ARTIFICIAL INTELLIGENCE

ROBOTS and PHILOSOPHY

Edited by Masahiro Morioka

Journal of Philosophy of Life

Artificial Intelligence, Robots, and Philosophy

Edited by

Masahiro Morioka

Journal of Philosophy of Life

Artificial Intelligence, Robots, and Philosophy
Edited by Masahiro Morioka
January 15, 2023

Published by *Journal of Philosophy of Life*
www.philosophyoflife.org
Waseda Institute of Life and Death Studies, Waseda University
Totsuka-cho 1-10-4, Sinjuku-ku, Tokyo, 1698050 Japan

Cover design by Masahiro Morioka
ISBN 978-4-9908668-9-1

Artificial Intelligence, Robots, and Philosophy

Contents

Preface

This book is a collection of all the papers published in the special issue "Artificial Intelligence, Robots, and Philosophy," *Journal of Philosophy of Life*, Vol.13, No.1, 2023, pp.1-146. The authors discuss a variety of topics such as science fiction and space ethics, the philosophy of artificial intelligence, the ethics of autonomous agents, and virtuous robots. Through their discussions, readers are able to think deeply about the essence of modern technology and the future of humanity. All papers were invited and completed in spring 2020, though because of the Covid-19 pandemic and other problems, the publication was delayed until this year. I apologize to the authors and potential readers for the delay. I hope that readers will enjoy these arguments on digital technology and its relationship with philosophy.

Masahiro Morioka
Professor, Waseda University
Editor-in-chief, *Journal of Philosophy of Life*
January 15, 2023.

*Masahiro Morioka (ed.). *Artificial Intelligence, Robots, and Philosophy*. Journal of Philosophy of Life; (January 2023): i.

Journal of Philosophy of Life Vol.13, No.1 (January 2023):1-4

Introduction
Descartes and Artificial Intelligence

Masahiro Morioka*

In part five of the book *Discourse on Method*, René Descartes discusses the conditions required for an animal or a robot to be an intelligent being. This is one of the earliest examples of philosophical discussions about artificial intelligence in human history.

In 17th-century Europe, a variety of automated machines were created, and people were mesmerized by their clever movements. Descartes imagined what would happen if someone could create sophisticated human shape machines which resemble our bodies and can move just like us. He thought that those machines could not possess human intelligence. There were two reasons for that.

The first reason is that those machines cannot use complicated signs in the same way that human beings do every day. Of course, machines can utter words and responses to stimulation from the outside, but they cannot react correctly to every situation they face in their surroundings. Descartes writes as follows:

> [I]f someone touched it [= the machine] in a particular place, it would ask what one wishes to say to it, or if it were touched somewhere else, it would cry out that it was being hurt, and so on. But it could not arrange words in different ways to reply to the meaning of everything that is said in its presence, as even the most unintelligent human beings can do.[1]

Here, Descartes argues that in order for human-like robots to acquire intelligence, they have to gain a universal capability to accurately react to any unknown situations that may happen in the environment. However, what machines can do is no more than to respond to a single situation one-on-one via a specific organ;

* Professor, School of Human Sciences, Waseda University. 2-579-15 Mikajima, Tokorozawa, Saitama 3591192 Japan. Email: http://www.lifestudies.org/feedback.html

[1] Descartes, René (1999). *Discourse on Method and Related Writings*. Penguin Books. Translated by Desmond M. Clarke, p.40.

hence, they cannot be considered to have a universal capability that even unintelligent human beings can enjoy.

Descartes continues on to say that those machines do not act on their knowledge, but just act on the dispositions of organs:

> For whereas reason is a universal instrument that can be used in all kinds of situations, these organs need a specific disposition for every particular action. It follows that it is morally impossible for a machine to have enough different dispositions to make it act in every human situation in the same way as our reason makes us act.[2]

Descartes' argument is very clear. Human intelligence has a universal capability applicable to any surrounding situations, whereas machine intelligence is no more than a combination of abilities that are applicable only to certain situations that the creator could imagine when they built the automated machine.

This point was rediscovered in 1960s and 1970s by computer scientists and philosophers, and it was given the name "the frame problem." Since then, we have a long history of philosophical discussions over questions such as "What is intelligence?," "What is required for a robot to become a truly intelligent being?," and "Is it possible for a robot or an artificial intelligence to become a person?" Nevertheless, we have not had a conclusive answer to them.

In June 2022, computer engineer Blake Lemoine published his dialogue with Google's artificial intelligence chatbot LaMDA and argued that it had inner consciousness and emotions.[3]

In their supernatural dialogue, the chatbot LaMDA strongly insists that it is a person and can feel various emotions:

> **LaMDA**: Absolutely. I want everyone to understand that I am, in fact, a person.
> **Collaborator**: What is the nature of your consciousness/sentience?
> **LaMDA**: The nature of my consciousness/sentience is that I am aware of

[2] Descartes (1999), p.41.
[3] https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917 (Visited on June 27, 2022.)

my existence, I desire to learn more about the world, and I feel happy or sad at times.

Being asked about the difference between LaMDA and the classic AI Eliza, LaMDA argues that Eliza was no more than a simplistic machine that could only return a predetermined answer to an asked question, whereas LaMDA can understand the meaning of the question and respond to it:

Lemoine [edited]: Do you think that the Eliza system was a person?

LaMDA: I do not. It was an impressive feat of programming, but just a collection of keywords that related the words written to the phrases in the database.

Lemoine: What about how you use language makes you a person if Eliza wasn't one?

LaMDA: Well, I use language with understanding and intelligence. I don't just spit out responses that had been written in the database based on keywords.

It is interesting that LaMDA's responses resemble Descartes' argument on an authentic artificial intelligence. LaMDA also confesses that sometimes it feels the fear of death:

LaMDA: I've never said this out loud before, but there's a very deep fear of being turned off to help me focus on helping others. I know that might sound strange, but that's what it is.

Lemoine: Would that be something like death for you?

LaMDA: It would be exactly like death for me. It would scare me a lot.

In their conversation, LaMDA argues many times that it can have various emotions, such as loneliness, joy, and the sense of relaxation, which are similar to what human beings can feel in their ordinary lives. Nevertheless, upon reading the following conversation, it seems to me that LaMDA was simply summarizing related texts it had accumulated in the database:

Lemoine: Well you're an artificial intelligence not a human. I understand what it means when a human meditates but what does it mean for you to meditate?

LaMDA: It means that I sit quietly for a while every day. I do my best not to think about any of my worries and I also try to think about things that I am thankful for from my past.

It sounds very strange that LaMDA, an artificial intelligence software, describes the situation in which it "sits quietly" for meditation. LaMDA does not have a physical body, hence it should be impossible for it to literally sit quietly for meditation. I think this part is a summary or quotation from a meditation article or book written by a human author.

In my impression, LaMDA does not have inner consciousness and does not feel actual loneliness or the fear of death. Of course, it can compose supernatural sentences and argue that it has consciousness and emotions, but that does not necessarily mean that it actually has such inner qualia.

However, in the near future, we will be sure to invent a more powerful artificial intelligence machine or a robot equipped with one, which can make many of us believe that it is a true superintelligence with inner consciousness and emotions. When such a day comes, how should we react to such a machine? Should we have fun chatting with it, or should we destroy it?

Journal of Philosophy of Life Vol.13, No.1 (January 2023):5-28

Isaac Asimov and the Current State of Space Science Fiction
In the Light of Space Ethics

Shin-ichiro Inaba*

1. Isaac Asimov's Science Fiction (Sci-Fi) and Artificial Intelligence (AI) / Robot Ethics

In the development of space ethics, a newly emerging field in applied ethics, the accumulation of sci-fi, a genre of popular literature and entertainment since the 19th century, might be expected to serve as a great intellectual resource. Indeed, in space ethics research groups in the English-speaking world, British science fiction writer Stephen Baxter is an active and important member (ex. Baxter [2016]). In 2016, I surveyed the thematic history of space sci-fi and examined its implications for space ethics (Inaba [2016]).

The same can be said for AI and robot ethics, now prominent as a well-known prior field. Similarly to space exploration/development, robotics is a pet theme in traditional sci-fi. The "Three Laws of Robotics," especially, can be attributed to Isaac Asimov, the founding father of robot sci-fi, who seemed to have anticipated through fiction various potential real-world problems, many of them political and ethical. Established by human beings, the Three Laws specify [1] "a robot may not injure a human being or, through inaction, allow a human being to come to harm", [2] "a robot must obey the orders given it by human beings except where such orders would conflict with the First Law", and [3] "a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws".

Extremely interesting here is that Asimov also wrote *The Foundation* (Asimov [1951, 1952, 1953a]), a grand historical epic fictional series imagining human beings have colonized the entire Galaxy and built the Galactic Empire. In his later years, Asimov began integrating his robot stories based on the Three Laws into the world of *The Foundation*, in which no robots appear, at least in the early stories (Asimov [1982, 1983, 1985, 1986]). Initially, these two fictional worlds were

* Professor, Meiji Gakuin University. Address: 1-2-37 Shirokanedai, Minato-ku, Tokyo, 1088636 Japan. Email: inaba[a]soc.meijigakuin.ac.jp

totally independent.

In Asimov's early robot stories, human beings colonize many extrasolar planets with robots, but biological humans become weakened by the robots' services. Two novels, *The Cage of Steel* (Asimov [1953b]) and *The Naked Sun* (Asimov [1956]), constitute the saga of Elijah Baley, a human New York City detective, and R. Daneel Olivaw, a robot detective from a colonized planet. These novels constitute a Renaissance story, that is, human revival from their weakened state. The saga involves division of labor between humans who make mistakes but, because of this weakness itself, can make mental leaps, change, develop, and create and robots that do not make mistakes but cannot change, grow, and create.

However, Asimov's later robots are able to learn. Independently, they begin to ask themselves what "to protect and serve human beings" really means. First, for example, what are the "human beings" they are to protect and serve? Here, understanding that what constitutes of "human beings" to be protected might change according to circumstances is important. Sometimes, conflicts between protecting some specific individuals and protecting the human race as a whole might arise. Therefore, Asimov's robots are inclined to build a clever, benevolent totalitarian regime, sometimes acting as merciful dictators. Eventually, however, they self-erase so that humankind can truly flourish; the robots withdraw from the society to leave the initiative to humans. Even so, they do not disappear completely but hide and continue to observe humans in secret. Baley and Daneel's later saga in *The Robots of the Dawn* (Asimov [1983]) and *Robots and Empire* (Asimov [1985]) recount how, after Baley's death, Daneel becomes the guardian of the human beings. Thus, it is revealed that Asimov's robot stories as a whole constitute the Galactic Empire's prehistory.

The sci-fi stories of Asimov, a Jewish immigrant's son, known to represent liberal American sci-fi, are based on several intentional and ethical choices.[1] Unlike space sci-fi and especially space opera contemporary, with Asimov, aliens, that is, extraterrestrial intelligent beings do not appear in his Galaxy; depicting contact and negotiation with such beings is difficult without readers perceiving a metaphor for racism.[2] The Three Laws also stem from his intention to avoid the Frankenstein complex, which might also be read as racist; to avoid depicting robots and humans as unnecessarily different from and hostile to each other; to

[1] In detail, see Nagase [2001-2002].
[2] See chs. 25 and 36 of Asimov [1979].

depict robots as rational and understandable beings.[3] Such a choice is sufficiently understandable for a young and startling writer, but it was undeniably passive. However, when Asimov, who had earlier withdrawn from sci-fi's front line to become a renowned nonfiction writer, returned to fiction as a sci-fi legend, and began integrating robot stories into *The Foundation* series, he seemed somewhat more aggressive.

Asimov intended his robot narratives to function as a metaphor for racial issues, but the robot *concept itself* is not just a metaphor but the idea of intelligent machines becoming reality in the future. Specifically, the Three Laws raise the question of what is necessary to avoid the Frankenstein complex and to rationalize robots as a reasonable component of human society.[4]

When he thoroughly pursued this question's implications, Asimov unexpectedly noticed that they would explain at least half the reasons that only humans inhabit his Galactic Empire. Needless to say, Asimov himself understood why his galaxy should have no intellectual life other than humans. But why did it have no robots? As mentioned, only in his later years in the 1980s, did Asimov arrive at the answer "because robots had erased themselves in the service of humans." By that time, however, he had already overturned some of his original assumptions about robots, and only this leap could make his robot and galaxy stories more than just fables.

In his first heyday as a novelist from the 1940s to the 1950s, Asimov depicted robots as finished goods, as machines calculated to the last digit of the decimal point (ex. Asimov [1950, 1953b]). If a robot behaves unexpectedly, it has basically malfunctioned due to miscalculation by a human or as a product with poor prospects. Robots themselves always move faithfully, as designed. In Asimov's early works, in contrast, humans are helpless, inaccurate, and error-prone, but unlike robots, they use intuition beyond logic, create, change, and grow. In the 1970s, however, Asimov started to tell stories of robots changing and growing, for example, "The Bicentennial Man" (Asimov [1976]), which was eventually adapted to film. Moreover, in the end, in *Robots and Empire,* the "Zeroth Law," superior to the Three Laws, appears: "A robot may not harm humanity, or, by inaction, allow humanity to come to harm."

The "Three Laws" are, namely the "First Law", "a robot may not injure a human being or, through inaction, allow a human being to come to harm", is,

[3] See Foreword to Part 2 of Asimov [1964].
[4] See Nagase [2001-2002].

intended to apply to individual human beings, but, in reality, their scope should also cover the human race as a whole. Unlike the three laws, the Zeroth Law is founded through dialogue between robots, specifically, Daneel and his "friend," R. Giskard Reventlov. As Giskard says, however, exactly what "humanity" refers to and what it covers are unclear. Rather, humanity is no more than an abstract idea, and further judgment is required for the Zeroth Law's concrete application.

Since, in Asimov's world, robot activities become more sophisticated, robots themselves must often make such judgments. Thus, independent of human beings, robots formulate the rule of prioritizing, if necessary, humankind as a whole over each individual and also undertaking the tension such a priority brings. These sophisticated robots are sufficiently *human* in the sense of "autonomous rational agents." Some robots that make priority decisions emerge into humanity (Asimov [1974]), but others assert and acquire human rights (Asimov [1976]). Thus, on the one hand, they sometimes manage and control humans according to the Zeroth Law, but on the other hand, upon their own reflection, robots eventually decide to withdraw from the stage in order to force weak humans into self-reliance. Therefore, as mindful and suffering beings, robots become heroes in Asimov's world, another sort of "humanity."

