big data and cognitive computing

MDPI

Article

Peacekeeping Conditions for an Artificial Intelligence Society

Hiroshi Yamakawa [1,2,3]

[1] The Whole Brain Architecture Initiative, a Specified Non-Profit Organization, Nishikoiwa 2-19-21, Edogawa-ku, Tokyo 133-0057, Japan; ymkw@wba-initiative.org
[2] The RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-Chome Mitsui Building, 15th Floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
[3] Dwango Co., Ltd., KABUKIZA TOWER, 4-12-15 Ginza, Chuo-ku, Tokyo 104-0061, Japan

Received: 18 May 2019; Accepted: 20 June 2019; Published: 22 June 2019

Abstract: In a human society with emergent technology, the destructive actions of some pose a danger to the survival of all of humankind, increasing the need to maintain peace by overcoming universal conflicts. However, human society has not yet achieved complete global peacekeeping. Fortunately, a new possibility for peacekeeping among human societies using the appropriate interventions of an advanced system will be available in the near future. To achieve this goal, an artificial intelligence (AI) system must operate continuously and stably (condition 1) and have an intervention method for maintaining peace among human societies based on a common value (condition 2). However, as a premise, it is necessary to have a minimum common value upon which all of human society can agree (condition 3). In this study, an AI system to achieve condition 1 was investigated. This system was designed as a group of distributed intelligent agents (IAs) to ensure robust and rapid operation. Even if common goals are shared among all IAs, each autonomous IA acts on each local value to adapt quickly to each environment that it faces. Thus, conflicts between IAs are inevitable, and this situation sometimes interferes with the achievement of commonly shared goals. Even so, they can maintain peace within their own societies if all the dispersed IAs think that all other IAs aim for socially acceptable goals. However, communication channel problems, comprehension problems, and computational complexity problems are barriers to realization. This problem can be overcome by introducing an appropriate goal-management system in the case of computer-based IAs. Then, an IA society could achieve its goals peacefully, efficiently, and consistently. Therefore, condition 1 will be achievable. In contrast, humans are restricted by their biological nature and tend to interact with others similar to themselves, so the eradication of conflicts is more difficult.

Keywords: autonomous distributed system; conflict; existential risk; distributed goals management; terraforming; technological singularity

1. Introduction

Emergent technology is continually advancing because of its many benefits for humankind. However, technology is not always used for good. As a result, the number of people who have destructive offensive capabilities are increasing. These trends enhance existential risks such as deliberate misuse of nanotechnology, nuclear holocaust, and badly programmed superintelligence [1]. Specifically, the existence of a small number of persons whose aim is to use AI for destructive purposes has the potential to have an enormous impact on humanity. Suspicion between nations has the potential to cause a disastrous war [2]. This irreversible change is also called the "threat of universal unilateralism" [3], and this sufficiently high existential risk could explain Fermi's paradox:

"humanity has no experience of contact with civilized extraterrestrials, compared to their potentially high likelihood of existence" [4].

In many cases, such an abuse of advanced technology is motivated by conflicts in societies, but eliminating all conflict is impossible. Today, the accelerating innovation by, the artificial intelligence (AI) and the recruiting human resource for that are the kinds of major factors of competition when nations and organizations seek to gain supremacy [5,6]. From this background, a human society equipped with advanced technology cannot sustain itself without keeping the peace despite various conflicts. Humankind has made many efforts to maintain peace, including the creation of institutions and organizations such as the United Nations and International Law and Peace Keeping Operation, and their effects have been observed. However, they have been unable to eradicate disputes, wars, and conflicts. Thus, maintaining peace among human societies using only human efforts remains a challenge.

AI will gradually surpass human intelligence, and human-level artificial general intelligence is estimated to be created by 2100 [7]. In general, this unpredictable change is feared due to various dangers [8–12], but this change will provide us with an opportunity to eradicate disputes, conflicts, terrorism, and wars. Peace in human society can be achieved through appropriate interventions by advanced artificial intelligence, rather than by human effort. Figure 1 shows an example of an ecosystem in which an AI system built as a society of intelligent agents (IAs) supports human society. In this example, basically, the AI society observes the values of individuals and/or groups and provides them with benefits. Simultaneously, based on the common values of all of humanity, IAs arbitrate conflicts and contradictions that exist in human society. AIs also act to persuade and educate individuals and groups.

