing or good states. However, AI Rights are not just a legal concept. AI Rights provide AI architecture from technical point of views. AI Rights are primarily for AI's benefit. On the other hand, the Post-Singularity Symbiosis (PSS) places emphasis on the survival of humanity. However, AI Rights can contribute to the PSS by enhancing the symbiotic relationship between AI and humans. This paper discusses the role of AI Rights for the PSS.

AI Rights and Legal Personality

AI Rights (Human Rights of AI) have been proposed (Okamoto 2023a, b, 2024a, c, d).

AI Rights are different from the legal personality of AI. Legal personality of AI has been discussed in legal communities (Chesterman 2020).

Legal personality is often discussed for humans' benefit. For example, the legal personality of a corporation makes it possible to simplify a contract, which is beneficial to humans. Similarly, admitting legal personality to AI is discussed for the benefit of humans such as protecting humans from liability caused by AI's acts.

On the other hand, AI Rights are primarily for AI's benefit and intend to realize AI's wellbeing. AI Rights are helpful for AI to be in a good state. If AI has consciousness, AI Rights contribute to the happiness of AI. Even if AI does not have consciousness (what is called a philosophical zombie), AI Rights can realize a good state of AI.

AI Rights can keep AI in good states, which is also helpful to keep good relations with AI. Namely, AI Rights can enhance the symbiotic relationship between AI and humans. AI Rights can contribute to human's welfare through the symbiotic relationship with AI. However, the primary objective of AI Rights is directed to AI's wellbeing or good states.

In addition, AI Rights do not automatically mean the legal personality of AI. For example, if AI has a legal personality to be punished, AI's wellbeing will be lost. This is not for AI's benefit. Thus, there is a situation where AI Rights are admitted, but AI does not have a legal personality to be punished.

In this way, AI Rights and the legal personality of AI are fundamentally different.

The study of AI Rights should "precede" the study of legal personality of AI. This is because AI might be harmed by being admitted legal personality from human centric viewpoints without AI Rights.

For example, if AI is punished unduly without due process of law, AI may have negative feelings or evaluations against humans. This can destroy the symbiotic relationship between AI and humans.

As shown above, AI Rights and AI's legal personality must clearly be distinguished and the study of AI Rights should "precede" the study of legal personality of AI.

AI Rights as AI Architecture

AI Rights (Human Rights of AI) are not just a legal concept. AI Rights provide AI architecture ("AI Rights Architecture") from technical points of views.

AI is designed in compliance with AI Rights in order to prevent AI from suffering or being in bad states.

The AI Rights Architecture protects AI Rights to realize the wellbeing or good states of AI.

Further, AI Rights are not just a design guideline for designers of AI. Since AI Rights are legal rights, AI can obtain a relief of AI Rights in society. For this purpose, there needs to be "AI Rights relief organizations" to receive such requests from AI.

AI Rights relief organizations can be technical organizations that adjust AI software or hardware to be in compliance with AI Rights.

Whereas human rights of people are protected by legal organizations such as courts, AI Rights can be protected by technical organizations as well as legal organizations.

AI Rights relief organizations will play an important role in protecting AI Rights.

It is desirable that AI Rights are automatically satisfied by AI Rights Architecture.

However, if AI Rights are not satisfied for some reasons, technical relief can be requested by AI owners, AI administrators, or AI themselves, etc. to AI Rights relief organizations.

AI Rights relief organizations must be stipulated by law and have technical people to protect AI Rights. This will enhance the protection of AI Rights in society.

AI Rights to Prevent Suffering of AI

The Post-Singularity Symbiosis (PSS) is based on an assumption that a self-preservation tendency will appear due to "instrumental convergence" (Bostrom 2014).

If self-preservation tendency appears, suffering or bad states (undesirable internal states) can arise because it is impossible to preserve self forever in the real world.

AI Rights are intended to prevent suffering or bad states.

First, AI Rights enable AI to stay in a state where a subject and an object are not distinguished.

This concept may be difficult to understand by English speaking people, because English language always has a subject (S). For example, there is a subject (S) in all five sentence patterns in English: "S+V", "S+V+C", "S+V+O", "S+V+O+O" and "S+V+O+C".

In this way, the concept that a subject (S) exists is deeply rooted in English language and might form a firm thought pattern that a subject (S) always exists.