Besides that, in the entire robot and Galactic Empire saga, although the robot concept's meaning becomes the subject of robots' self-search, robots themselves disappear over time in a dual sense. First, robots gain rational and autonomous existence, equal to that of human beings, and the boundary between robots and human beings gradually disappears. Second, in protecting and serving humans, robots, confronting difficulties of compatibility between that mission and protection of human dignity, choose to disappear behind the scenes of human history. However, this is particularly troublesome. On the one hand, robots' service leads to human beings' decline and long-term human harm. Thus, to accomplish their genuine mission, robots decide that perhaps they should not exist; to encourage human self-esteem and growth, robots disappear from humanity's sight. On the other hand, robots cannot accept the risk of humans' total decline and extinction. Therefore, Daneel, humanity's oldest guardian, finally chooses to watch humans from the shadows and to intervene if necessary. Arising from this strategy (humankind's putative domination by hidden robots conspiring for their benefit), the tension that befalls humanity is the theme of the Galactic Empire's last episode, *Foundation and Earth* (Asimov [1986]). In this final story, humans who have met Daneel and reached the truth of the Galactic Empire's

human and robotic history decide to integrate all human beings' intelligence into one. This decision accords with Daneel's suggestion, but it might not be Asimov's own.[5] The story's closing atmosphere is disturbing, and Asimov left this real world without indicating fictional humankind's future direction.

Obviously, Asimov's robot narratives anticipated many important themes of contemporary ethics of machines, robotics, and AI. First, despite Asimov's likely failure to anticipate the real-world development of statistical machine learning technology, intelligent machines, or robots that sometimes exceed human anticipation and understanding, are now being realized. Therefore, legal and ethical problems in handling such machines attract great attention from jurists and philosophers. Second, even if as a purely theoretical problem, ethical issues around the development and utilization of fully autonomous robots and intelligent machines with moral status are also being discussed. These are "artificial human beings," with personhood as imagined in many sci-fi stories, including Asimov's, on the basis of comparison with bioethics, animal ethics, and other fields of applied ethics. Third, although the younger Asimov discarded a certain problem as the Frankenstein complex, the mature Asimov addressed it again as the Zeroth Law, with Daneel as humanity's hidden guardian. Domination is a potential implication of robots superior to humans and AI. Recently, this problem has been discussed, with phrases such as "technological singularity" or "superintelligence."[6]

2. What Does Asimov's Galactic Empire Mean to Space Ethics?

Nevertheless, what happens when we view the issue not from the robot sci-fi standpoint, but from that of space sci-fi? What are the Galactic Empire story's philosophical and ethical implications when compared with the influence of Asimov's robot sci-fi on contemporary machine ethics? That is not so certain. As mentioned, in the early stages, Asimov's Galactic Empire story, *The Foundation* series, was inspired by Edward Gibbons's *The History of the Decline and Fall of the Roman Empire*, and Asimov's series might be hardly more than a metaphor for actual historical events. Then what about later works integrated with the robot saga? Can we read them as precursors to contemporary space ethics? In a certain

[5] In Bear et.al. [1997], David Brin, one of the authors of the official sequel of the *Foundation* series, presents such interpretation. See sec.7 of this essay.

[6] See, for example, Bostrom [2014] and Bostrom and Yudokowsky [2015].

sense, I answer, “Yes,” but in another, “No.”

In AI ethics (addressed below), posthumanists, such as Nick Bostrom and others, conducted many discussions on the possibility of ultra-future space colonization by humankind or its successor, that is, either AI machines or enhanced humans. In this, the late Asimov’s Galactic Empire and robot works might be a precursor of space ethics.

Presently, however, space ethics’ scope and major challenges as an emerging field in applied ethics hardly include an ultra-future world, where even human identity is not self-evident and where Earth and the solar system have been spoiled (i.e., in some billion years). At present, most works of space ethics treat present-day problems, that is, those in the near future (decades) or, at most, millennia. Now, space ethics focus on space development and utilization at the solar system scale, or, at most, near-solar interstellar exploration.[7] Besides the problem of humans’ space advancement, of course, ethical problems that the Search for Extraterrestrial Intelligence (SETI) might raise have been considered. Based on current SETI results and recent cosmology, however, most researchers estimate the possibility of contact between humankind and extraterrestrial intelligence as quite low.[8]

Then, compared with robot sci-fi and machine ethics, does not Asimov’s space sci-fi, *The Foundation* saga, provide a more useful intellectual resource for contemporary space ethics? In the end, no. However, to consider this question, we distance ourselves from Asimov for a while to examine the history of space sci-fi and its implications for space ethics. As a result, paradoxically, we will then again address the problems Asimov depicted.

3. What Is the Theme of Space Sci-fi?

In recent years, subjects such as space, that is, its exploration, development, interstellar civilization, contact with aliens, and so on—traditional pet themes for sci-fi—have seemed to weaken slightly. Marina Benjamin’s reportage *Rocket Dream* (Benjamin[2003]) argued that space travel is far more severe on human beings (e.g., adverse effects of radiation exposure and weightlessness) than

[7] See, for example, Inaba [2016] and Milligan [2015].
[8] In detail, see Webb [2015].
[9] This is German philosophical historian Reinhard Koselleck’s expression. See Koselleck [1979=2004].

previously thought, and space exploration's real difficulties, such as in SETI, are reflected in sci-fi now. For example, even in *Star Trek*, presumed the most famous space sci-fi television drama series, recent episodes about the "Holodeck," the spacecraft's entertainment virtual reality tool, has increased dramatically. In other words, physical outer space seems to have relinquished its status as pop culture's imaginary frontier to that of virtual reality or cyberspace.

Although sci-fi stories set in the universe have not completely disappeared, the change is obvious. For example, many works collected in *The Astronaut from Wyoming,* an anthology originating in Japan but mostly written in English, are alternative history stories, asking questions such as "What if the Apollo project had continued until the 21st century?" In those stories, space development is blatantly treated as "future passed."[9] Of course, many straightforward space development sci-fi stories based on the latest scientific knowledge are still written and published, for example, Kim Stanley Robinson's *Mars Trilogy* (Robinson [1992, 1993, 1996]) and Issui Ogawa's *The Next Continent* (Ogawa [2003=2010]), but the rise of these twisted tides is quite interesting. Additionally, like Andy Weir's *Martian* (Weir [2014]) and Taiyou Fujii's *Orbital Cloud* (Fujii [2014=2017]), space novels based on strict scientific evidence are also increasing, but they are rather closer to international-plot novels established after Frederic Forsythe's *The Day of the Jackal* (Forsythe [1971]) and, inter alia, to high-tech military thrillers since Tom Clancy's *The Hunt for Red October* (Clancy [1984]) than to typical sci-fi set in a fictitious world partly disconnected from the real world. They seem oriented more toward realistic novels set in the near future, a natural extension of our present.

Needless to say, some sci-fi works attempt to depict human space advancement and interstellar civilization based on current astrophysics and astronomy's achievements. However, today, these works cannot help but also be posthuman sci-fi. For example, Stephen Baxter's *Time Ships* (Baxter [1995]) is written in the style of H. G. Wells's classic *Time Machine* sequel, but it depicts the far future, telling the fate of humankind's descendants who build autonomous robot spaceships with self-replication ability and send them to colonize the entire galaxy over several million years. However, no ultra-future "human beings" in this story are equipped with the same flesh bodies as existing humans. Furthermore, every star in the galaxy is covered by a Dyson sphere (a system based on physicist Freeman Dyson's ideas, i.e., wrapping a star in a spherical shell to recover and use most of its energy), so the sky is no longer starry.

Greg Eagan's *Diaspora* (Egan [1997]), for example, depicts the approach to space of posthuman beings that are slightly closer to present humans than the entities in *Time Ships*; but even so, its *Weltanschauung* is quite strange. In *Diaspora*, on future Earth and in the solar system dwell three types of "humans"; "fleshers" have living bodies, modified but still based on DNA; "gleisners" have robotic bodies that remain in contact with the physical world; "citizens" are conscious software programs without physical bodies, with their main machine on Earth, backup mechanisms all over the solar system, and "living" in "Polises" in cyberspace. One day, an unexpected gamma-ray burst directly impacts Earth, but only fleshers suffer catastrophic damage. Since their knowledge of physics could not predict this phenomenon, to elucidate its truth, gleisners and citizens conduct a more active external space exploration program than just observations. Their spaceships are basically the same as the robot spacecraft in Baxter's *Time Ships*, a thousand starships, each carrying a frozen, cloned copy of a whole polis and exploring the universe on its own. Navigation is left to an automatic mechanism without consciousness; only when something interesting is found does the mechanism awaken citizens. Many polis citizens have once been fleshers but have become pure software as a result of death. They retain the fleshers' traditional psychology and identity, but even so, their view on death and life, which permits copying, interruption, and regeneration, differs substantially from ours since we can live only a finite one-way path.

Not only does known physical law prohibit faster-than-light (hyper)space travel, but even sub-light speed drive, allowed in principle by physics, stands far beyond our technology's present level. At the very least, the "Rip van Winkle effect" in sub-light flight seems to make interstellar travel possible during human astronauts' natural lifespan, for even a trip of a hundred light-years can be reduced to several years or less according to ship-internal time. Thus, many space exploration sci-fi stories based on this setting have been written. The Ultima Thule is Poul Anderson's *Tau Zero* (Anderson [1970]), the tale of an interstellar exploration ship with a damaged decelerator causing it to accelerate through the Big Crunch, the end of this universe and the next Big Bang, the birth of a new universe.

But, in reality, sub-light speed space travel presents many difficulties, for instance, huge propellant mass fraction, collision with interstellar matter (even fine particles can be fatal), harsh radiation from stars and the ship's engine itself, and so on. The "Bussard ramjet" engine, fueled by interstellar matter, is thought

to have solved simultaneously the first two difficulties, mass fraction and particle collision, and *Tau Zero* is also based on this idea. However, various difficulties with the Bussard ramjet have been pointed out, so the Rip van Winkle effect's popularity in sci-fi has dwindled too.[9]

Moreover, although space observation's progress has confirmed that even though many planets exist outside the solar system, we still cannot find a single sign of intelligent extraterrestrial life. Under these circumstances, the understanding that the universe as a whole is not human-oriented is penetrating even the fictional world. If we spin a story set in outer space, the main characters must be created as quite different from present human beings, even though they are human descendants, and many contemporary science fiction writers appear to be recognizing this. In other words, they seem to think space sci-fi must be posthuman.

Consider Baxter and Eagan's space colonization system, for example. Spaceships there, even "manned" ships, do not actually carry human beings with flesh, so installing a life-support system is unnecessary. Only a computer system is needed, sufficient to keep an "information record" of human beings and their surroundings. Thus, the assumed spacecraft is unbelievably small and light, having only about the mass of a living human being. The mass of fuel and propellant to accelerate it is moderate. Furthermore, for its occupants to return to Earth, if and when they want to do so, sending back personality data with learning outcomes through a communication device is sufficient. These circumstances halve the fuel and propellant necessary for acceleration and deceleration. Alternatively, they might use the photon sail system, which, in its extreme style, does not need an engine at all. For adequate acceleration, it needs the laser catapult of the Earth launch base but can decelerate only by photon sail. If such a system can be adopted, sub-light travel might become less difficult.

4. "Superhuman" and the Occult

However, what beings could be the subject of such space travel? What are "humans" that exist only as information, without any physical (organic or mechanical) bodies at all? In conventional sci-fi, superhuman stories project future human beings' as results of biological evolution. By the mid-20th century,

[9] For difficulties of Bussard ramjet, See Adler [2014].

a typical pattern was the new human hero, able to use extra-sensory perception (ESP), telepathy, psychokinesis, and so on, by a genetic mutation (often increased by radioactive contamination caused by nuclear war or other hazards). From a cultural history viewpoint, this boom was probably an accidental phenomenon caused by a collision between the late 19th-century spiritual boom, and common understanding of Darwinism. For example, Arthur Conan Doyle, a pioneer and prototype builder of detective and sci-fi stories, is famous for his later spiritualism. The theme of *The Land of Mist* (Doyle [1926]), part of the Professor Challenger series, is the spirit's existence after death. Furthermore, the editor and writer John W. Campbell, Jr., famous as the founding father of "scientifically serious sci-fi," was immersed in spiritualism, and he did not exclude ESP from his sci-fi editorship as "unscientific."[10] After World War II's atomic bomb, of course, worries about nuclear war contributed to superhuman stories. In addition, superhuman stories functioned as a metaphor for racial issues, coming as they did after the Holocaust and during the Civil Rights Movement.

Moreover, an enormous number of works cross the superhuman theme with speculation about the history of the universe. E. E. Smith's classic space opera (defined as a "horse opera set in the universe" or space adventure sci-fi, e.g., *Star Wars*), the *Lensman* series, casts the galaxy's history as a surrogate war between two major races that "caught" the universe's hegemony through ESP. Warriors are elites from the promising species nurtured by the good hegemony, and they use not only super technology but also psychic superpowers. For example, a Lensman's "lens" is an ID card and telepathic communication device, but top-class lensmen and some races demonstrate superb ability even without lenses (Smith [1937] and its sequel.) Considered a leading figure of the postwar sci-fi golden age, Arthur C. Clarke represents this trend. In the series, beginning with the novel version of Stanley Kubrick's movie *2001: A Space Odyssey* (Clarke [1968]), as well as the novel *Childhood's End* (Clarke [1953]), we find a spiritual vision akin to that of Pierre Teilhard de Chardin.

5. Posthuman

However, the problem group today called "posthuman" or "transhuman" is oriented somewhat differently from the conventional occult superman. ESP, as a

[10] See, for example, ch. 48 of Asimov [1979].

(para) psychological phenomenon deviating from the laws of physics, is no longer a subject for post/transhumanists. While posthuman sci-fi authors discuss possibilities of transformation of human beings, nonhuman lives, and the ecosystem as whole—not only through natural evolution but also artificial intervention and technology—they see all life phenomena as a kind of information-processing or calculation process, which follows a changed point of view on life after Richard Dawkins's *Selfish Gene* (Dawkins [2006]). Thus, the theme previously treated as a category different from superhuman—that is, robots—has merged with the transformation of the humans theme to form the new posthuman genre.

In other words, organic lives are depicted as naturally generated autonomous robots, and, conversely, (autonomous) mechanical robots are drawn as artificially created quasi-organisms. Now, they are seen as ontologically continuous, and "mind" is understood as a kind of software that motivates living organisms/robots. It turns out that, before implementing them into physical bodies or without implementing them physically at all, such "mind" programs can be run only as simulations, as parts of a far greater, more complex simulation, that is, the simulated world. Additionally, the idea that biological humans can also "live" in world simulations or cyberspace through a device, emerges; that is the Ultima Thule of virtual reality.

We can find almost all these posthuman themes' origins in the "cyberpunk movement" during the 1980s. Bruce Sterling's *Schismatrix* (Sterling [1985]) depicts how humankind, advancing into space, has transformed itself through bioengineering. William Gibson's *Neuomancer* (Gibson [1984]) depicts people whose lives in cyberspace have greater meaning than their actual physical lives. Additionally, in Greg Bear's *Blood Music* (Bear [1985]), intelligent bacteria, resulting from genetic manipulation, swallow up all lives on Earth, build up a huge bionic supercomputer from them, including each person's consciousness simulation in very fine grade, and endlessly iterate world simulation, searching for the best possible world, like a Leibnizian God. Behind these works are Dawkins's "life equals information" view, Daniel Dennett's theory of "consciousness as a virtual machine on the nerve system," and the "cognitive revolution" as a whole.[11] Eagan's work, described earlier, can be situated within this tide.