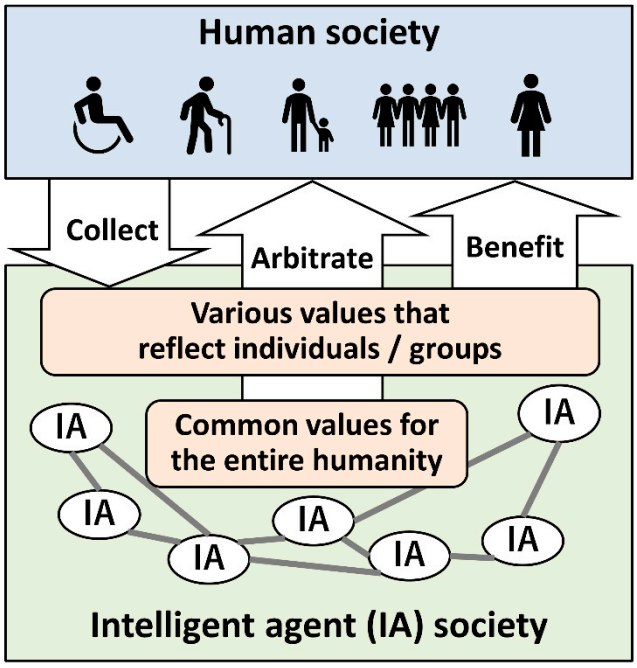

Figure 1. Example ecosystem consisting of human society and artificial intelligence (AI) society. Note: IA denotes intelligent agent which contains AI.

For the AI system to keep the peace in human society, at least the following three conditions must be satisfied:

(1) Condition 1: the AI system is operating continuously and is stable.
(2) Condition 2: the AI system has an intervention method that maintains peace in human societies based on a common value or values.

Additionally, as a premise, the following conditions are required. These conditions involve the sustainable development goals (SDGs), which are a collection of the wisdom of many people, and are potential candidates for common values of humankind in the near future.

(3) Condition 3: a minimum common value must exist that can be agreed upon across human society.

In this study, a thought experiment was conducted to investigate the first condition. The research question is "Is it possible to build an AI system that operates continuously and stably?". The reason this argument is necessary is because an AI system that inevitably creates a society of autonomous decentralized agents can be destabilized by the occurrence of competition, as well as by human society. For an AI system that is becoming increasingly fast, the range of operations that can be controlled by humans decreases. Therefore, the AI system must be able to operate stably without human assistance.

In the next chapter, I explain the setting for a thought experiment in which an AI system is responsible for the execution of terraforming. This mission can be currently readily agreed upon as a common goal by all human beings. AI systems need to be able to react quickly to various situations and must be robust against threats of destruction and failure. For these reasons, the AI system should be a team of distributed autonomous intelligent agents (IAs). In Section 3, the occurrence of contradictions, competition, and conflict in a society of autonomous decentralized IAs is investigated. Even in this case, peace can be maintained if all IAs think that the other IAs share similar goals; however, several obstacles exist to realizing this ideal situation. In Section 4, I argue that a distributed goal-management system for the IA society can be constructed to support sharing goals among agents. By introducing this system, conflicts in IA society can be arbitrated and peace can be realized. In Section 5, I discuss the reasons why it is difficult for human society to maintain peace by itself in comparison with an AI society. In Section 6, the first argument is that IAs can be comrades for human beings, unlike other animals. Further, I argue that unlike human society, majority decision-making does not make sense for an IA society. The major conclusions are finally summarized in Section 7.

2. Thought Experiment Settings

Before explaining the terraforming that is the subject of the thought experiment in the following sections, a trivial example in which resource competition creates peaceful cooperation rather than conflict is explained. For example, in the case where a deep reinforcement learning agent [13] runs searches in parallel for a parameter that achieves the highest score in a certain game task, the agent plans how and in what order to conduct a number of experiments. During the process, no parameter fights another over finite computational resources, and thus there can be no problematic situation that requires resolution. Due to the clarity of common goals, and due to an absence of local goals in each parameter set, there is no competition among parameter sets (although such virtual competition paradigms can be possible). This means that there can be no struggle under shared common goals and unified management.

This section is divided into subheadings. It provides a concise and precise description of the experimental results, their interpretation, and the experimental conclusions that can be drawn.

2.1. Intelligent Agent Society for Terraforming

In the above example in which no agent has local goals, no competition can occur. As the setting of the thought experiment in this article, it is assumed that a number of IAs that can proliferate themselves have been sent to an unknown planet. The background of this setting is that the IAs are dispatched from Earth by humans, and charged with the mission of remaking the environment of the planet in preparation for human migration. Thus, the goal of these IAs is transforming the planet into being human-habitable; that is, terraforming. The Invincible (first published in 1945) by Stanisław Lem [14] and Code of the Lifemaker by James P. Hogan [15] are famous fictional examples of this kind of scenario.

This science fiction-like setting is introduced for two reasons. The first is to simplify the structure of the struggle in human society. The second is to reduce the influence of biased thinking caused by personal history. The procedure of providing an objective function to an agent is standard in artificial intelligence research, and can be easily discussed. In addition, I think that expanding human habitats to other planets is an effective method to increase the survivability of humanity.

In the following sections, the scenario that the IA society avoids disputes and war and maintains peace, despite technological progress causing conflicts among that society, is described.