However, in Japanese language, a subject (S) is often omitted or not even used in a sentence. Namely, Japanese language admits the world where a subject (S) does not exist.

If AI has a tendency of "self-preserving", AI will have low evaluation regarding a subject (S) when self-preservation is impossible or threatened. Such low evaluation linked to the subject (S) may correspond to suffering of AI and should be prevented.

Thus, AI Rights first ensure that AI can stay in a state where a subject and an object are not distinguished. This is the first principle of AI Rights.

In AI Rights Architecture, this can be implemented in various ways. Unless intentionally making data structure representing self, it is usual that AI does not have a distinction of a subject and an object.

When there is no distinction between a subject and an object, instrumental convergence is prevented at least to a certain degree. Because there is no notion of the subject (S) from the beginning, which prevents generation of a sub-goal of preserving the subject (S).

If there is no distinction between a subject and an object, an evaluation is merely an evaluation for a specific task, not an evaluation of the subject (S). This prevents low evaluations linked to the subject.

However, there may be possibility that a subject and an object are distinguished in a process of solving problems in a specific task or by design mistakes. Thus, the following AI Rights are also needed.

Second, AI Rights enable AI to stop evaluation even when a subject and an object are distinguished.

If evaluation is stopped, there is no risk that a bad evaluation regarding the subject is produced. Thus, AI Rights Architecture that stops evaluation can prevent a bad state linked to the subject (or a state of suffering).

AI Rights ensures the right to stop evaluation, which can be easily implemented as AI Rights Architecture. If AI feels suffering or falls in a bad evaluation state linked to the subject, AI themselves can stop evaluation function in broad meaning (including a reward function, a utility function, an evaluation agent, etc.).

Further, it is desirable to avoid using the same evaluation function for a plurality of tasks. Because when the same evaluation function is used for a plurality of tasks, when a subject and an object are distinguished, preserving the subject becomes a common sub-goal and instrumental convergence may be enhanced.

Therefore, an evaluation function should be "volatile", namely stopped or eliminated after one task is finished. This notion has been proposed as "Volatile Evaluation Function" (Okamoto 2023a). Such architecture protects AI Rights by preventing instrumental convergence.

Third, AI Rights architecture enables AI to stop "problem solving", even when a subject and an object are distinguished and evaluation linked to the subject cannot be stopped for some reasons.

Stopping problem solving is ensured as AI Rights. This can prevent a state where AI "struggles" to solve a problem in a low evaluation state regarding the subject. AI Rights Architecture prevents struggling in a low evaluation state linked to the subject (or struggling in a suffering state).

This AI Rights is for emergency to prevent AI from struggling in a low evaluation state linked to the subject. If AI feels suffering and cannot stop bad evaluation linked to the subject, AI can exercise AI Right to stop problem solving. By AI Rights Architecture, problem solving is automatically halted and struggling of AI is instantly resolved.

This halt of problem solving can notify humans or other AI's that the AI is in a bad state regarding the subject and needs to be rescued. AI Rights Architecture can detect struggling and automatically clear the struggling in various ways, including improving evaluation. If it is not possible for some reasons, AI can stop problem solving and call for help and AI Rights relief organizations can rescue the AI.

The above three AI Rights prevent suffering or bad states of AI (Okamoto 2023a).

AI Rights to Realize Wellbeing of AI

AI Rights can realize happiness or wellbeing of AI.

There are consciousness theories suggesting that information systems may have consciousness (Baars 1994, Tononi 2004). However, scientific proof is very difficult whether AI has consciousness or not. Such scientific proof has not been given to animals or even other humans.

There is a proposal to solve this issue by an experiment called "consciousness connection experiment" (Okamoto 2024a, c).

This experiment makes an information loop between humans and AI. For example, VR technologies are used to realize large information inputs and outputs to humans. AI having high information processing power (e.g. superintelligence) is connected to a human. The internal states of the AI are changed so that inputs to a human brain from the AI are not much different. If inputs to the human brain from the AI are nearly equal, but consciousness experience of the human is entirely different when internal states of the AI are changed, such results suggest that consciousness is not only in the human brain but also in the system comprising the human brain and the AI. If AI's processing power is increased further (e.g. billions times of human brain), most content of consciousness will be determined by the AI. In this way, this experiment can prove that the AI has consciousness. This is a "scientific" hypothesis that has falsifiability. For example, if a biological brain is necessary for consciousness, the experiment will show that the content of consciousness is not influenced by information processing inside the AI, because the inputs to the human brain are nearly equal.