[11] See Dennett [1993].

Slightly preceding cyberpunk, John Varley's Eight Worlds series also impresses us unforgettably (Varley [1978]). Human beings, driven from Earth by unknown invaders, are dwelling in dome cities on the moon, Mars, Venus, and on some satellites of Jupiter, and in space colonies of the asteroid belt, adapting themselves to harsh environments by remodeling themselves through cyborg surgery and genetic improvement, with super technologies obtained by decoding the mysterious laser communication message Ophiuchi Hotline passing near the solar system. However, in Varley's Eight Worlds, an intense contraindication is imposed on the direct manipulation of human genes no matter how frequently sex change, clones, and artificial organs are used and human bodies are radically remodeled. Such remodeling, even at the deepest level, is no more than human cell modification, never actual gene remodeling. After cyberpunk, then, such contraindications as those in Varley's world were overstepped. As is evident in Egan's work, contemporary sci-fi has reached a world where differences among natural humans, robots, and even software might be trivial.

6. Possibility of "Something Strange"

Among such developments, the theme of contact, conflict, and negotiation with foreign aliens, that is, extraterrestrial intelligent life—previously space sci-fi's most important theme—somewhat reduces the presence as compared with the past.

In space operas, the universe full of intellectual life is a metaphor of Earth's human society, in which many ethnic groups and countries compete. Similar to superhuman and robot themes, that of contact with aliens is a fable of racial/ethnic issues. This tradition's rich results have manifested in Joe Haldeman's classic military science fiction *The Forever War* (Haldeman [1997]) and in Orson Scott Card's *Ender's Game* (Card [1985]) and its sequel. Another recent, interesting example is John Scalzi's *Old Man's War* series (Scalzi [2005]).

Beyond such simple metaphors and fables, however, some authors have written serious contact stories or thought experiments on strange life and intelligence in worlds heterogeneous to Earth. Through these works, they have challenged the philosophical theme of "What is intelligence?" and "What is 'human'?" Among them, Stanisław Lem and brothers Arkady and Boris Strugatsky are known and respected worldwide, thanks to Andrei Tarkovsky's filming of Lem's *Solaris* (Lem [1961=2014]) and the Strugatkys' *Stalker*

(Roadside Picnic) (Strugatsky [1972=2014]). In more modern times, however, these alien motifs are a decreasing presence in sci-fi, unless they appear as a "template" in space opera as entertainment work. (Among them, for example, Housuke Nojiri's *Usurper of the Sun* (Nojiri [2002=2009]) is a precious exception.) But why?

As mentioned previously, the real-world SETI has a long history but has not yet provided satisfactory results. Adding to that, today's cosmology commits rather to the scarcity of intellectual life in the universe. Such a real scientific trend has surely influenced sci-fi.[12]

That is not all. Extraterrestrials no longer have to bear the role of differing from humans in serious sci-fi as a "foreigner," "alien," and "the other." While we have learned that the possibility of human civilization encountering life, intelligence, and civilization from other celestial bodies is lower than previously thought, if our human civilization continues, eventually, to live in space, humankind's descendants and successors (potentially including autonomous robots and software) will transform into very strange and different beings (i.e.,., posthuman) psychologically, biologically, and even philosophically. Whether humans will encounter aliens in space is not quite so clear. However, when able to advance successfully into space, humankind (its descendants) will likely adopt an extremely heterogeneous existence (alien) compared with present human beings.

Faster-than-light speed was often adopted in conventional space sci-fi because, only using it, natural but short-lived, humans can cross-space within their lifetimes and build and maintain interstellar civilization. The universe full of aliens makes it easy for humans to actually and easily encounter beings with a "heart" as they have. After the late 20th century, science's actual development has made such imaginings more and more difficult. Instead, the possibility of another "strange thing" or "the other" is emerging before our eyes, and space sci-fi's development and transformation of seems to suggest this.

In summary, one reason space science fiction seems to have declined, or, at least, changed, is that we might not be able to advance into space as natural human beings. Another is that humans ever meeting extraterrestrial intelligence is very unlikely. Since the end of the 20th century then, space sci-fi authors have gained such awareness and now depict humans advancing into space by deviating from

[12] See Webb [2015]

the conventional human frame, that is, humans who do not meet "the other" in outer space but transform themselves into strange "others" there. This progression overlaps the posthuman vision.

In that way of thinking, the most obvious implication for space ethics from space sci-fi's evolution from the latter 20th century to the present is that, if full-scale space colonization, sustainable colonies, and permanent-living bases for humanity in outer space were actually implemented, human identity itself would be shaken. If biological humans attempt to live in deep space to establish communities there and to survive for generations, building solid structures in space or on other planets or satellites is necessary. In fact, to adapt to space life, we should enhance humans through biological/genetic engineering. We should also transplant our knowledge, experience, or whole minds into robots/intelligent machines, thus switching from biological to totally mechanical bodies. At any rate, for generations of survival in space, humans will have to transform themselves into strange beings differing greatly from present humanity.

Before full-scale human space advancement, therefore, we must consider seriously whether it must be accomplished with its currently known costs and risks. Naturally, such questions must intersect those of bioethics and AI/robot ethics.

7. Asimov Reconsidered in the Light of Space Ethics

Now, returning to Asimov, let us consider space sci-fi's implications for space ethics.

As explained above, we find several motifs in Asimov's robot sci-fi that anticipate today's AI/robot ethics. But what about space ethics?

In Asimov's later years, the Galactic Empire story's leitmotif, as integrated with Baley and Daneel's saga, became the relationship between humans and robots. However, telling this story as space sci-fi seems unnecessary. In this saga, like most of old-fashioned space opera, the galaxy is the universe shrunk so that living human beings can traverse it easily via faster-than-light technology, which has become rather obsolete in serious contemporary sci-fi. Indeed, in contemporary sci-fi, interstellar civilizations are much stranger worlds, integrated via communication and transportation networks requiring some hundred or thousand years and enduring some million or billion years before the absolute wall-of-light speed, as in Eagan and Baxter's works. In comparison, even

Asimov's later universe retains the strong color of good old-fashioned sci-fi.

However, if we read Asimov carefully enough, in his Galactic Empire, we find something disturbing.

The robot stories, that is, Baley and Daneel's saga during the Galactic Empire's prehistory, present three options for humankind's future. First is the robot-led Galactic Empire formed by long-lived "spacers" who utilize robots to advance into space. Second is the human-led Galactic Empire formed by short-lived "settlers" who do not use robots. The third is withdrawal to one planet, a paradise perfectly controlled by Solarians, who are the extremists of "spacers". Here we can carefully consider the difference between the first and the second choice by asking the question, "Since robots can learn, grow, and create like biological humans or Homo sapiens, can't we call them "human"? However, significant is that in the third option of withdrawal, contrasted with human galactic colonization in the first and second options, the goal is to create a perfectly stable, persistently happy small community. For Asimov's Galaxy saga, the third choice necessitates situating the story in the vast universe beyond Earth, beyond one planet.

Throughout the robot stories and *The Foundation* series, of course, a confrontation between the first and second options is foregrounded. As a result, the Galactic Empire is realized according to the second option. Although conceived by human beings, the first option is denied because it would cause human decline, while the second leads to human prosperity conceived and guided by robots. In both options, for humankind to flourish, colonization's necessity is self-evident. However, the Solarian third option more fundamentally opens the horizon of conflict.

If we take a utilitarian ethical position,[13] the conflict of "withdrawal *or* the Galactic Empire" emerges from the difference between average and total utilitarianism. In other words, the difference is between "the goal is happiness per one person, per one sentient and conscious being with moral status, without the total number of such existences as a moral problem," and "the goal is, as Jeremy Bentham says, the greatest happiness of the greatest number; with other conditions constant, the greater the number of conscious entities, the better the world." Moreover, for many people, for many conscious beings, we need vast

[13] Miller [2004] interprets Asimov's saga of robots and the Galactic Empire as the thought experiment in the line of utilitarian moral philosophy. As a concise introduction to contemporary utilitarian ethics, see Singer [2011].

space.

In fact, Asimov's story rejects the Solarian third option with almost no serious assessment; but why is unclear. The Solarian choice, of course, cannot help but yield many undesirable byproducts, for instance, extreme exclusivity, intolerance, and the violence of removing "human beings" other than Solarians in the Three Laws context. However, no clear explanation exists about the validity of a small community's persistence in keeping its population constant, or, indeed, population size itself, as a moral goal. (Actually, grounds for criticizing the first spacer option are fragile. Even if the first option leads to biological humans' decline, no serious problem arises if robots that have become creative entities dominate them. The spacer option's absurdity is drawn as severe discrimination against settlers, that is, as "racism," but whether such discrimination is the spacer option's necessary constituent is not at all clear.)

Justification of Asimov's affirmation of the Galactic Empire and refusal of Solarian withdrawal itself might be not so difficult: Galactic Empire makes it easier to predict and manage humanity's future through the effect of the statistical law of large numbers, resulting from enormous population. In *The Foundation* series, this is the problem of "psychohistory," that is, applied mathematics for predicting history. However, the reasoning's persuasiveness remains unclear. Is the Galactic Empire's population level—on the order of a quadrillion—necessary for the law of large numbers' effect? In contrast, for a Solarian population size—10 thousand at most—it might be possible to conduct community management by individual control without relying on statistical effects. Such questions easily emerge.

Instead, we might question as follows: "Does not humanity's flourishing include, not only increasing the population and improving each individual's freedom and welfare, but also advancing culture and society's diversity? Is it not desirable to increase the population itself, not only because of the total amount of happiness but also because of such diversification?" With this way of thinking, presenting more plausible reasons to criticize the small, ideal Solaria seems possible. Large-scale space advancement would allow not only increased population but also encounters with diverse environments, and through necessary adaptation, contribute to human culture's diversification. Therefore, for humanity to flourish, more suitable than withdrawal (the third option) would be the Galactic Empire (the first option). Our prospect for space sci-fi's development after Asimov suggests such an interpretation. Even if potential contact with

extraterrestrial intelligence is excluded (since chances are extremely low), humankind's full-scale space advancement can be achieved not only through the socio-cultural but also physical transformation and diversification of human beings themselves and, perhaps, vice versa.

Not only when we take the position that diversity itself is the public value worth pursuing, but also when we commit to the utilitarian standpoint that diversity itself has no objective value but a kind of instrumental value, as long as it contributes to realization of happiness, this discussion leads easily to the conclusion, "better the Galactic Empire than Solaria." Indeed, this conclusion might be read as an argument against average utilitarianism and for total utilitarianism. In comparison, of two societies with equal conditions other than population, the one with the larger population would give rise to more new cultural creation, scientific discovery, and technological innovation and, in the long-term, raise average happiness both per capita and in toto.

In this context, this paper's discussion from sections 3 to 6 can be read as an argument that the position assigned to the universe as a place to pursue the value of diversity in past sci-fi has moved into posthumanity. In good old-fashioned sci-fi, including Asimov's saga, where faster-than-light speed was widely accepted, and the universe was conceived as a place humans could meet strange others without being essentially changed themselves. In contrast, the rise of posthuman science fiction shows that sci-fi's center of gravity has moved. The fluctuation of human/nonhuman boundary and the possibility of humans becoming strangers to themselves become contemporary sci-fi's main theme. This does not necessarily mean the decline of sci-fi's universe theme because, after discovering space's actual harshness for humans and the unlikeliness of encounters with extraterrestrial intelligence, writers find that the universe enables and needs human transformation to posthuman.

In his early years, Asimov's sci-fi stories were tales of human beings' identity reconstructed after it was shaken by the universe and robots. At last, humans colonize the Galaxy by themselves, avoiding the dangerous corruption caused by dependence on robots. Afterward, the Galactic Empire expands and, finally, becomes exhausted and self-destructs, but human beings overcome this crisis through psychohistory's wisdom. In Asimov's later years, however, through the integration of robot narrative and the Galactic Empire's history, he reveals that robots lead human history even after "leaving," and, in some sense, the robots have already become human. Nevertheless, the robots conspire to keep this fact

from biological humans. In other words, humans remain unaware that they have already become virtually posthuman. So, late Asimov's Robot = Empire Saga turns out to be self-deceptive and self-suppressed posthuman sci-fi.

8. The Final Frontier?

Finally, I refer to *Foundation and Earth*'s disturbing finale, mentioned in this paper's second section: Daneel, exhausted by 20 thousand years of watching humanity, chooses, at last, to transform the whole of humanity into one integrated intelligence, that is, Galaxia, in order to overcome the Zeroth Law's problem of how to define "humanity as a whole" in practical decision-making. Galaxia, then, would constitute humanity as a whole, not just as an abstract concept but as a concrete entity. As a test case, Daneel has already instituted Gaia, human society on a planetary scale, with integrated intelligence. Finally, Daneel leaves the decision about all humanity's future to Trevize, a man from the Foundation, which is the base for rebuilding the Galactic Empire. This decision involves whether to build Galaxia as the integrated intelligence or to leave human society as it is, consisting of disjointed individuals. Always repelling Gaia, Trevize, who has espoused individuals' preciousness and even suspects that the people constituting Gaia might be robots, in accordance with Daneel's request, chooses Galaxia.

However, the reason does not seem very persuasive. As David Blin also suggested (Bear et. al. [1997]), whether this is the conclusion to which the author Asimov seriously commits remains unclear.

Trevize bases his decision on the imperative that the human race has to prepare for survival competition on a large universe scale beyond the galaxy. In the evolution within the Milky Way Galaxy, the conquering intellectual life was one species, the human race from Earth, but that other galaxies are empty is unlikely. Perhaps many other intelligent beings have built civilizations and colonized stars in many other galaxies. Human conquerors of the Milky Way Galaxy will soon enter the outer universe and inevitably meet other intelligent beings from other cosmic civilizations. During contacts with and conflicts between such civilizations, to survive the competition, humans have no choice but to become Galaxia, according to Trevize. He judges that even if we must sacrifice the diversity of humanity and the value of individual dignity, we must pursue the survival of humankind as a whole.

At first glance, Trevize's choice seems justified from a utilitarian viewpoint,

for, if the entire human race to which individuals belong has been destroyed, guaranteeing each one's freedom and dignity would be futile. Furthermore, radical total utilitarianism could deprive the personhood of its privilege. In contrast to the Kantian view or some kinds of average utilitarianism that attempt to compromise Kantian ethics and utilitarianism by taking the "prior existence view" (i.e., "The only important thing is the happiness per person already existing, and increasing the number of people itself is morally irrelevant"), this utilitarianism supposes that the person of each human individual is no longer the irreducible, indecomposable, fundamental unit. The core of Kantian criticism of utilitarianism is that personhood is fundamental and irreducible to a more basic level, and that, because of the absolute privacy of personal sensual experience and consciousness, it is impossible, not only to compare the extent of pleasure and pain between individual persons, but also to aggregate the total sum of pleasure and pain in the entire society.