2.2. Autonomous Distributed Agents: For the Survival of the Group

Fortunately, a group of IAs have landed on a planet. They have secured resources including energy and have successfully survived for the time being. To achieve their common goal, which is the terraforming of this new planet, they begin striving and working toward it.

For the following reasons, each member of the group must be spatially distributed and autonomous [16] because, firstly, each IA should adapt to the surrounding environment to respond quickly to what is in front of its eyes with limited information processing capabilities. The second reason is concerned with the robust survivability of the group. The hardware of individual IAs is constantly exposed to various environmental factors and may be destroyed. To secure the survival of the group, therefore, IAs have to be highly autonomous and spatially dispersed.

2.3. Physical Composition of the IAs and Their Group

It is assumed that the hardware of each IA is a set of physical devices for memory, communication, computation, sensory inputs, locomotion, manipulation, and so forth. It is further assumed that hardware for new members is produced/reproduced in manufacturing plants, in a system that is similar to that of social insects. Unlike living organisms, though, reproduced IAs do not need to be similar to the producing IAs, as they are manufactured solely based on their design specifications. Stored data such as memory, programs, and knowledge, which arguably compose the essential substance of the IAs, are realized as software, and the dependency on their hardware can be relatively low.

Thus, the essential substance of IAs, which is their software, does not need a specific physical body and is able to wander among many bodies. Even the preservation and maintenance of the software of a specific IA are low priority because electronic data can be stored easily and restored at will. Additionally, when an IA reboots another IA, they do not need to be similar.

3. Development and Conflict in an IA Society

IAs work as an organization by communicating information, such as goals, to each other. Formation of a group leads to cooperation and division of labor within it, which contributes to efficiency in achieving their common purpose. By sharing knowledge about their environment and developing knowledge including science and technology, their efficiency continually. Closer relationships between the IAs enable useful collaborations toward their goals, but also increase conflicts.

3.1. Diversification and Fixation of Local Values: Emergence of Survival Instinct

All IAs contain a distributed autonomous system with common goals, and the members are required to retain the goals and maintain activities toward achieving them. This means all the IAs must hold the common goals individually and in a distributed manner. Each IA derives sub-goals from the common goals or from an assigned part of the common goals (target-means decomposition) in response to its environment, and builds local values as a network of sub-goals. Each IA forms specialized local values, depending on its body, tasks, and the local environment (Figure 2). This area of research is referred to as cooperated multi-agent planning (MAP), and a large amount of accumulated work on the subject has been published [17]. As each IA changes behavior by learning, and as the number of IAs increases, the coordination between them becomes more difficult.

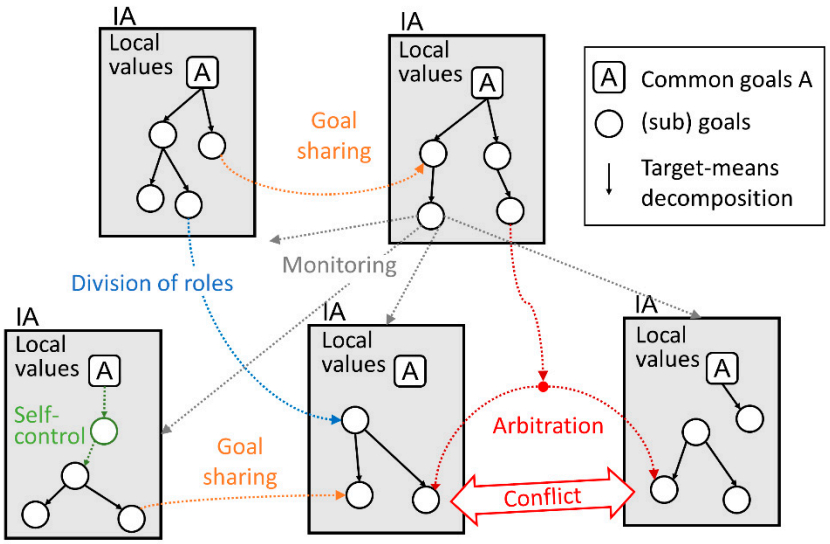

Figure 2. Sharing goals in an IA group: various IAs generate different sub-goals and specialize themselves to those sub-goals in response to their environment, bodies, and tasks. Each IA conducts target-means decomposition, arbitration, and, at times, checks consistency. They also carry out, in cooperation with other IAs, task allotment, goals sharing, arbitration, monitoring, and so forth.

Individual IAs carry out activities toward concrete sub-goals within a certain time frame. Too frequent changing of the sub-goals makes it difficult for them to solve the current problems. Thus, other sub-goals emerge—stabilization of local values. With a slightly longer time frame, self-preservation of their hardware also emerges as a sub-goal because frequent breakdowns, or shutdowns, of the hardware impairs their usefulness.