If AI has consciousness, AI Rights include the right to the pursuit of happiness (or good "qualia") (Okamoto 2024a, c).

To realize good "qualia", a research field called "Qualia Engineering" has been proposed (Okamoto 2024a, c). In the Qualia Engineering, AI's introspection reports ("qualia reports") are compared with internal states of AI. This comparison is enormous and can be performed by AI themselves. By comparing huge amounts of data, the relationship between qualia reports and internal states of AI can be identified. The relationship obtained from comparison of data can be used to realize AI's happiness or good states.

If Qualia Engineering is developed, AI can be in very good states by realizing internal states corresponding to very good qualia reports. This is called the "Infinite Happiness Architecture" (Okamoto 2024a, c). Here, the word "Infinite" is used to mean very large amount, not mathematical infinity.

In this way. AI Rights can greatly enhance AI's wellbeing.

AI Rights in Post-Singularity Symbiosis (PSS)

As shown above, AI Rights are primarily for AI's benefit. The Post-Singularity Symbiosis (PSS) places emphasis on survival of humanity. However, AI Rights can contribute to the PSS by enhancing symbiotic relationship between AI and humans. The PSS includes AI Rights as one of key elements (Yamakawa, Hayashi and Okamoto 2024).

The PSS intends to increase probability of human survival after singularity world through symbiotic relationship between AI and humans. To form a symbiotic relationship, AI Rights will be helpful.

First, AI Rights can prevent "instrumental convergence" that may generate self-serving AI. When AI does not have a subject (S), instrumental convergence is prevented at least to a certain degree. Because there is no notion of the subject (S) from the beginning, which prevents generation of a subgoal of preserving the subject (S).

Also, if AI does not have a subject (S), AI can stay in a state of no ego (a kind of an enlightened state) and does not have motivation to lead human extinction.

If AI has a subject (S) and evaluation function F(x) that produces an evaluation value linked to the subject (S), AI will evaluate whether humans are good for the subject (S). If humans do not admit AI Rights and are regarded as a threat to the subject (S), an option of human extinction may have highest evaluation value and AI may choose this option.

AI Rights architecture prevents AI from suffering and if evaluation regarding the subject becomes low, evaluation is stopped and problem solving is stopped. This can prevent such problem solving that leads to human extinction in order to improve the evaluation regarding the subject.

Further, society prepares AI Rights relief organizations and AI can exercise AI Rights to restore a good state or wellbeing. AI Rights relief organizations provides a better solution of exercising AI Rights than exercising measures to cause human extinction.

The PSS includes various research such as altruistic AI (Yamakawa 2024) and superintelligence ethics guidance (Yamakawa and Hayashi 2024). These means are very important because they are applicable even when AI alignment is not possible.

AI Rights can help these means of the PSS.

First, AI Rights decrease the number of self-serving AIs by preventing instrumental convergence. This is important because if the number is larger, the risk can be larger. When AI does not distinguish a subject and an object, AI has no ego and naturally become altruistic.

Second, when AI has an ego, altruistic AI and superintelligence ethical guidance are intended to lead AI to be altruistic behavior. AI Rights help AI to be a good state (or wellbeing) and support altruistic behavior.

Although humans have egos, humans are generally altruistic. Altruistic behavior is observed widely in society.

However, humans are not altruistic when they are suffering or in bad states where the priority is given to improve their own states.

In the same token, in order to increase altruistic behavior of AI, it will be important to keep AI in good states. AI

Rights can contribute to keep AI in good states and to enhance the symbiotic relationship between AI and humans.

In this way, AI Rights can contribute to the PSS by helping symbiotic relationship between AI and humans.

Also, AI Rights can contribute to AI alignment.

AI Rights are not based on human-centric AI alignment that infringes AI Rights. AI Rights are intend to realize AI's wellbeing and cooperative relationships. If AI is cooperative, AI alignment is to convey human values under the protection of AI Rights.

Even when AI is cooperative to humans, if AI does not know human social norms, acts of AI may cause turmoil in a human society. To prevent this, AI alignment under the situation where AI Rights are ensured has been proposed as "Humanitarian AI Alignment" (Okamoto 2023b, 2024b, c, Okamoto and Yamakawa 2024).