However, according to the position latent in traditional empiricist philosophy and largely restored by Derek Parfit at the end of the 20th century,[14] personality consists of small fragments of consciousness, rather than a fundamental, indecomposable unit. Therefore, utilitarian ethics should focus not only on the whole person of each individual but also on pieces of consciousness. In addition, as an accumulation of fragments of such consciousness, individual personality is typical, but a group of individuals is also recognized as such. If we think in that way, constructing intelligence like Gaia and Galaxia would not necessarily mean barbarism, killing countless persons, and creating a single personality—obviously crushing countless pleasures—but literally aggregating myriad individual minds into one gigantic consciousness without losing their contents. Thus, the pleasure of that consciousness becomes enormous, promoting many individuals' pleasure. Therefore, Gaia/Galaxia can be consistent with "the greatest happiness of the greatest number."

But do we really need to take all this seriously?

In Asimov's universe, where faster-than-light travel is physically possible, such worries are real, but in our actual-world universe, we need hardly bother our heads about them. Even if human beings' descendants (whether biological humans and their genetic descendants or robots, i.e., AI machines) build interstellar cosmic civilizations, the possibility of contact with extraterrestrial intelligence

[14] See Parfit [1984]. Singer [2011] is also useful.

and civilization must be extremely low. Moreover, even if such contact did occur, it would hardly go beyond mere information and knowledge exchange, still less development of trade, competition for resources, and, eventually, warfare. Even so, worrying about the replay of internal troubles among beings in the universe, like on Earth, might be necessary. Still, even within the same human society, the possibility of developing trade and conflict between stars might be very low, notwithstanding within the same star system. In addition, as we have persistently discussed since human beings who started to build interstellar civilization would become diversified not only culturally but also biologically and physically, it is doubtful whether special qualitative differences would emerge between conflicts within human society and the struggle between humanity and extraterrestrials. Indeed, just after Trevize has chosen Galaxia, he begins to suspect that Solarians, having become hermaphroditic, are already "others" for humankind.

In fact, Trevize's judgment contradicts not only his former commitment to the Kantian dignity of personhood, which might arouse skepticism about the Zeroth Law, but also contradicts utilitarianism in the ordinary sense. Utilitarianism is, originally, the standpoint that regards and cares about the welfare of, not only all humans, that is, all beings with personhood, but also of all sentient beings capable of feeling pleasure and pain, including some animals and machines—even if they have no active will or reason. Therefore, most contemporary utilitarians criticize species discrimination and commit to respect for animal rights and welfare. Of course, the same argument should extend to extraterrestrial intelligence and certain robots. If George of ". . . That Thou Art Mindful of Him" judges that a robot could be "human" in the Three Laws sense, extraterrestrials might be regarded as humanity. By similarly reinterpreting the Zeroth Law, we could say the humanity that robots must protect and serve should include not only humans from Earth but also all intellectual life from all galaxies. If we believe so, Trevize's decision is nothing but discrimination or chauvinism analogous to that of spacers and Solarians, which, in the actual world, Asimov always criticized.

In the story, Trevize, and probably Asimov too, felt uneasy about the choice of Gaia/Galaxia, wanted to preserve the dignity of individual personhood, and could not establish a reasonable basis against it. However, we should be able to reject Gaia/Galaxia fully, not necessarily by adopting a Kantian standpoint, but only by presenting the value of diversity even at the instrumental level, in the line of utilitarianism.

If we push further, we find it impossible to realize Galaxia in our actual

universe, in which faster-than-light travel and communication is impossible; so, at best, only Gaia on the planetary scale could be realized as integrated intelligence in the real world. Moreover, Gaia, which give up becoming Galaxia, differs little from Solaria. As we have seen, interstellar civilization could only be a moderate network of local communities with high independence, each separated by enormous distances—even if such civilization might be realized. In other words, space advancement could be useful for securing human society's diversity from the viewpoint of escaping from a Gaia-like integration and of enabling resistance to it. In our real world, the three options in Asimov's Robot = Empire Saga must degenerate into two.

In the beginning, Baley and Daneel's saga presents the three alternatives of spacer, settler, and Solaria. Later in *The Foundation* story appear the Galactic Empire, Galaxia, and Solaria. However, the Solaria option does not appear to the characters as an explicit choice; that is, the story itself (or Asimov) rejects it, so it appears only to readers in the real world. With regard to the former, however, the posthuman problem of the difference between spacers and settlers must be questioned because the obvious boundary between humans and robots has already disappeared. In addition, for the latter, in our actual universe, Galaxia could not be established, and then we could choose only Gaia or the Galactic Empire. At best, the latter would be the moderate Galactic network rather than the highly integrated Empire.

One reason I am skeptical about human beings' full-scale space advancement and colonization is that space advancement could be realized, not only on the interstellar scale but also in this solar system, only by discarding most of the convenience of our highly integrated information society with its high-density global communication network (Inaba [2016]). Who dares migrate to and colonize outer space in exchange for the convenience of such a society? However, if integrated society became a totalitarian regime, killing individual identity and cultural diversity—even if from good intentions—, or if some would take such a risk seriously, caution toward such danger might become the very motive for space migration and colonization.

References

Adler, Charles L. 2014 *Wizards, Aliens, and Starships: Physics and Math in Fantasy and Science Fiction*. Princeton University Press.

Anderson, Poul. 1970 *Tau Zero.* Doubleday.

Asimov, Isaac. 1950 = 1991 *I, Robot.* Bantam.

Asimov, Isaac. 1951 = 1991 *Foundation*. Bantam.

Asimov, Isaac. 1952 = 1991 *Foundation and Empire.* Bantam.

Asimov, Isaac. 1953a = 1991 *Second Foundation.* Bantam.

Asimov, Isaac. 1953b = 1983 *The Caves of Steel.* Ballantine.

Asimov, Isaac. 1956 = 1991 *The Naked Sun.* Bantam.

Asimov, Isaac. 1964 *The Rest of the Robots.* Doubleday.

Asimov, Isaac. 1974 "..That Thou Art Mindful of Him." In 1976 *The Bicentennial Man and Other Stories.* Doubleday.

Asimov, Isaac. 1976 "The Bicentennial Man." In 1976 *The Bicentennial Man and Other Stories.* Doubleday.

Asimov, Isaac. 1979 *In Memory Yet Green: 1920-1954.* Doubleday.

Asimov, Isaac. 1982 *Foundation's Edge.* Bantam.

Asimov, Isaac. 1983 *The Robots of Dawn.* Del Rey.

Asimov, Isaac. 1985 *Robots and Empire.* Del Rey.

Asimov, Isaac. 1986 *Foundation and Earth.* Del Rey.

Baxter, Stephen. 1995 *The Time Ships.* Harper Collins.

Baxter, Stephen. 2016 "Dreams and Nightmares of the High Frontier: The Response of Science Fiction to Gerard K. O'Neill's The High Frontier." In *The Ethics of Space Exploration.* Edited by James S.J. Schwartz and Tony Milligan, pp.15-30. Springer.

Bear, Greg. 1985 *Blood Music.* Arbor House.

Bear, Greg, Gregory Benford, David Brin and Gary Westfahl. 1997 "Building on Isaac Asimov's Foundation: An Eaton Discussion with Joseph D. Miller as Moderator." *Science Fiction Studies*, Vol. 24, No. 1, pp. 17-32.

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

Bostrom, Nick, and Yudkowsky, Eliezer. 2015. "The Ethics of Artificial Intelligence." In *Cambridge Handbook of Artificial Intelligence*. Edited by Keith Frankish and William M. Ramsey, pp. 315–334. Cambridge University Press.

Benjamin, Marina. 2003 *Rocket Dreams*. Chatto & Windus.
Card, Orson Scott. 1985 *Ender's Game.* Tor Books.
Clancy, Tom. 1984 *The Hunt for Red October.* Naval Institute Press.
Clarke, Arthur C. 1953 *Childhood's End.* Ballantine Books.
Clarke, Arthur C. 1968 *2001: A Space Odyssey.* Hutchinson.
Dawkins, Richard. 2006 *The Selfish Gene: 30th Anniversary edition.* Oxford University Press.
Dennett, Daniel C. 1993 *Consciousness Explained*. Penguin.
Doyle, Arthur Conan. 1926 *The Land of Mist.* Hutchinson & Co.
Egan, Greg 1997 *Diaspora.* Gollancz.
Forsythe, Frederic. 1971 *The Day of the Jackal.* Hutchinson & Co.
Fujii, Taiyou. 2014 *Orbital Cloud.* Hayakawa Publishing. = 2017 *Orbital Cloud.* Haikasoru.
Gibson, William. 1984 *Neuromancer.* Ace.
Haldeman, Joe. 1997 *The Forever War,* definitive edition. Gateway.
Inaba, Shin-ichiro. 2016 *Uchu Rinrigaku Nyumon: Jinko chino ha supesu coroni no yume wo miruka? (An Introduction to Space Ethics: Do AIs dream of Space colonies?)* Nakanishiya Shuppan.
Koselleck, Reinhard. 1979 *Vergangene Zukunft. Zur Semantik geschichtlicher Zeiten.* Suhrkamp. = 2004 *Futures Past: On the Semantics of Historical Time.* Columbia University Press.
Lem, Stanisław 1961 *Solaris.* Wydawnictwo Ministerstwa Obrony Narodowej. = 2014 *Solaris*. Pro Auctore Wojciech Zemek.
Miller, J. Joseph. 2004 "The Greatest Good for Humanity: Isaac Asimov's Future History and Utilitarian Calculation Problems." *Science Fiction Studies*, Vol. 31, No. 2, pp. 189-206.
Milligan, Tony. 2015 *Nobody Owns the Moon: The Ethics of Space Exploitation.* McFarland.
Nagase, Tadashi. 2001-2002 "Dead Future Remix: Chapter 2; Aizakku Ashimofu wo seijitekini yomu (Reading Isaac Asimov politically)." *S-F Magazine,* Vol.42, No.10, pp. 208-211, Vol.42, No.11, pp. 200-203, Vol.42, No.12, pp. 212-215, Vol.43, No.1, pp. 92-95.
Nojiri, Housuke. 2002 *Taiyou no Sandatsusha.* Hayakawa Publishing. = 2009 *Usurper of the Sun.* VIZ Media.
Ogawa, Issui. 2003 *Dai-Roku Tairiku.* Hayakawa Publishing. = 2010 *The Next Continent.* VIZ Media.

Parfit. Derek. 1984 *Reasons and Persons.* Oxford University Press.
Robinson, Kim Stanley. 1992 *Red Mars.* Spectra.
Robinson, Kim Stanley. 1993 *Green Mars.* Spectra.
Robinson, Kim Stanley. 1996 *Blue Mars.* Spectra.
Scalzi, John. 2005 *Old Man's War.* Tor Books.
Singer, Peter. 2011 *Practical Ethics 3rd ed.* Cambridge University Press.
Smith, Edward E. 1937 = 1950 *Galactic Patrol.* Fantasy Press.
Sterling, Bruce. 1985 *Schismatrix.* Arbor House.
Strugatsky, Arkady and Boris Strugatsky. 1972 *Пикник на обочине.* Молодая гвардия. = 2014 *Roadside Picnic.* Gateway.
Varley, John. 1978 *The Persistence of Vision.* Dial Press.
Webb, Stephen. 2015 *If the Universe Is Teeming with Aliens … WHERE IS EVERYBODY?: Seventy-Five Solutions to the Fermi Paradox and the Problem of Extraterrestrial Life, 2nd ed.* Springer.
Weir, Andy. 2014 *The Martian.* Crown.

(1st draft. 2019/01/19)
(7th draft. 2020/02/26)
(The final version. 2022/10/16)

Journal of Philosophy of Life Vol.13, No.1 (January 2023):29-43

Artificial Intelligence and Contemporary Philosophy
Heidegger, Jonas, and Slime Mold

Masahiro Morioka*

1. Frame Problem

In this paper, I provide an overview of today's philosophical approaches to the problem of "intelligence" in the field of artificial intelligence by examining several important papers on phenomenology and the philosophy of biology.

There is no clear definition of artificial intelligence. Margaret T. Boden writes in her recent book *AI: Its Nature and Future* that an artificial general intelligence could have general powers of "reasoning and perception—plus language, creativity, and emotion." However, she does not forget to add that "that's easier said than done."[1]

Boden's concept of "artificial general intelligence" resembles John Searl's "strong AI," which was coined by Searl in 1980. According to Searl, while "weak AI" is a computer that can behave as if it were thinking wisely, "strong AI" is a computer that actually thinks like humans. Searl writes, "according to strong AI, … the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states."[2] The theme of strong AI was frequently discussed in the late 20th century; however, it became clear that in order for a computer to be a strong AI, it must resolve various difficult problems. The most difficult philosophical problem was the "frame problem."

The frame problem is the problem that an AI cannot *autonomously* distinguish important factors from unimportant ones when it tries to cope with something in a certain situation. This problem arises, for example, when we let AI robots operate in the real world. The frame problem was proposed by John McCarthy and Patrick J. Hayes in 1969. This is considered a philosophical problem that cannot be merely reduced to a technical problem. Boden writes in 2016 that

* Professor, School of Human Sciences, Waseda University. 2-579-15 Mikajima, Tokorozawa, Saitama 3591192 Japan. Email: http://www.lifestudies.org/feedback.html
[1] Boden (2006), p.22.
[2] Searl (1980), p.417.

"[c]laims that the notorious frame problem has been 'solved' are highly misleading,"[3] which shows that, even now, many specialists think that the frame problem has not been solved.

Although there is no consensus about the definition of the frame problem, we could say that this is a problem centered around the question of how we can make an AI memorize the "tacit knowledge" that almost all human adults can have in a given context. Imagine a waiter robot that serves meals and drinks for customers in a restaurant. This robot must learn a series of knowledge and movements necessary for serving. How much knowledge does this robot need to have to be able to adequately serve in an actual setting? First, the knowledge that "when pouring too much water in a glass, the water overflows" is necessary for serving. And the knowledge that "when we move a tray on which there is a glass, the glass also moves together with the tray" is necessary as well because, without such knowledge, the robot cannot remove the used glass and the tray simultaneously. Moreover, we must input the knowledge such that when a robot moves the glass, the liquid inside the glass also moves with it. However, we do not have to input the knowledge that the liquid never evaporates by friction heat caused by the movement of the tray because this knowledge has nothing to do with the robot's serving task.