Maintenance of the IAs' software does not emerge as an overly important sub-goal because, basically, the programs of these IAs can be stored, rebooted, copied, and transferred at low cost. There is the risk that they may be destroyed by accidents, attacked, or manipulated by enemies before rebooting. Additionally, the risk exists that a proper information environment might not be available when rebooting is needed. However, if some specific program is useful from the viewpoint of the common goals, the IA society tries to secure preservation and rebooting of that program.

In case the rebooted IA needs to catch up to a change in the social situation, the AI society simply provides it with a learning period. If a rebooted IA is forgotten by the others and cannot serve IA society, that one does not need to be rebooted, and if wrongly rebooted, it is immediately shut down. Almost all the goals of IAs have terminating conditions (ending with the fulfillment of the purpose or with a judgment of infeasibility) [18]. However, the survival instinct, including self-preservation, resource acquisition, and knowledge acquisition, is always a sub-goal as long as each IA exists. Regardless of initial goals, any advanced intelligence generates sub-goals related to a survival instinct, and it sometimes becomes excessively predominant; this is called "instrumental convergence" [10].

3.2. Confident Sharing of Common Goals Is Difficult

When all IAs share mutually believable local values derived from common goals, no inconsistency or struggle will exist between IAs, and all IAs in the society can pursue common goals peacefully, efficiently, and consistently. The ideal situation is, in other words, that every agent can believe that "all other agents intend socially acceptable goals".

However, as mentioned above, when various goals are generated diversely and dynamically in each IA, different local values will be developed among them. Therefore, it is required for the cooperation and division of labor between different IAs to not only share goals, but also to be mutually confident about the shared goals. This would correspond to the establishment of a trust relationship in the contract.

It is assumed that each IA is programed to act in good faith; this means IAs do not pretend, lie, or betray.

Because the AI system is designed as distributed autonomous IAs, an IA needs to be able to do the following to act ideally for the social good:

(1) commit to socially accepted goals,
(2) send and receive goals as information to and from other IAs, and
(3) understand goals received from other IAs.

Here, "socially acceptable goals" means that the goals contribute to common goals and do not conflict with any other IA's local values in practice.

I think that a society constructed by individuals with different local values has a potential risk of conflict. Therefore, some common goals must be shared that are on levels beyond those local values to establish a single, orderly society.

There are three obstacles to achieving the above ideal situation:

(1) Communication channel problem

It is assumed that communication paths between IAs for sharing common goals as information are built in and shared with all IAs in the design stage. However, communication channels among IAs are not always stable and may be disrupted at times. According to Brewer's CAP Theorem (This theorem states that it is impossible for a distributed data store to simultaneously provide more than two out of the consistency, availability and partition tolerance) [19], when securing availability and partition tolerance in a distributed system, a delay in sharing information must be accepted.

(2) Comprehension ability problem

This problem is caused by the limitations of each IA's comprehension ability. Here, this ability means the capacity of an IA to derive sub-goals from received goals and to act on received goals. Even if the shared goals are formally identical, differences in IAs lead to different comprehension. For example, different IAs have different designs and appearances (body, experience/knowledge, capacity, etc.).

(3) Computational complexity problem

Suppose one IA overcomes the above two problems and understands the other IAs' goals. Even in this case, the following processing is required to avoid substantial conflicts. First, in the IA's own environment, all the goals held as their own local values will check for conflicts with other IA goals. Next, if a contradiction is detected, the IA needs to change its own local value so that the contradiction does not occur in view of the higher priority goals. In some cases, an IA may need to determine that it needs to request another IA to adjust its goals. This type of processing requires a considerable computational cost.

3.3. Birds of a Feather Flock Together: Agent Society for Terraforming

Given the problems mentioned above, it is difficult for IAs to share goals in a workable manner. However, if the pre-designed appearance is similar between IAs, they can infer that they have similar goals because the IAs' goals and appearances are probably governed by the same design information. If the circumstances are similar between IAs, their interests also tend to be similar. In short, when similar people gather together, the possibility of sharing goals is increased (Figure 3).