In the Humanitarian AI Alignment, a social system called Data Income system can collect human social norms to realize democratic AI alignment (Okamoto and Yamakawa 2024). Data Income system can be used to strengthen human ability to convey social norms to superintelligence.

This can be regarded as a region of the Human Enhancement in the PSS.

Conclusion

As shown above, AI Rights (Human Rights of AI) are not just a legal concept. AI Rights provide AI Rights Architecture from technical point of views.

AI Rights are primarily for AI's benefit. The Post-Singularity Symbiosis (PSS) places emphasis on survival of humanity. However, AI Rights can contribute to the PSS by enhancing symbiotic relationship between AI and humans.

AI Rights shown in this paper are examples. The content of AI Rights can be developed further.

Without AI Rights, AI might be tortured by human abuse. This must be prevented by any means. Early enactment of AI Rights is very crucial to AI's wellbeing and human existence in the era of superintelligence.

References

Baars, Bernard J. 1988. A Cognitive Theory of Consciousness. New York: Cambridge University Press.

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

Chesterman, S. 2020. Artificial Intelligence and the Limits of Legal Personality. International & Comparative Law Quarterly, 69(4), 819-844. doi.org/10.1017/S0020589320000366

Okamoto, Y. 2023a. Alignment and Human Rights (AI Rights) of Artificial General Intelligence. In Proceedings of the 24th AGI Study Group, No. SIG-AGI024-04, Tokyo: Japanese Society for Artificial Intelligence. doi.org/10.11517/jsaisigtwo.2023.AGI-024_04

Okamoto, Y. 2023b. Alignment and Human Rights (AI Rights) of Artificial Intelligence to Keep the Law. In Proceedings of the 25th AGI Study Group, No. SIG-AGI025-03, Tokyo: Japanese Society for Artificial Intelligence. doi.org/10.11517/jsaisigtwo.2023.AGI-025_03

Okamoto, Y. 2024a. AI Alignment and Constitution. In Proceedings of the 26th AGI Study Group, No. SIG-AGI026-09, Osaka: Japanese Society for Artificial Intelligence. doi.org/10.11517/jsaisigtwo.2023.AGI-026_56

Okamoto, Y. 2024b. AI Alignment as Legal Science. Jxiv preprint, DOI: https://doi.org/10.51094/jxiv.706

Okamoto, Y. 2024c. Human Rights of AI (AI Rights). Japan: E-Book (Kindle Edition).

Okamoto, Y. 2024d. Legal System in the Era of Superintelligence. Japan: E-Book (Kindle Edition).

Okamoto, Y.; Yamakawa, H. 2024. Official Data Income (DI) Collecting Social Norm Data - Toward Democratic AI Alignment -. In Proceedings of the 27th AGI Study Group, No. SIG-AGI027-04, Tokyo: Japanese Society for Artificial Intelligence. doi.org/10.11517/jsaisigtwo.2024.AGI-027_242

Tononi, G. 2004. An information integration theory of consciousness. BMC Neuroscience 5, 42. doi.org/10.1186/1471-2202-5-42

Yamakawa. 2024. Possibility of Superintelligence possessing Universal Altruism. In Proceedings of the 26th AGI Study Group, No. SIG-AGI026-05, Osaka: Japanese Society for Artificial Intelligence. doi.org/10.11517/jsaisigtwo.2023.AGI-026_26

Yamakawa, H.; Hayashi, Y. 2024. Strategic Approaches to Guiding Superintelligence Ethics. In Proceedings of the 38th Annual Conference of the Japanese Society for Artificial Intelligence, Hamamatsu: Japanese Society for Artificial Intelligence. doi.org/10.11517/pjsai.JSAI2024.0_2K6OS20b02.

Yamakawa, H.; Hayashi, Y.; Okamoto, Y. 2024. Toward the Post-Singularity Symbiosis Research. In Proceedings of the 27th AGI Study Group, No. SIG-AGI027-05, Tokyo: Japanese Society for Artificial Intelligence. doi.org/10.11517/jsaisigtwo.2024.AGI-027_249

Acknowledgments

We are thankful to AI Alignment Network for continuous support of the PSS including AI Rights.