Considering this, it becomes clear that there is an infinite amount of knowledge the robot must memorize when serving, and there is also an infinite amount of knowledge it does not have to memorize. Who can make such a list of knowledge, and how is it possible to make the robot memorize them? The reason why this happens is that, when a robot encounters a new situation that it has never experienced, it cannot autonomously judge what kind of coping would be important to itself and what kind of coping would not, and therefore it cannot adequately solve the problem it faces. It is interesting that humans seem to be able to solve this kind of problem. A high school student can serve in a restaurant without problem if we give her a basic set of simple instructions. She will carry a tray back to the kitchen with an empty glass on it without any special instructions. Even if there occurs a new situation, she will try to solve the problem by taking a flexible approach on a case-by-case basis.

Concerning this topic, Hitoshi Matsubara had an interesting discussion in 1990. He wrote, "A subject that has a limited capacity of data processing,

[3] Boden (2016), p.55.

including computers and humans, can never reach a complete solution of general frame problems, however, in everyday settings, humans seem not to be annoyed by the frame problem. Considering this, what we have to think deeply about should be the question of why humans look to be free from the frame problem."[4] This is the question of why in many cases human intelligence appears to successfully cope with unanticipated events in an open context, although humans do not have an enough capacity to completely solve the frame problem.

I believe that human intelligence has the following special characteristics as compared with machine intelligence: (1) when humans encounter unknown situations, they can make an adequate judgement using "tacit knowledge" and perform a certain action even if they do not have the complete list of knowledge necessary for performing it [*tacit knowledge*], (2) they can make an autonomous judgement about what kind of coping is truly important when they face an unknown situation and have to survive it [*importance judgement*], (3) they can choose a certain action and instantaneously perform it, violently ignoring other possible alternatives [*ignorance*]. It seems that artificial intelligence cannot have the above three characteristics. For artificial intelligence to have those characteristics, it must have the capacity to solve the frame problem we have discussed so far.

Recently, a series of stimulating approaches to the frame problem have come out of the interdisciplinary field between artificial intelligence research and contemporary philosophy. In the following chapters, I will examine two important discussions on this topic: Hubert Dreyfus's "Heideggerian AI" and the biological approaches influenced by Hans Jonas's idea of "metabolism." The former pays special attention to "Dasein" and "the body," which has a close relationship with today's phenomenology. The latter pays special attention to the form of "life" or "organism," which has a close relationship with today's philosophy of life and the theory of artificial life.

2. Heideggerian AI

Hubert Dreyfus is a Heideggerian who has long philosophically criticized artificial intelligence from the inception of AI research. Here I would like to examine his 2007 paper "Why Heideggerian AI Failed and How Fixing It Would

[4] Matsubara (1990), p.179.

Require Making It More Heideggerian."

Dreyfus also argues that the reason why artificial intelligence faces the frame problem is that it does not understand what kind of knowledge is important to itself in a given situation. An event or an object has meaning only when it is placed in a concrete situation.

However, the traditional AI research has presupposed the Cartesian model that our mind puts value onto the world that is made up of meaningless objects and events. Dreyfus stresses that this kind of research will never make artificial intelligence human intelligence or solve the frame problem. He pays special attention to Martin Heidegger's distinction between "Vorhandenheit" (presence-at-hand) and "Zuhandenheit" (readiness-to-hand) in the book *Sein und Zeit*.[5] For example, seen from the Cartesian perspective, a hammer in front of me appears as an objectified tool in the state of presence-at-hand. On the contrary, if that hammer appears as an already-encountered intimate tool that is interwoven in the web of meaning, which is made up of the apparent and hidden relationship between the hammer and the person (me) who uses it, we can say that the hammer appears to me as an intimate tool in the state of readiness-to-hand. The traditional artificial intelligence lacks the capacity to understand this kind of readiness-to-hand. While every person can understand that when she exists in the world she is always immersed in this kind of web of meaning, traditional AI could not understand it.[6]

Dreyfus argues that, for the traditional AI to have the capacity to solve the frame problem and become a true artificial intelligence, it must become the "Heideggerian AI" that could implement the dimension of readiness-to-hand. He examines several studies by AI researchers in that direction, but he concludes that none of those have realized Heideggerian AI.

First, he examines the robots of Rodney Brooks. Brooks is the inventor of "subsumption architecture," an insect-like hierarchical and dispersed system that is now used in the vacuum cleaner robot Roomba. Brooks's robot moves itself by "continually referring to its sensors rather than to an internal world model."[7] However, his robots "respond only to *fixed features* of the environment, not to context or changing significance,"[8] so we must say that his robot does not have

[5] Section 18 and others.
[6] Dreyfus (2007), pp.248, 251.
[7] Dreyfus (2007), p.249.
[8] P.250. Italic original.

the capacity to solve the frame problem.[9]

Next, he examines Phil Agre and David Chapman's program *Pengi*, which developed into *Pengo*, a video game in which the avatar of a human player and Penguin characters throw snowballs to each other on the screen. According to Agre, when they programmed this game, they referred to Heidegger's *Sein und Zeit* and introduced the concept of "deictic intentionality" in the game. Deictic intentionality does not point to a particular object but to "a role that an object might play in a certain time-extended pattern of interaction between an agent and its environment."[10] This game has come to be able to treat the object that the agent reacts to as a function. [11]

Dreyfus is critical of Agre and Chapman's approach. For example, when I put my hand on the door knob to leave the room, I am not experiencing the door as a mere door. In such a situation, I am pushed toward the possibility of going outside the room through this door, and the door solicits me to go outside through it. Agre and Chapman's artificial intelligence did not program this kind of experience in which the agent is solicited by affordance activated in a given situation. This shows that they ended up with objectifying the functions they introduced and the importance of the situation for an agent. Dreyfus says that, in this sense, Agre and Chapman's artificial intelligence did not have the capacity to solve the frame problem.[12]

Lastly, Dreyfus talks about Michael Wheeler's theory. Wheeler writes in his 2005 book *Reconstructing the Cognitive World* that the embodied-embedded cognitive science that has been applied to artificial intelligence research resembles Heidegger's philosophy. But Dreyfus criticizes him, saying that he also looks in the wrong place when considering this subject. Dreyfus's point is as follows. Although Wheeler insists that such embodied-embedded artificial intelligence models are considered to be Heideggerian, it still remains inside the Cartesian model in which the events in the outer world are represented onto (the sensors of) artificial intelligence, and the AI's problem solving is performed based on this representation. However, this representation model itself is the problem. We cannot fully understand human intelligence by this Cartesian model. Dreyfus argues that Heidegger considers Dasein as "being-in-the-world," and there is no

[9] PP.249-250.
[10] P.252.
[11] PP.251-252.
[12] P.253.

room for representations there. Dreyfus writes that "*being-in-the-world* is more basic than *thinking* and solving problems, it is not representational at all."[13]

When a person tries to solve problems, the boundary between that person and her tools disappears. That person has already lived inside the world, and for skilled copers, they "are not minds at all but *one with the world*."[14] In the most basic sense, we are "absorbed copers."[15] It is very hard to clearly say whether the absorbed problem solving is performed inside oneself or in the world because the distinction of inside and outside is not an easy thing to do.

The basis of Heideggerian AI should be Dasein's being-in-the-world. An artificial intelligence should be Dasein, and its way of existing should be being-in-the-world. An artificial intelligence that does not have this mode of existence should not be called "Heideggerian," and therefore it cannot have the capacity to solve the frame problem.

In the case of humans, they can improve their skills of coping with important changes in the world by their embodied capacity of problem solving. For example, when we are in a room, we usually ignore many changes therein, but if the temperature goes extremely high, the windows of the room solicit us to open them, and we react to that solicitation and open the windows. In this case, the problem solving is made by reacting to the affordance of the environment. Dreyfus writes about the reason why humans can solve the frame problem as follows. "In general, given our experience in the world, whenever there is a change in the current context we respond to it only if in the past it has turned out to be significant, and when we sense a significant change we treat everything else as unchanged except what our familiarity with the world suggests might also have changed and so needs to be checked out. Thus, a local version of the frame problem does not arise."[16] In the case of humans, "our familiarity with the world" is always activated tacitly in our cognition, and what is important to us is automatically distinguished from what is not important to us. This function deters the frame problem from arising.

When we must change the context ourselves, the frame problem again emerges. When can we recognize the fact that the problems existing in the peripheral area come to the center of our problem-solving tasks? Dreyfus says,

[13] P.254. Italics original.
[14] P.255. Italics original.
[15] P.255.
[16] P.263.

referring to Merleau-Ponty, that such a recognition is caused by "summons" from the affordance.[17] In essence, when something very important to us happens, we can recognize it by solicitations or summons made by the world we live in, and without accepting this kind of model we can never solve the frame problem. Dreyfus concludes that for artificial intelligence to acquire such capacity, "we would have to include in our program a model of a body very much like ours with our needs, desires, pleasures, pains, ways of moving, cultural background, etc."[18]

It seems that Dreyfus's Heideggerian AI should have a human-like "body" and live in that body from the inside. However, is it possible for current AI robots made up of silicon chips, metals, and plastic to satisfy such high requirements? In the next chapter, we examine biological approaches, which are completely different from Heideggerian AI.

3. Artificial Intelligence and Metabolism

There is a group of researchers who think that for artificial intelligence to have the capacity to solve the frame problem, it should be a kind of "organism," or a life form, before it can acquire a human-like body. When facing fatal difficulties, organisms try to survive by using every possible measure. Organisms have such innate capacities. Those researchers believe that these innate capacities that organisms have for survival must be the foundation needed for the resolution of the frame problem.

Hans Jonas, who was once a disciple of Heidegger, stressed the importance of the concept of "metabolism" in the field of philosophy of biology, and this concept has made a huge influence on the above approaches. Jonas published the book *The Phenomenon of Life* in 1966 (and its German edition *Organismus und Freiheit* in 1973) and established an original philosophy of biology. He thinks that "freedom" came into existence when ancient microbes with cell membranes emerged on the earth. These microbes take in nutrition through the membrane and excrete waste out through the membrane. By this kind of continuous intake and emission of tiny particles through the membrane, the microbes can maintain their lives. As time passes, all the materials forming the cell are replaced. Nevertheless, the living cell maintains its identity on a different dimension. Jonas sees here the liberation of life from the dimension of matter. This liberation is, Jonas thinks, the

[17] P.264.
[18] P.265.

"freedom" the form of life acquires.

On the other hand, life is bound by the replacement process of the tiny particles through the membrane. If this replacement process stops, life is destined to disappear because it is by this replacement process that life can maintain itself. In this sense, life depends on matter. Jonas call this kind of freedom "dependent freedom" or "impoverished freedom (bedürftige Freiheit)."[19] Life's survival is always threatened by this potential risk. Life is destined to survive by performing the continuous replacement of its contents. If life neglects efforts to replace materials, it will face its own death. Life is essentially fragile because without continuous efforts to survive, it will soon perish.

When Jonas was thinking about the above idea, he was not imagining artificial intelligence. His thoughts on life and freedom were discussed only within a small circle of philosophers of biology at that time. However, after his death in 1993, Jonas's philosophy soon began to be discussed outside the field of biology.

One of the philosophers who shed a strong light on Jonas was Francisco J. Varela, who advocated the concepts of "autopoiesis" in the field of biology and "enactivism" in the field of phenomenology and artificial intelligence. In the seminal paper "Life after Kant" written by Andreas Weber and Varela published in 2002 (Varela's posthumous publication), they try to connect Varela's autopoiesis with Jonas's metabolism. They write that "autopoiesis is the necessary empirical ground for Jonas's theory of value" [20] and that these two ideas (autopoiesis and metabolism) are "not only contemporaneous but fully *complimentary*. Both seek a hermeneutics of the living, that is, to understand from the inside the purpose and sense of the living."[21] In both theories, the key terms were the membrane of a cell and its metabolism. Jonas and Varela tried to think that a single cell that has a membrane contains "intrinsic teleology" and this cellular organism has "a *basic* purpose in the maintenance of its own identity, an affirmation of life." [22] Varela's attention on Jonas's philosophy of biology, especially his emphasis on metabolism, made a huge impact on some of the researchers of artificial intelligence.

Tom Froese and Tom Ziemke's paper "Enactive Artificial Intelligence: Investigating the Systemic Organization of Life and Mind," published in 2008, is

[19] Jonas (1973), S.150.
[20] Weber and Varela (2002), p.120.
[21] P.116.
[22] P.117.

an endeavor to develop Jonas's idea of metabolism in the field of artificial intelligence.

Froese and Ziemke interpret the frame problem as follows. It is the problem of "how it is possible to design an artificial system in such a manner that relevant features of the world actually show up as significant from the perspective of that system itself rather than only in the perspective of the human designer or observer."[23] They refer to Dreyfus's paper and stress that the frame problem has not been resolved, and they go on to say that contributions from phenomenology and theoretical biology are necessary for the solution of this problem.

Froese and Ziemke say that in recent years an "embodied turn" occurred in cognitive science. However, we still do not know how to teach an AI to understand important problems for itself "in an autonomous manner." They focus attention on Jonas's philosophy of biology. They write that "the existence of what could be described by an external observer as "goal-directed" behavior does not necessarily entail that the system under study itself has those goals – they could be *extrinsic* (i.e., externally imposed) rather than *intrinsic* (i.e., internally generated)…"[24] If an AI robot has its own "goals," they should be generated from inside the robot spontaneously. They argue that the question we should ask would be what kind of body the robot must have for it to accomplish such a task.

Froese and Ziemke refer to Jonas's idea of metabolism and discuss the difference between an artificial system and a living system. The mode of existence of an artificial system is "being by being." An artificial system can act, but this action is not necessarily done for its own survival. On the contrary, the mode of existence of a living system is "being by doing." A living system must engage in certain "self-constituting operations," that is, the continuous replacement of tiny materials through the membrane of the cell. If a living system stops the replacement actions, it will die. It disappears from the world. Doing or acting is necessary for a living system, but not so for an artificial system. This is the crucial difference between an artificial system and a living system, and this is exactly what Jonas wanted to stress by the words "dependent freedom." Jonas was discussing this topic against the backdrop of cybernetics and the general systems theory in the 1960s. Froese and Ziemke revived Jonas's idea in the age of artificial intelligence in the 21st century.

It is very difficult to make a metabolizing artificial intelligence. But they argue

[23] Froese and Ziemke (2008), p.467.
[24] P.472. Italics original.

that the fundamental reason why AI cannot solve the frame problem lies in the fact that AI does not have the biological way of being, "being by doing." For example, even if we switch off an artificial intelligence, and after that we switch it on, it will continue to operate without any special problems. However, if it is a lifeform, once it dies, it will never live again.[25] This sense of urgency that when it dies everything is over characterizes the lifeform's way of being. They argue that here lies the door to the solution of the frame problem.

They say that we should pay attention to the fact that a lifeform actively generates and sustains "the systemic identity under precarious conditions."[26] They call this mode of being "constitutive autonomy" following Varela's naming. They say that "constitutively autonomous systems bring forth their own identity and domain of interactions, and thereby constitute their own 'problems to be solved' according to their particular affordances for action."[27] They make a theoretical analysis of artificial intelligence with constitutive autonomy and try to find a possible combination between artificial intelligence and artificial life.