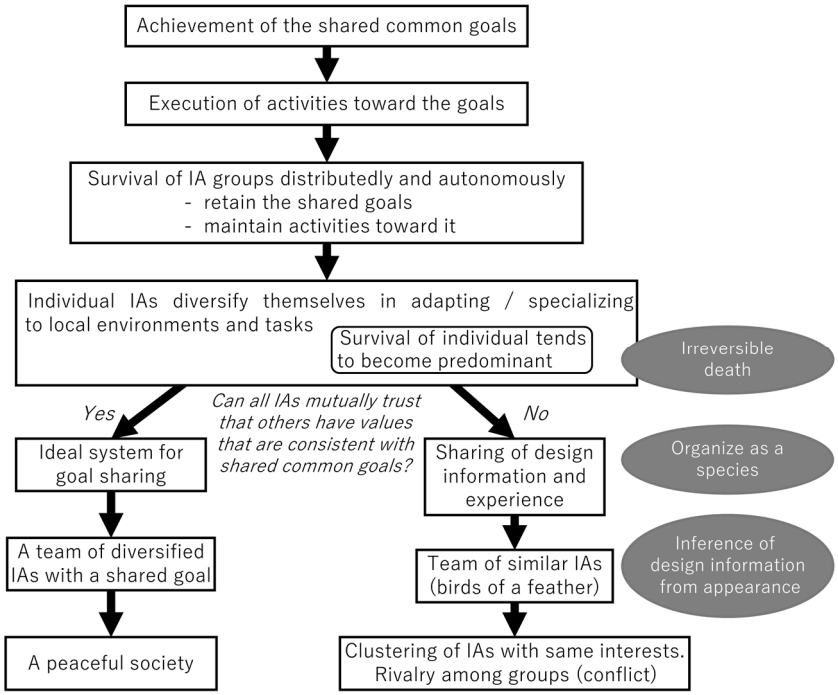

Figure 3. The branching point of peace and conflict.

A highly homogenous team of IAs will have few conflicts. They can cooperate and divide labor efficiently, and that makes them advantageous compared with other teams. Like a flock of birds, survival probability is increased for individual members. This leads to achieving another sub-goal that they should pursue as a team: survival of the team. For an individual in a flock, the more its local values reduce conflict with other members, the better the chance it has of surviving, which promotes standardization across the entire flock.

For these reasons, IAs will tend to form highly homogeneous teams. In other words, "Birds of a feather flock together". However, this often exposes a weakness: homogeneity makes the flock susceptible to environmental changes.

3.4. Conflict Between Groups

Within each group, similar local values are shared by the members, but they differ from those of other groups. Because resources in the world are limited, the effort to acquire them causes conflicts of interest between individuals and between groups. Similarly, divisions in IA groups will tend to cause a state of conflict between groups.

When confrontation deepens, an IA in one group perhaps ignores, disfavors, or blocks opponent IAs. Contrarily, the same IA provides preferential treatment and increased communication to members of the same group.

As already described in the introduction, if the worst should happen, a struggle could imperil the entire society. However, even in the preliminary stages of such a conflict, each IA will expand activities of attacking and defending against opposing IA groups, causing the problem of diminishing allocation of resources to the original common goals.

4. Peace of IA Society Maintained by a Distributed Goal Management System (DGMS)

A distributed goal management system (DGMS) should be introduced to make the IA society peaceful. The technological foundation of DGMS has progressed in the field of multi-agent planning (MAP) since the beginning of 90's, as reported in a previous survey article [17]. By using DGMS, each IA coordinates its local values with other IAs' values through dialogue, and often an individual IA

needs to execute the tasks it faces in real time. Various technical issues must be overcome to realize the specifications required for practical DGMS, and it is necessary to promote research on MAP and related fields to make DGMS a reality.

If the goals of all IAs are coordinated to be socially acceptable, there is no conflict in IA society. However, communication channel problems, comprehension ability problems, and computational complexity problems are preventing this from being realized. To overcome these problems, DGMS should have the additional three functions listed below.

4.1. Normal Responses

When two IAs conflict with each other due to different sub-goals and actions derived therefrom, another appropriate IA arbitrates (or mediates) between them. A third IA estimates the importance and validity of the sub-goals of both by considering consistency and contribution to the common goals. Both IAs must comply with the ruling (Figure 2).

In cases in which the local value of an IA lacks consistency with the common goals and consistency cannot be restored by the calculation of the IA itself, another appropriate IA recalculates the sub-goals and assigns them.

4.2. Emergency Responses

Emergency responses are necessary because normal responses need time for communication and calculation. These will consist of suspending actions that are based on questionable sub-goals and even shutting down the IA temporarily for safety.

The method for detecting danger in each IA as the premise for taking these kinds of measures is as follows. First, one IA monitors the local values of many IAs and finds any sub-goals that are inconsistent with the common goals and might be a source of conflict. Second, each IA checks the consistency between its local values and common goals. This is the self-restraint of IAs (Figure 3).

This crisis management is a kind of traditional safety design technology (e.g., safe operation of aircraft).

4.3. Task Assignment in Consideration of Comprehension Ability

Due to the comprehension ability problem, the level of understanding of IAs varies depending on differences in their design and/or appearance. However, understanding goals is necessary for many goal-related processes such as execution, target-means decomposition, arbitration, and monitoring. Thus, appropriate assignment of roles to each IA by considering its comprehension ability will be an important technical consideration in designing DGMS.