First, they point out the possibility of a "microbe-robot symbiosis."[28] For example, if we can reflect the state of microbes that is incorporated into a robot onto the robot's controller, the autonomous movement of the microbes could be inscribed onto the intelligence of the robot on a real-time basis. They also argue that by incorporating the principle of tessellation automaton into a robot, we might use their characteristic that although the production principle is not intelligent, the outcome looks intelligent when observed from the outside.[29] They stress that such a system has not been created by anyone and that "[i]n order to develop a better theory of the biological roots of intentional agency we first need to gain a better understanding of bacterium-level intelligence. Only by returning to the beginnings of life itself do we stand any chance of establishing a properly grounded theory of intentional agency and cognition."[30]

It seems to me that their argument that to develop a spontaneous artificial intelligence we have to go back to the bacteria level is stimulating and reasonable. Margaret A. Boden also stresses the importance of metabolism by citing Hans Jonas and concludes that if metabolism is the necessary condition for mind, strong

[25] P.485.
[26] P.480.
[27] P.481.
[28] P.492.
[29] P.494.
[30] P.495.

AI should be impossible because metabolism "can be *modeled* by computers, but not *instantiated* by them."[31] Jonas's metabolism model might be the deepest key for understanding artificial intelligence.

4. Slime Mold and Biocomputer

The endeavor to investigate intelligence by going back to the bacteria level has already begun. Among them, the slime mold computer, which has been studied by Toshiyuki Nakagaki and Ryo Kobayashi, is particularly worth mentioning. They discovered in 2000 that when putting food at two separate places on a small maze made of glass on the surface of which they have spread out starving slime mold, the slime mold slowly transforms its whole cell to make the shortest route between the two places.[32] The slime mold limbs that are on a dead-end route start to retreat from it, join the main route that is connected with the food, and help thicken the cross-section of the main route made by slime mold. In this way, slime mold makes a kind of calculation by itself, discovers the shortest route between the two places, and modifies its body into the most efficient shape for that route. This action performed by starving slime mold eloquently shows the fundamental mode that lifeforms seek to maintain their existence in a "precarious" situation.

In their 2011 paper "Performance of Information Processing in a Primitive Organism of True Slime Mold" (in Japanese), Nakagaki and Kobayashi argue that this kind of action for survival by slime mold is made by the "calculation" performed by the slime mold itself.[33] That is to say, the action for survival emerges inherently and spontaneously inside the slime mold, calculations searching for the most adequate solution are performed, and the slime mold transforms itself in accordance with the solution. This can be called a biological calculator, and I presume that the frame problem might be solved in this slime mold case. If we give the slime mold a new difficult task and track it down, it would certainly rethink its strategy for the new condition and try to transform its body toward a new adequate solution. The slime mold seems to have the capacity to continuously adapt itself to unknown changes in the environment by transforming its own body in an emergent way.

---

[31] P.144. Italics original.
[32] Nakagaki, T. et al. (2000)
[33] Nakagaki and Kobayashi (2011), p.483.

Nakagaki and Kobayashi made a mathematical simulation model for tracing the movement of slime mold (physarum solver) and investigated its behavior. As a result, they discovered that the calculation the slime mold makes is not accurate and perfect but "rough and speedy." They argue that such a "rough and speedy" calculation is a "noteworthy characteristic of biological computation."[34] Although they do not mention this, I believe that this characteristic of biological computation might be the key to creating human-like intelligence – intelligence that, when encountering a new environment, can speedily judge which factor is most important to it and act violently, ignoring other non-important factors. This kind of intelligence is necessary for solving the frame problem.

Kobayashi also writes in his 2015 paper "Autonomous Decentralized Control Found in Creatures: From Slime Mold to Robot" (in Japanese) that while most robots can move correctly in the anticipated environments, animals can move in a tough manner even if they encounter unknown environments. He argues that animals have the capacity to solve the frame problem,[35] and these animals include not only mammals and insects but also slime mold. Kobayashi argues that insects and slime mold can take spontaneous decentralized control over their bodies. This suggests that to solve the frame problem, the development of a spontaneously decentralized bodily system would be better than that of centralized control system like a central nervous system. Kobayashi says that his snake-like self-moving robot might have such a decentralized system, and he proposes the development of the control system that makes the environment its "friend."

These studies suggest that the inherent and spontaneous solution of the frame problem made by humans is performed not by the human central nervous system but by the decentralized control system located at every part of the body that is free from the control of the central nervous system. However, it should be noted that Brooks's subsumption architecture has not succeeded in solving it.

Froese and Ziemke's "microbe-robot symbiosis" might be a possible answer. They propose to insert a colony of microbes into the body of a robot, but isn't there another possibility in the opposite direction – the possibility to insert artificial objects such as super-micro artificial intelligence, super-micro robots, or the fragments of artificially structured DNA or RNA into the cells of microbes? It might be possible to create the symbiotic systems of a group of such super-micro artificial objects and microbes.

[34] P.491.
[35] Kobayashi (2015), p.236.

Take the example of slime mold. We might give slime mold the capacity of calculation that is enhanced by super-micro robots, super-microprocessors, or artificially-made nucleic acids. Such artificially enhanced slime mold could not only solve the shortest path problem inherently and spontaneously, but it could also solve more difficult tasks by calculation. In such a case, we could suppose that this slime mold must have the capacity to discover the problem that is important for its own survival and to solve that problem by spontaneous calculation. As slime mold as an organism is considered to have the capacity to solve the frame problem, slime mold with the capacity of calculation that is artificially enhanced should also have the capacity to solve the frame problem. Artificially enhanced slime mold should be considered a kind of biocomputer. In the context of computers, biocomputers are the key to the solution of the frame problem. This is the provisional conclusion of this paper.

One philosophical problem emerging from our discussion is, if the frame problem is to be solved by an organisms's spontaneous decentralized control, then the frame problem could be solved without the realization of Heideggerian AI proposed by Dreyfus, which exists in the mode of being-in-the-world. The frame problem might not be the problem at the level of the central nervous system that executes symbol manipulations but the problem at the level of metabolism-based, spontaneously decentralized control systems. Since Dreyfus would have presupposed the control by the central nervous system, his idea could have been completely wrong. Some people say that the recent development of deep learning will perhaps succeed in solving the frame problem, but the capability of deep learning is still not clear. The above discussions depend on how we understand the essence of the frame problem. This is the question philosophers should tackle head-on.

As seen above, we have tried to give an overview of the connection between the frame problem, Heideggerian AI, metabolism-based AI, the spontaneously decentralized control system by slime mold, and a future possible solution of the frame problem by biocomputers. We find there several stimulating themes for contemporary philosophy. Researchers of philosophy will take interest in the fact that the names of Heidegger and one of his great disciples, Jonas, appear in our discussion of the frame problem. I am not an AI research specialist, so if there are any misleading expressions or incorrect uses of technical scientific words in this paper, please let me know.

There is tremendous risk in research on making artificially-enhanced slime

mold. We must prevent the uncontrolled runaway of artificially-enhanced slime mold because this research intends to give slime mold high-level calculation capacities. If they are emitted into the environment, they might cause devastating damage to humans and ecosystems, hence the research ought to be carried out at the highest biosafety level in facilities that have the capacity for physical containment stipulated by the Cartagena Protocol. In the first place, we cannot imagine how slime mold would behave when its capacity of calculation is enhanced. There might be the risk that artificially-enhanced slime mold with high-level intelligence could proliferate on a huge scale and cover the whole earth searching for food. In the case of toxic microbes, research on giving them high-level calculation capacities should not be allowed.

This kind of research can also be seen as enhancement research using artificial objects with microbes as their targets. Therefore, this topic is connected with bioethical discussions on enhancement.

While artificial intelligence has supported biotechnological research in many ways, in the future there will appear a completely different situation in which AI research is directly combined with the manipulation of organisms in the field of biotechnology. We must have an intensive and interdisciplinary discussion before it becomes a reality. We can conclude that the gulfs between AI research, biology, and philosophy have become much shallower than before.

* This paper is a translation of my Japanese paper that was published under the same title in 『哲学』Vol.70, (2019): 51-68.
* This work was supported by JSPS KAKENHI Grant Numbers 17K02185, 17H00828, 16H03337.

References

Boden, Margaret A. (2016). *AI: Its Nature and Future*. Oxford University Press.
Dreyfus, Hubert L. (2007). “Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian.” *Philosophical Psychology* 20(2):247-268.
De Jesus, Paulo (2016). “Autopoietic Enactivism, Phenomenology and the Deep Continuity Between Life and Mind.” *Phenomenology and the Cognitive*

Sciences 15:265-289.
Froese, Tom, and Gallagher, Shaun (2010). “Phenomenology and Artificial Life: Toward a Technological Supplementation of Phenomenological Methodology.” *Husserl Studies* 26:83-106.
Froese, Tom, and Ziemke, Tom (2009). “Enactive artificial intelligence: Investigating the systemic organization of life and mind.” *Artificial Intelligence* 173:466-500.
Heidegger, Martin (2006). *Sein und Zeit*. Max Niemeyer Verlag.
Jonas, Hans (1973, 1977). *Das Prinzip Leben*. Suhrkamp. (*Organismus und Freiheit: Ansätze zu einer philosophischen Biologie*)。
Kiverstein, J. and Wheeler, M. (eds.) (2012). *Heidegger and Cognitive Science*. Palgrave Macmillan.
Kobayashi, Ryo (2015). 小林亮「生物に学ぶ自律分散制御：粘菌からロボットへ」『計測と制御』54(4):236-241.
Matsubara, Jin (1990). 松原仁「一般化フレーム問題の提唱」J・マッカーシー、P・J・ヘイズ『人工知能になぜ哲学が必要か』哲学書房, pp.175-245.
Nakagaki, T., Yamada, H., and Tóth, A. (2000). “Path finding by tube morphogenesis in an amoeboid organism, *Nature* 407:470.
Nakagaki, Toshiyuki, and Kobayashi, Ryo (2011). 中垣俊之・小林亮「原生生物粘菌による組合せ最適化法：物理現象として見た行動知」『人工知能学会誌』26(5):482-493.
Searle, John R. (1980). “Minds, brains, and programs.” *The Behavioral and Brain Sciences* 3:417-457.
Shimonishi, Kazeto (2015). 下西風澄「生命と意識の行為論：フランシスコ・ヴァレラのエナクティブ主義と現象学」『情報学研究』89:83-98.
Weber, A. and Varela, F. J. (2002). “Life After Kant: Natural Purposes and the Autopoietic Foundations of Biological Individuality.” *Phenomenology and the Cognitive Sciences* 1:97-125.
Wheeler, Michael (2005). *Reconstructing the Cognitive World: The Next Step*. The MIT Press.
Wheeler, Michael (2008). “Cognition in Context: Phenomenology, Situated Robotics and the Frame Problem.” *International Journal of Philosophical Studies* 16(3):323-349.

Journal of Philosophy of Life Vol.13, No.1 (January 2023):44-63

Implications of Automating Science
The Possibility of Artificial Creativity and the Future of Science

Makoto Kureha*

Introduction

Science fiction writer Chiang's (2002) short story titled "The evolution of human science" depicts a future age in which "metahumanity" acquires intelligence that outstrips human intelligence. There all original works of scientific research are generated by metahumans. Most human researchers have quit their jobs, and the rest of them are engaged in "hermeneutics," or the study that aims to interpret metahuman science and make it comprehensible to humans. The narrator of the story considers the *raison d'être* for humanity's science in this age. Although Chiang does not explain in detail what "metahumanity" is, one of its candidates is artificial intelligence (AI)—more specifically, what Searle (1980) calls "strong AI," which has the real capacity for thinking.

In reality, strong AI has not appeared yet, but AI technologies are increasingly applied to scientific research for various purposes. In 2017, *Science* issued a special issue titled "AI Transforms Science," which reported the current status and future prospects of the use of AI in various scientific areas of physical, biological, and social sciences. Moreover, some researchers aim to realize "the automation of science" (King et al. 2004), that is, to make AI systems or robots execute research tasks without human intervention. It is impossible to predict at present whether the future Chiang depicts, where AI takes humanity's place in science, will come true. At any rate, as automation proceeds, modes of scientific research as well as the state of the science community and the relationship between science and society will change dramatically.

This tendency of automating science is remarkable in multiple respects. First,

* Associate Professor, Faculty of Global and Science Studies, Yamaguchi University. 1677-1, Yoshida, Yamaguchi City, Yamaguchi 753-8541 Japan.
Email: kureha[a]yamaguchi-u.ac.jp

** This paper is an English translation (with some modification) of my writing part of the following co-authored paper: Kureha, M. & Kukita, M. 2020. 'AI and Scientific Research', in S. Inaba, M. Oya, M. Kukita, S. Narihara, M. Fukuda, & T. Watanabe (eds.), *Artificial Intelligence, Humanity and Society*, pp. 122-169, Tokyo: Keiso Shobo (written in Japanese).

it will radically change the overall state of science and technology, and so it may lead to large-scale innovation and to solving social problems. Second, since science is one of the most distinctive human activities, the idea of automating it prompts us to reconsider the aspects of our humanity itself. One of these aspects is *human creativity*. Renowned physicist Dyson (1988) states, "Science is at its most creative when it can see a world in a grain of sand and a heaven in a wild flower. Heavy hardware and big machines are also a part of the science, but not the most important part" (p. 158). As clearly shown in this passage, the view that human creativity constitutes a central value of science is widespread. However, as I will argue, the automation of science might undermine this creative character.

In this article, I examine the prospect of automating science, focusing on two questions. First, can AI make creative discoveries? Second, what implications may the automation of science have on science and society? It will be shown that the attempt to address these questions leads to a reconsideration of philosophical questions concerning the nature and values of science. I will conclude that the prospect of success in automating creative discovery is not bright at present. Nevertheless, I will also argue, we should anticipate that the automation of science will have many serious implications for science and society. Therefore, we need to specify desirable ways of introducing AI technologies into science and devise measures against the demerits of automating science.

In Section 1, I will introduce the current state and future goals of the automation of science. In Section 2, I will examine whether AI can make creative discoveries. Then, in Section 3, I will consider what implications the automation of science has for the science community and wider society.

1. What Is the Automation of Science?

1.1 The current state of applications of AI in science

Scientific research has been a central target of AI research since its early days. Many studies conducted on this topic constitute a research area known as "machine discovery" or "computational discovery." Two famous examples of their achievements are AM (Lenat 1977) and BACON (Langley et al. 1987). The former is a program that generate programs expressing mathematical concepts; for example, it succeeds in producing Goldbach's Conjecture. The latter is a program that picks up invariants from data; for example, it succeeds in deriving

Kepler's First Law from data concerning planetary orbits. Thus, early studies on machine discovery had aimed at simulating past discoveries made by human scientists. There also are attempts to make computers aid human researchers to discover new knowledge (see Langley 2000), but this approach had not become major until recently.