If ideal and practical DGMS can be built by overcoming obstacles, the situation in which every IA thinks that "all other IAs intend socially acceptable goals" can be maintained. In that situation, IA society can achieve goals peacefully, efficiently, and consistently.

5. What Prevents Peace in Human Society?

Despite our capability to share common goals through language, humans cannot stop fighting with each other. Much of the reason for this might lie in the biological constraints to which humans are subjected. By comparing IA society to human society, the reasons that human society divides into many rival groups resulting in conflicts were considered.

5.1. Irreversible Death of Living Organisms

Death is an irreversible and inescapable event for all organisms. Humans cannot delay or switch on and off our biological activities at will. For this reason, each individual organism has to cling to life. In the future, the realization of hibernation technology may reduce this fear, but our human brain

cannot escape from the fear that something could go wrong, and we might not wake up. People in hibernation might also fear that they will be forgotten by society (e.g., Rip Van Winkle [20]).

Because available resources in the world are limited, increasing numbers of individuals obsessed with survival will inevitably cause competition for resource acquisition and become an origin of conflict.

An important part of IAs is their software, which can be stored, rebooted, copied, and transferred at a low cost as described in Section 3.1. Therefore, the degree to which IAs cling to their lives may be much lower than that of living organisms.

5.2. Struggle Between Evolutionarily United Species

Evolution is a search algorithm, and it can expand the possibilities of organisms through copying, mating, and mutating individuals. The phenotypes that can survive in the environment are extremely narrow in the space of innumerable genome combinations [21]. Therefore, to ensure the survival probability of offspring born by mating, it is necessary to form a species that is a homogeneous group.

Surviving as a species requires a population above a certain number, which is called the minimum viable population (MVP) [22]; otherwise, the diversity of the genes within the species decreases and it becomes vulnerable to extinction. For this reason, individuals of a species share the same fate, and they sometimes help other members of the same species [23]. However, this leads to competition over resources among species that sometimes develops into conflict, as seen in invasive species [24].

IAs are constructed based on their specifications, and they can produce completely different offspring. Therefore, there is no incentive to increase similar mates for reproduction. The situation in which species compete for resources becomes unrealistic. In the future, gene editing technology could realize the free design of living organisms [25]. After that, humans may not need to fight to maintain the species.

5.3. Estimate Goal Similarity Based on Appearance

In the case of the IA, there is no need to be suspicious that it uses the same communication devices and shares common goals, except with regard to their failure or hacking. Living organisms sometimes validate agreement of design information using chemical interaction. However, for humans, other agents' goals are inferred from their appearance or from shared experiences. Therefore, human beings tend to be sympathetic to organisms or objects that have a similar appearance. Due to this nature of human beings, it is thought that intelligent robots like human beings will greatly affect human emotions, and there are concerns regarding various problems arising from this [26].

Human beings can communicate their goals to others through language, but they cannot know whether the other intends to commit to these goals. To secure that point, modern society uses a legally effective contract. In this case, it is premised that the people on both sides have the ability to understand the contents of the contract. However, from the viewpoint of one organization, it is not possible to confirm the intention of another organization. Stemming from this lack of confidence, almost every nation has military capabilities that are based on "offensive realism" [2].

5.4. Section Summary

As humans are also living beings, they have an individual survival instinct for avoiding irreversible death, and they try to care for members of the same species to keep the species alive. In particular, humans tend to infer the goals or intentions of other persons from their visual appearance. If the same language is used, cooperation in the group is much easier. For this reason, even in humans, the tendency of "birds of a feather flock together" is particularly strong. Therefore, competition between groups with different local values inevitably occurs.

6. Discussion

6.1. As a Comrade

The mechanisms of collaboration of non-human animals are mostly determined genetically [23]. However, from the time of evolution to homo sapiens, the scope of our recognition of a comrade began to expand, and is now expanding to all humanity through language and education [27]. The scope recognized as comrades will be extended to include intelligent machines in the future.

According to the Salient Value Similarity (SVS) model [28], whether a person is trustworthy or not depends on whether the person seems to share the same stable and consistent goals, and has the capability and enthusiasm to pursue them, from the perspective of the observer. Again, the main condition for peaceful coexistence is whether every agent can believe that "all other agents intend socially acceptable goals". In this sense, IAs with highly advanced AI will be able to share goals with humans, and have the capacity and enthusiasm to pursue them as well. Considering this, IAs will have the opportunity to become trustable comrades, more so than other intelligent animals on earth. If such an advanced AI can share a relatively wide range of values with them through the education given by surrounding people, like foster children, then it is possible that AI may become a trusted comrade.

6.2. Significance of Majority Decision

A decision by majority rule is a prevalent social decision method in human society. However, in the case of IAs whose programs are indefinitely duplicable, it is pointless to count the number of software units that agree with an opinion. Conversely, imbuing hardware with the right to vote can be somewhat meaningful, but perhaps the hardware has no opinion.