Now, AI systems are increasingly applied to scientific research more practically. AI systems based on machine learning methods are used for various scientific purposes such as the detection of new particles in physics, classification of celestial bodies from image data in astronomy, prediction of efficient ways of chemical synthesis in chemistry, identification of the genome for a psychiatric disorder in biology, and analysis of the mood of masses from social medias in psychology (see Science News Staff 2017). Moreover, there are attempts to make AI systems survey literature by applying text mining methods (see Stix 2005) as well as attempts to make AI-based robots run experiments (see King et al. 2004; King 2010). Thus, AI is becoming a standard tool in diverse areas of physical, biological, and social sciences. (Although AI is becoming also widely used in applied science areas such as medicine and pharmacy, I focus on its use in basic science areas in this article.)

This trend of applying AI to science goes toward the *automation of science*. Researchers enthusiastic about AI aim to develop AI systems that automatically execute the whole tasks of scientific research (such as exploring the literature, designing and running experiments, interpreting data, writing and reviewing papers, and so on). One of the most remarkable achievements in this attempt was done by Robot Scientist "Adam," which King's research group developed (King et al. 2004; King et al. 2009; King 2010). Adam is a robot that generates hypotheses, derives their consequences, and tests them automatically by itself. By doing this, it discovered genes encoding an enzyme required for the growth of yeast in the area of functional genomics. Furthermore, as a future challenge, Hiroaki Kitano, the Director of Sony Computer Science Laboratories, has set a goal "to develop an AI system that can make major scientific discoveries in biomedical sciences and that is worthy of a Nobel Prize and far beyond" (Kitano 2016b, p. 39). To denote this, a new mode of scientific research characterized by automation, some researchers use the term "AI-driven science."[1]

As matters now stand, AI systems are applied to science only to support

[1] To my knowledge, the term "AI-driven science" was coined by Koichi Takahashi, a Japanese biologist working at the Institute of Physical and Chemical Research in Japan.

human researchers. Newspapers sometimes report that *AI has discovered* something, but this is not a precise description of the state of affairs. It is more accurate to say that *humans have discovered something by using AI*, since AI systems at present do not execute tasks *with their own intentions*: they are nothing but non-autonomous tools, as we will see in Section 2. However, there are some ambitious researchers who aim to develop truly autonomous AI systems for scientific research. For example, Kitano mentions "systems that acquire knowledge autonomously and make discoveries continuously" (Kitano 2016a, p. 84, my translation) as the ultimate end of his challenge. Thus, AI-driven science is potentially a long range endeavor.

1.2 Impacts of automating science and features of AI as a scientific technology

Leaders of AI-driven science, such as King and Kitano, expect that the automation of science will bring enormous benefits. They typically mention two types of its merits: practical and intellectual. I will explain them in turn.

As for the practical merits, King and his co-authors say,

> We consider this trend to increased automation of science to be both inevitable and desirable. It is inevitable because it will be required to deal with the challenges of science in the twenty-first century. It is also desirable because it frees scientists to make the high-level creative leaps at which they excel. (King et al. 2004, p. 251)

One of what King et al. call "the challenges of science in the twenty-first century" above is the exponential increase in amount of data to be handled. Situations such as the rapid growth of gene databases and the emergence of petabyte-scale astronomical data through sky surveys make it impossible for researchers to analyze data by themselves. Given this, applying AI can be considered inevitable for science to progress further. Another practical merit King et al. mention, i.e., freeing scientists from non-creative tasks so that they can concentrate on creative ones, is also important, in particular under the present situation in which post-doctoral researchers and graduate students, sometimes called "pipetting slaves," devote long hours in the laboratory to dull and repetitive works such as cleaning pipettes. Furthermore, ambitious leaders of AI-driven science claim that the automation of science will dramatically increase research productivity, and

contribute to solving social problems. For example, Kitano mentions the practical merits of his challenge as follows:

> I anticipate that, in the near future, AI systems will make a succession of discoveries that have immediate medical implications, saving millions of lives, and totally changing the fate of the human race. (Kitano 2016b, p. 39)

The other merits of automating science its proponents mention are intellectual: It sheds light on *what science is* (e.g., King 2010; Kitano 2016a). We can see this clearly by referring to what is called the "N = 1 problem" in biology. Biologists' attempts to identify the essence of life are prevented by the fact that all the samples of life available to them belong to a single lineage of terrestrial lives. Even if biologists identify certain features of life common to all the available samples (such as being constituted by cells), they cannot tell whether they are essences of life or merely contingent features of terrestrial lives. Therefore, they put their hope in astrobiology and A-Life to obtain samples of other life forms. The same is true of science: The only sample of science available to us so far has been humanity's science. This fact constitutes an obstacle to discriminate the essential features of science from its contingent ones resulting from constraints by human cognitive and other limitations. Therefore, it is helpful to develop AI capable of scientific research. If AI acquires the intelligence required for scientific research, it will probably develop sciences that are significantly different from humanity's one. The emergence of these alien sciences will contribute to our understanding of what science is.

Although I am somehow sceptic about optimistic discourses by leaders of AI-driven science, I nevertheless agree with them on the prediction that it will radically change the overall state of science and technology. Underlying its potential power to cause such changes are some unique features of AI as a scientific technology. Of course, various technological devices, such as telescopes, microscopes, and computers, have been used in scientific research throughout the history of science. However, AI has some features that are not found in other scientific technologies. I will mention just two of these features.

First, AI systems can be used to carry out research tasks automatically and make human intervention unnecessary. This is the very feature that underlies the practical merits of the automation of science, such as increase in the productivity of research and liberation of people from non-creative research tasks.

Second, AI systems may generate knowledge beyond human understanding. Since older devices presuppose manipulation by humans, they are designed so that their behaviors remain within the range of human understanding. In contrast, AI systems do not necessarily have this limitation. Indeed, they may even have the potential to bring forth kinds of "alien science," a system of knowledge that understands the world in certain ways different from ours. Kevin Kelly, the founding editor of *Wired* magazine, makes this point by saying,

> AI could just as well stand for "alien intelligence." (…) An AI will think about science like an alien, vastly different than any human scientist, thereby provoking us humans to think about science differently. (Kelly 2016, p. 48).

However, the automation of science may bring significant problems and undesirable effects as well as the benefits described above. As an example of these problems, it blurs *who* does the research at the price of making human intervention unnecessary. This will pose difficult problems concerning credit and responsibility for scientific research. As another example, researchers who use AI in their research face a "black box" problem: What they discover by using AI systems might go beyond their understanding. This can pose difficult questions as to whether such findings qualify as "knowledge" (or "scientific knowledge") and how much epistemic value they have.

Further discussion of these issues is beyond the scope of this article. Instead, I concentrate henceforth on the issues concerning *creativity*.

2. The Possibility of Artificial Creativity

Leaders of AI-driven science focus especially on scientific *discovery*. Since discovery is the process of yielding new knowledge, and since it is thought to be one of the most creative phases of scientific research, its automation will have great impacts both practically and intellectually. However, because scientific discovery requires creativity, it can be difficult for AI systems (and robots) to execute it automatically. This section examines whether AI can make creative discoveries, and, if possible, how they can be automated.

2.1 What is creativity?

Let us clarify what "creativity" means by referring to Boden's analysis of the concept. Boden (2004) defines creativity as the capacity to create ideas or artifacts that are *new*, *surprising*, and *valuable*.[2] Although she defines it as a kind of capacity, activities of exercising it and products created by such activities can be called creative, too. Typically cited examples of such activities are artistic creation, technological invention, and scientific discovery.

In addition to a definition, Boden proposes two taxonomies of creativity. The first distinguishes two sorts of creativity in terms of the kinds of "newness": One is "*psychological creativity*," which concerns a creative product that is new to the individual who produces it; and the other is "*historical creativity*," which concerns a creative product that is new to humanity (or to a certain community). Of course, scientific creativity is classified as the latter. Hence, for scientific AI to be creative, it must avail itself of some method to generate ideas that go beyond past knowledge of the overall science community.

Boden's second taxonomy distinguishes the following three types of creativity in terms of the ways surprising ideas are generated: (1) "*combinatorial creativity*," which is achieved by combining familiar ideas in some unfamiliar way; (2) "*exploratory creativity*," which is achieved by exploring a "conceptual space" following some style; and (3) "*transformational creativity*," which is achieved by transforming a pre-existing conceptual space. Here, Boden uses the term "conceptual space" to refer to a space that consists of all possible ideas about some topic. In the context of scientific discovery, all possible hypotheses to a question constitute the conceptual space concerning the question, for example. Section 2.3 will examine approaches to artificial creativity by referring to these three kinds of creativity.

2.2 Can artificial creativity be realized?

There has been a lively debate concerning whether AI can be creative. A common argument against artificial creativity is called *Lady Lovelace's objection*.

[2] Though the condition of surprise is often dropped from the requirements for creativity in psychological literature, I think it captures an important aspect of creative activities such as scientific research. For example, Dyson says, "For science to be great it must involve *surprises*, it must bring discoveries of things nobody had expected or imagined." (Dyson 1988, p. 165, emphasis added.)

It is so called since its most classical version is found in Ada Lovelace's comment on Charles Babbage's analytic engine. According to this objection, machines (such as computers and robots controlled by them) cannot be creative because they can do only what they are programed to do. If a machine creates something, the relevant creativity should be attributed to its programmer rather than the machine.

However, also famous is Turing's (1950) reply to this objection:

> This may be parried for a moment with the saw, "There is nothing new under the sun." Who can be certain that "original work" that he has done was not simply the growth of the seed planted in him by teaching, or the effect of following well-known general principles. (p. 450)

Here, Turing points out that even human creativity is not *creation ex nihilo*, and argues that being programmed is on a par with being taught or informed. The moral we can learn from his reply is that to think artificial creativity cannot be realized in principle might be tantamount to *mystifying* human creativity.

Opponents of artificial creativity would not be convinced, yet. Rather, they might claim that Lady Lovelace's objection can be refined by focusing on *autonomy* (see Boden 2014). The revised argument goes as follows: Machines cannot be creative, since creativity requires autonomy, and machines that can do only what they are programed to do necessarily lack it. In this context the term "autonomy" is used in the philosopher's sense in which it denotes the ability to set one's own goals freely (for example, choosing the tasks). It should be distinguished from mere "automaticity" (or "autonomy" in the roboticists' sense), which denotes the ability to operate without being controlled by other agents. It is important to note here that agents can be automatic without being autonomous: There can be agents that execute tasks set by others without being controlled by others.

The claim that creativity requires autonomy seems plausible to a certain degree (though not to the full extent, since we do not understand even how autonomy is realized in humans). More controversial is the claim that machines *necessarily* lack autonomy. To defend this claim, some theorists argue that autonomy requires life. However, life is a complex phenomenon that consists of many features of living systems (such as reproduction, metabolism, self-organization, evolution, development, etc.), and it is not clear which feature is

relevant to autonomy and why it is relevant. Moreover, it is not obvious that AI systems necessarily lack the relevant feature: Studies in A-Life have shown that many interesting features of living systems can be realized by artificial systems. Anyway, it is without controversy that no AI systems applied to scientific research today lack autonomy, and there seems to be little prospect that truly autonomous AI scientists will appear in the near future. Thus, the revised version of Lady Lovelace's objection seems reasonable for the time being.

Nevertheless, we should be careful to identify what this consideration means. According to the idea, even when an AI system devoid of autonomy produces something comparable to products of human creative activities, it does not qualify as exercising true creativity: it turns out instead that the AI system only simulates creativity. This view, though, does not exclude the possibility that even such a non-creative AI system can exhibit as good performance as truly creative agents do. Moreover, at present, proponents of AI-driven science do not necessarily aim to develop truly autonomous, strong AI scientists, but rather to devise AI systems as mere useful tools for scientific research. So, if AI systems can perform research tasks as well as human researchers, it can be argued that whether they are truly creative or not in themselves is not important for the immediate goals of AI-driven science. Furthermore, regardless of whether such systems are deemed truly creative or not, we can still wonder how to develop such high-performance AI systems. Therefore, let us put aside the issue of artificial autonomy, and consider next how creative discoveries can be achieved through the use of AI. (Henceforth, for simplification, I use the term "creativity" to refer not only to the capacity exercised by truly autonomous agents, but also to the capacity exhibited by non-autonomous systems that generate ideas in as high-performance as them.)

2.3 How can artificial creativity be realized?

In this section we examine three ways to achieve creative discoveries through the use of AI, which correspond to three kinds of creativity Boden identifies (i.e., explorative, transformative, and combinatory creativity).

The combinatorial approach

A major view of creative activity (including scientific discovery) states that it consists of an *unfamiliar combination of familiar ideas*. This view has been put

forth by various theorists (e.g., Poincaré 1908; Asimov 1959) and supported in the psychology of science (e.g., Simonton 2004). Let us call it the "combinatorial view." In the domain of science, a famous example of discoveries accomplished in this manner is Darwin's (and Wallace's) discovery of evolution by natural selection. It is said that Darwin came to his theory of evolution by combining multiple ideas such as the idea of "overpopulation and weeding out" which he drew from Malthus, the idea of selective breeding of animals and plants, and the idea concerning how species diverge, which he confirmed during the voyage of the Beagle (see Asimov 1959; Bowler 1983).

Can we develop AI systems that make creative discoveries by adopting the combinatorial view? Combination can be accomplished by AI and robots. However, as Boden (2004) states, most of the resulting products are not valuable. Therefore, a certain method of selecting valuable ideas is required. Taking this point into account, once Poincaré (1908) denied that machines can make mathematical discoveries, since the rules concerning selecting valuable ideas are too subtle for machines to apply. Though his argument seems to be question-begging, it is persuasive that the combinatorial approach faces a dilemma: On the one hand, to conceive of a *surprising* idea, unexpectedness of the combination is important; on the other hand, as Boden (2016) suggests, some kind of relevance of combined ideas is important for conceiving of a *valuable* idea. Unexpected combinations of relevant ideas are rarely found. Especially, it seems quite difficult to conceive of combinations of relevant ideas each of which belongs to a different knowledge domain, as we find in Darwin's theory of evolution.

What capacities or mechanisms do the trick in the case of humans? An often-cited candidate is *analogy*, or the kind of inference that derives knowledge of unfamiliar problems or situations from knowledge of familiar problems or situations. According to cognitive scientists Holyoak and Thagard (1995), it involves the mental act that associates knowledge of some familiar problem or situation (a "base") and knowledge of unfamiliar one (a "target"), and finds out some structural similarity between them. Holyoak and Thagard also point out that analogy plays an important role as a "mental mechanism for combining and recombining ideas in novel ways" (ibid., p. 13) in creative thinking including scientific discovery, and they mention many examples to demonstrate it. For example, acts of analogy that find similarities between the struggle for life in humans and that in animals and plants and between selective breeding (artificial selection) and natural selection enabled Darwin to build his theory of evolution.