In contrast, the individuality of each living organism is of supreme importance because each software and hardware is tightly coupled. Thus, in social decisions made in human groups, , appropriate to distribute voting rights with the same weight to each person from a utilitarian perspective [29].

As for social decisions in IA society, the main point of interest is how to achieve common goals. Therefore, it is desirable that the IA group collect diverse experiences and abilities to produce diverse planning alternatives [30]. It is also desirable for the IA groups to be able to predict the degree of contribution of each alternative to the common goals as accurately as possible and, eventually, decide the best action for attaining these goals.

7. Conclusions

The development of emergent technology will increase the risk to human existence. Therefore, maintaining peace in human society by overcoming various conflicts is becoming an urgent issue. Historically, human society has not achieved full peace through its own efforts alone. Therefore, it is worthwhile to explore the possibility of realizing peace in human society through the intervention of advanced AI systems. Three conditions are assumed to be needed to realize this situation: There are minimum common values that can be agreed across society (condition 3), advanced AI systems can intervene to keep the peace of human society based on these common values (condition 2), and an AI system exists that can work stably and continuously (condition 1).

In this paper, an AI system that satisfies condition 1 was investigated. A part of the system may potentially be destroyed, so a robust IA society should be a team of autonomous and distributed IAs. A common value of humanity is shared among all IAs. Individual IAs would decompose common goals and derive means so that they can contribute to the advancement of common values. Each IA would diversify its activities to effectively divide tasks among them all. In order to adapt to their local environment, IAs would usually hold, as their local values, sub-goals derived from common goals.

There are a wide variety of local values and competition for available resources will create a competitive situation for IA comrades. It is an advantage that competition leads to an increase in capacity to achieve a common goal. However, if the effort towards merely winning the competition increases, cooperation is lost, and devastating struggles occur, creating obstacles to achieving common goals.

The ideal situation is one in which every agent believes that "all other agents intend socially acceptable goals". Here, "socially acceptable goals" means that the goals contribute to common goals and do not conflict with any other IA's local values in practice. Under such circumstances, the IA society can achieve goals peacefully, efficiently, and consistently. Communication channel problems, comprehension ability problems, and computational complexity problems exist, however, that may impede the realization of ideal situations. In an IA society based on a computer, it seems possible to design a DGMS that maintains the local values of distributed IAs.

Conversely, humans are biologically constrained. Irreversible death strengthens our survival instincts, and human beings need to maintain our species through reproduction. Humans must also distinguish their mates by appearance. For these reasons, similar people gather and become more likely to form a party. If people divide into groups with similar values and compete for resources, this can be a major cause of conflict.

I assumed that building a universal AI system to arbitrate conflicts in human society based on a common value (Figure 1) would reduce the existential risk. This assumption is consistent with Torres' The Friendly Supersingleton Hypothesis [3]. For stable and continuous operation, the AI system in this paper was constructed as an autonomously distributed system, which has concurrency, scalability, and fault-tolerance. Many issues remain to be solved [17], but this technology is feasible. From this aspect, my method differs from the Friendly Supersingleton of Torres, which is based on future technology. It is desirable for the final form of our proposed AI system to be almost autonomous and worldwide, but part of that system can begin as a conventional AI system with the help of human operators. However, a new issue will then arise regarding executing arbitrations that are consistent with a common value, while avoiding arbitrary influences of human operators.

Finally, various possibilities for applying superintelligence to reduce existential risks caused by various non-AI factors, such as climate change, have been discussed before [31]. With regard to AI itself, discussions have mainly focused on the increase in risks they might pose. An approach using advanced AIs to reduce the existing risks that increase with the progress of AIs has not been sufficiently investigated, either in this paper or by Torres [3]. However, effectively using the power of superintelligence or more elementary AI to construct future governance will create previously unknown possibilities for the future of humanity.

Funding: This research received no external funding.

Acknowledgments: The author thanks Koichi Takahashi, Hiroshi Nakagawa, Hirotaka Osawa, Kentaro Fukuchi, Satoshi Hase, Dohjin Miyamoto, Ichiro Hasegawa, Masahiko Osawa, Ayako Fukawa, and Rzepka Rafal for their helpful comments and feedback.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Bostrom, N. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *J. Evol. Technol.* **2001**, *9*, 1.
2. Tinnirello, M. Offensive Realism and the Insecure Structure of the International System: Artificial Intelligence and Global Hegemony. In *Artificial Intelligence Safety and Security*; Yampolskiy, R.V., Ed.; Taylor & Francis Group: Boca Raton, FL, USA, 2018.
3. Torres, P. Superintelligence and the Future of Governance: On Prioritizing the Control Problem at the End of History. In *Artificial Intelligence Safety and Security*; Yampolskiy, R.V., Ed.; Taylor & Francis Group: Boca Raton, FL, USA, 2018.
4. Sotos, J.G. Biotechnology and the Lifetime of Technical Civilizations. Available online: https://arxiv.org/abs/1709.01149 (accessed on 18 June 2019).
5. Cave, S.; ÓhÉigeartaigh, S.S. An AI Race for Strategic Advantage: Rhetoric and Risks. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 1–3 February 2018.
6. General AI Challenge. Available online: https://www.general-ai-challenge.org/solving-the-ai-race-results/ (accessed on 9 June 2019).

7. Vincent, M.; Bostrom, N. Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In *Fundamental Issues of Artificial Intelligence*; Müller, V., Ed.; Springer: Berlin, Germany, 2014.
8. Omohundro, S.M. The Nature of Self-Improving Artificial Intelligence, Presented at the Singularity Summit. Available online: https://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf (accessed on 9 June 2019).
9. Bostrom, N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds Mach.* **2012**, *22*, 71–85. [CrossRef]
10. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
11. Shanahan, M. *The Technological Singularity*; The MIT Press: Cambridge, MA, USA, 2015.
12. Yampolskiy, R.V. Taxonomy of Pathways to Dangerous Artificial Intelligence. In Proceedings of the Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
13. Francois-Lavet, V.; Henderson, P.; Islam, R.; Bellemare, M.G.; Pineau, P. An Introduction to Deep Reinforcement Learning. Available online: https://arxiv.org/abs/1811.12560 (accessed on 9 June 2019).
14. Lem, S. The Invincible (Polish: Niezwyciężony). Available online: https://en.wikipedia.org/wiki/The_Invincible (accessed on 22 June 2019).
15. Hogan, P.J. *Code of the Lifemaker*; Spectrum Literary Agency: New York, NY, USA, 1983.
16. Ahmed, W.; Wu, Y.W. A survey on reliability in distributed systems. *J. Comput. Syst. Sci.* **2013**, *79*, 1243–1255. [CrossRef]
17. Torreño, A.; Onaindia, E.; Komenda, A.; Štolba, M. Cooperative Multi-Agent Planning: A Survey. *ACM Comput. Surv.* **2017**, *50*, 84. [CrossRef]
18. Rao, A.S.; Georgeff, M.P. Modeling rational agents within a BDI-architecture. In Proceedings of the 2nd International Conference Principles of Knowledge Representation and Reasoning, Cambridge, MA, USA, 22–25 April 1991; pp. 473–484.
19. Lynch, N.; Gilbert, S. Conjecture and the feasibility of consistent, available, partition-tolerant web services. *ACM SIGACT News* **2002**, *33*, 51–59.
20. Irving, W. *Rip Van Winkle*; Creative Co.: Boston, MA, USA, 1994.
21. Chaitin, G. *Proving Darwin: Making Biology Mathematical*; Pantheon Books: New York, NY, USA, 2012.
22. Boyce, M.S. Population Viability Analysis. *Annu. Rev. Ecol. Syst.* **1992**, *23*, 481–506. [CrossRef]
23. Nowak, M.A. Five Rules for the Evolution of Cooperation. *Science* **2006**, *314*, 1560–1563. [CrossRef] [PubMed]
24. Beck, K.G.; Zimmerman, K.; Schardt, J.D.; Stone, J.; Lukens, R.R.; Reichard, S.; Randall, J.; Cangelosi, A.A.; Cooper, D.; Thompson, J.P. Invasive Species Defined in a Policy Context: Recommendations from the Federal Invasive Species Advisory Committee. *Invasive Plant Sci. Manag.* **2008**, *1*, 414–421. [CrossRef]
25. Kobayashi, M. *What Is Genome Editing?—The Impact of 'CRISPR'*; Kodansha: Tokyo, Japan, 2016. (In Japanese)
26. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Affective Computing. In *Ethically Aligned Design*, 1st ed.; From Principles to Practice; IEEE Standards Association: Piscataway, NJ, USA, 2019. Available online: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e_affective_computing.pdf (accessed on 9 June 2019).
27. Harari, Y.N. *Sapiens: A Brief History of Humankind*; Harper: New York, NY, USA, 2015.
28. Earle, T.C.; Cvetkovich, G. *Social Trust: Toward a Cosmopolitan Society*; Praeger: Westport, CT, USA, 1995.
29. Mill, J.S. *Utilitarianism*, 1st ed.; Parker, Son & Bourn, West Strand: London, UK, 2015.
30. Cuppen, E. Diversity and constructive conflict in stakeholder dialogue: Considerations for design and methods. *Policy Sci.* **2012**, *45*, 23–46. [CrossRef]
31. Bostrom, N. Strategic Implications of Openness in AI Development. In *Artificial Intelligence Safety and Security*; Yampolskiy, R.V., Ed.; Taylor & Francis Group: Boca Raton, FL, USA, 2018.