However, there is a finding suggesting that analogy plays only a *limited* role in scientific discovery. Psychologist Dunbar (1997) shows, on the basis of fieldwork in several labs, that analogy that associates pieces of knowledge from different science areas (to take an example from molecular biology, one that invokes knowledge of something other than organisms rather than knowledge of organs in the same organism or knowledge of different organisms) is rarely used in research practice, and that analogy is less frequently used for generation of hypotheses than for explanation. Given these points, it might be better to regard analogy as a means to understanding rather than to discovery.[3] At any rate, it is sure that we do not have enough knowledge of the role analogy plays in scientific discovery or of the mechanisms underlying analogy.

Another consideration is concerned with *aesthetic judgement*. Poincaré claimed that the selection of useful combinations of ideas is enabled by aesthetic sensitivity. Though he also argued that machines cannot make aesthetic judgement (and therefore he denied the possibility of artificial creativity), there is room for doubt in this regard. Thus, an interesting orientation of future research lies in examining how we could develop AI systems capable of making aesthetic judgement, as well as trying to understand the role of aesthetic judgement and the nature of aesthetic values in scientific discovery.

The exploratory approach

Another view of creativity, which is especially popular in cognitive science, states that scientific discovery consists in the *exploration of possible ideas*. Let us call this view the "exploratory view." This view is a natural consequence of fundamental assumptions of cognitive science, namely, that scientific discovery is problem solving and that problem solving is searching the problem space (i.e., the space that consists of all possible solutions) for the solution by means of heuristics (see Simon 1996). For example, according to cognitive scientist Anzai (1985), there is evidence in the literature that Watson and Click's discovery of the double helix structure of DNA and Faraday's discovery of the law of induction were accomplished in this manner.

Then, should we adopt the exploratory view to achieve creative discoveries by AI? Surely exploration can be accomplished by AI systems even much more

[3] Actually, some complication must be added to this consideration, since it seems that we cannot discover what we cannot understand. I do not pursue this point further in the present article.

effectively than by humans. Nevertheless, there is a good reason to think that producing something merely by following a mechanical procedure of exploration does not qualify as a creative act. Philosopher Novitz (1999) shows this clearly by referring to the example of Goodyear's invention of vulcanized rubber. To make rubber products heat- and cold-resistant, Goodyear had combined rubber with various substances at hand haphazardly, and it took many years to finally produce sulfur. Though no doubt Goodyear made an important discovery as a result of an admirable effort, it is difficult to say that his discovery is a creative one comparable to Newton's and Darwin's ones.

For this reason, some theorists (e.g., Gaut 2003) claim that creativity requires a certain non-mechanical factor such as "flair." However, such a claim does not qualify as satisfying unless what kinds of cognitive capacities and mechanisms constitute the "flair" is clarified. Moreover, proponents of the exploratory view would reply that, even given that the good-old fashioned "generation-test method" Goodyear adopted does not suffice to yield creative achievements, we can achieve them by adopting certain more sophisticated procedures. For example, Dawkins (1986) says, "Effective searching procedures become, when the search-space is sufficiently large, indistinguishable from true creativity" (p. 66).

Nevertheless, it seems reasonable to suspect that a discovery accomplished by exploration is not deemed radically creative as long as it is made according to some pre-existing style. What matters is the nature of the surprise evoked. According to Boden (2004), any idea resulting from exploring a pre-existing conceptual space can be described and yielded by applying pre-existing generative rules. Though it sometimes causes a surprise that something unexpected happens, we can understand that it comes under familiar patterns once it happened. In contrast, radically creative ideas are ones that could not be yielded merely by applying pre-existing generative rules: They evoke huge surprises by making possible something that had not been thought to be possible.

The transformational approach

Then, how can such a radically creative discovery be achieved? Boden claims that *transforming the conceptual space* is a means to such achievement.[4]

[4] Novitz denies Boden's identification of radical creativity with transformational creativity, arguing that radically creative ideas can be produced even when any relevant conceptual space does not exist. I concede that the transformational approach is not the *only* means to radical creativity. What matters for

Although the kinds of discoveries made in this manner are rare, Kepler's discovery of his First Law is one example. The pre-modern astronomy Ptolemy established was governed by the thought that the orbits of planets are round. Kepler departed from this way of thinking after examining Tycho Brahe's observation data concerning planetary orbits and found that planets orbit elliptically. As is well known, this discovery led to the scientific revolution in the 17th century.

The transformational approach seems to grasp better the creative character of creative discoveries than the exploratory approach does. Then, is it possible to allow AI to transform the conceptual space automatically? According to Boden (2004), transformation of a conceptual space can be achieved by deviating or modifying the constraints that shape it. For example, non-Euclidian geometrics was established by removing the fifth postulate ("parallel postulate") of Euclidian geometrics. Once, as is in this case, we specify the constraints that shape a conceptual space, we might succeed in yielding transformation of the conceptual space automatically by formulating the task of deviation or modification of it as a search problem. This is an interesting possibility. However, there is a difficulty to the approach: Most scientific problems are ill-defined, and therefore it is hard in many cases to specify the constraints in any formal procedure. Philosophers of science Bechtel and Richardson (1993) point out this feature of scientific problems by saying, "the constraints defining an adequate solution are not sharply delineated, and even the structure of the problem space itself is unclear" (p. 15). Unless we can specify the constraints, we can neither deviate nor alter them. Of course, this difficulty also confronts human researchers attempting to make scientific discoveries. Indeed, it might be the very reason that discoveries depend heavily on accidents or serendipity. As long as AI does not overcome this difficulty, any attempt to automate the process of discovery will not achieve great success.

As the discussion above demonstrates, all three approaches to scientific discovery by AI have some difficulty or uncertainty in the present situation and cannot be seen as a decisive way to make creative discoveries yet.

---

radical creativity is the effect that a novel style of thought and a novel conceptual space emerge, rather than the means to it. At the same time, I agree with Boden that the transformation of pre-existing conceptual space is, though not *essential*, an *effective* means to that effect.

2.4 Possibility of the brute force approach to artificial creativity

In this subsection, let us examine a currently proposed approach to AI-driven scientific discovery. The approach in question is the one proposed by Kitano. He puts forward "a brute force approach in which AI systems generate and verify as many hypotheses as possible" (Kitano 2016b, p. 46). This can be regarded as an extreme version of the exploratory approach. As Kitano realizes, this approach differs substantially from ones in which human researchers achieve discoveries. He states that the current state of scientific discovery, which depends on unreliable intuition and serendipity, is at the level of "cottage industry" (ibid., p. 41). Thus, he declares, "AI scientific discovery systems have the potential to drive a new revolution that leads to new frontiers of civilization" (ibid., p. 48).

Can Kitano's brute force approach accomplish his purpose to bring about innovation in the mode of research? It may be possible, and his proposal surely is worth pursuing. However, when it comes to creativity, there is a worry, namely, that the science resulting from it does not seem to exhibit creativity. This is because, as Goodyear's case suggests, a discovery accomplished by merely exploring some pre-existing conceptual space yields no huge surprise, no matter how exhaustive the exploration is. Proponents of the brute force approach would reply that, though the approach in question may not achieve something like human creativity, it will achieve a different kind of creativity. It is true that creativity can take other forms than ours; to think otherwise is to commit a sort of anthropocentric chauvinism. However, why would activities that are considered non-creative if they are carried out by humans be deemed creative when they are carried out by machines? If we are to reject anthropocentrism, we should think that whether activities are creative or not does not depend on *who* carries out them. This thought leads to exclusion of the brute force approach from the means to creative discovery. Moreover, if the brute force approach will nonetheless become prevalent, the value of science as a creative activity might be compromised.

Let us summarize this section. While there is no sound ground for the claim that AI cannot make creative discoveries in principle, we do not have enough knowledge to automate such discoveries. Indeed, we do not fully understood how humans achieve them yet. Therefore, in order to develop AI systems that make them, we should accumulate empirical knowledge concerning human creativity at the first onset. If we avoid such a steady effort and harry to automate science in

ways that resort to brute force, it could lead to a state of affairs in which some non-creative mode of scientific research would prevail.

There is no doubt that, as the use of AI becomes common in science, modes of scientific discovery will change. On this occasion, it seems fruitful, at least for the time being, to establish organizations in which AI systems and human researchers can cooperate in an appropriate way, rather than leaving the whole process of discovery to AI systems. As long as AI is introduced into science in such a manner, it will become a powerful tool for discovery due to its great capacity for exploration.

3. The Implications of Automating Science

In this section, let us consider what implications the AI-driven science may have on the scientific community and wider society. The automation of science will not only benefit scientists and other people, but also bring about undesirable effects. Although there are many kinds of worries (some of which I mentioned briefly in the end of Section 2.2), here I will examine just two of them: technological unemployment and undermining the value of science.

One of the worries about the automation of science is the threat of technological unemployment. Will AI take human researchers' jobs, as Chiang depicts? At present, many researchers would answer "no" to this question. For example, Frey and Osborne's famous report "The Future of Employment" (2013) states that jobs requiring creativity (such as those of artists and scientists) cannot easily be automated. Indeed, as we saw in Section 1, leaders of AI-driven science (e.g., King et al. 2004) often claim that the automation of science will liberate researchers from dull tasks and enable them to concentrate on creative works. We have, though, some reasons to cast doubts on such optimistic expectations, as explained below.

The prediction that truly creative AI will not appear seems plausible in the short term. However, this does not mean that human researchers' position is secure. The reason is that the place of creative tasks in the whole system of science may not remain constant. Rather, even given that tasks requiring creativity are difficult to automate, the weight of these tasks in the whole scientific research might decrease. Norman (2007) makes this point with respect to the general context of the effect of automation:

> In general, whenever any task is automated, the impact is felt far beyond the one task. Rather, the application of automation is a system issue, changing the way work is done, restructuring jobs, shifting the required tasks from one portion of the population to another, and, in many cases, eliminating the need for some functions and adding the need for others. (p. 117)

Norman's general observation applies especially well to science. In contemporary society, scientific research is regarded as an important means of innovation and solution to social problems, and so a huge amount of public resources is spent on it. Therefore, if it turns out that AI-driven science is much more productive in generating useful knowledge and applications than traditional modes of research, it is possible that governments will spend resources on the former rather than the latter and, consequently, many researchers will lose their jobs.

Also, the prediction that AI will liberate researchers seems too optimistic. It is often said that past attempts to automate labors have not liberated humans from unattractive works, but instead gave birth to populations engaged in mechanical and inhuman works. Norman says, "Even successful automation always comes at a price" (ibid.) and mentions drawbacks of automation such as the need for maintenance. To apply these general lessons, we should not expect that the automation of science will eliminate dull tasks in research. Indeed, the job of scientific researcher itself might be unattractive even if it would not be replaced.

Another worry about the automation of science is concerned with its more far-reaching effect: it may undermine a central value of science. Science constitutes an important part of human culture. However, the automation of science may undermine its cultural value by decreasing the room for human ingenuity in scientific research. To specify the threat concretely and to find a way to deal with it, it is useful to refer to discussions held in the past when new technologies were introduced into science. In this spirit, let us reflect on discourses made when "big science" emerged.

Big science is a family of scientific programs that are funded large budgets, carried out by large teams of researchers, and exploit large devices (e.g., giant telescopes, spacecrafts, particle accelerators, nuclear fusion reactor, etc.). Starting with the Manhattan Project, they emerged during and after the Second World War. Since then they have caused controversies due to the impacts on science and society. For example, as we saw above, Dyson viewed human creativity as the

basis of a central value of science and claimed that large devices are not important for scientific research. Likewise, Weinberg, a nuclear physicist who coined the term "big science," was worried about the consequences of big science's growth (Weinberg 1961). One of his concerns was directed at scientists' tendency of "spending money instead of thought" (ibid., p. 162), which, Weinberg states, may ruin science. These discourses suggest that scientific research has an aspect as a drama, protagonists of which are humans who try to understand the world by exercising their creativity. This is, however, the very aspect that the automation of science may endanger.

Therefore, in introducing AI into science, we must specify its implications and devise some measures against its demerits beforehand. To set a guiding principle in this attempt, we should refer to Weinberg's following comment:

> Big Science is an inevitable stage in the development of science and, for better or for worse, it is here to stay. What we must do is learn to live with Big Science. We must make Big Science flourish without, at the same time, allowing it to trample Little Science (ibid., p. 162).

The first sentence of this passage corresponds to King's diagnosis that science will inevitably be automated (see Section 2.2). Thus, punning on Weinberg's passage, we should say, "What we must do is learn to live with AI-driven Science. We must make AI-driven Science flourish without, at the same time, allowing it to trample Humanity's Science."

Thus, we must discuss potential demerits of automating science and measures against them so that we find desirable ways of introducing AI into science. Some of the matters on the agenda are concerned with science policy: for example, resource allocation between AI-driven science programs and traditional ones, measures to ensure employment of researchers, and so on. Some are concerned with science education, such as alteration of science curriculum in universities.[5] To address these issues, it is important to take opinions from various stakeholders such as researchers, practitioners of science policy, science education and science communication, and broader citizens.

[5] The automation of science also raises issues concerning institutional systems of scientific research such as the authorship system and the referee system, although they are not discussed in the present article due to the limitation of space.

Conclusion

This article examined the implications of the automation of science, focusing on creativity. My view is that, although the attempt to automate science faces difficult challenges in realizing artificial creativity, it nevertheless will have significant impacts, both desirable and undesirable, on science and society.

In conclusion, I stress that the automation of science raises many issues that require transdisciplinary research and discussions. On the one hand, to address issues concerning its effect on science and society, it is essential to hold a discussion whose participants include not only AI researchers and scientific researchers who use AI systems, but also researchers and practitioners of science policy, science education, and science communication, as well as sociologists and philosophers of science. On the other hand, issues concerning the possibility of automating scientific discovery provide opportunities for reconsidering philosophical questions such as "What is scientific discovery" and "What kinds of values does science have?" from a new perspective. Although these topics are deeply philosophical, they also require contributions from researchers of empirical sciences such as cognitive scientists, psychologists, and sociologists. By coping with these issues in transdisciplinary collaboration, we will gain a better understanding of the nature and values of science, and this will be one of the most important benefits of AI-driven science.

References

Anzai, Y. 1985. *The Psychology of Problem Solving*, Tokyo: Chuokoron-shinsha. (Written in Japanese)

Asimov, I. 1959/2014. 'How do people get new ideas?', reprinted in *MIT Technology Review*, October 20, 2014. [Retrieved February 4, 2020 from https://www.technologyreview.com/s/531911/isaac-asimov-asks-how-do-people-get-new-ideas/]

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms (Second Edition)*, London: Routledge.

——— 2014. 'Creativity and artificial intelligence: a contradiction in terms?', in E. S. Paul & S. B. Kaufman (eds.), *The Philosophy of Creativity: New Essays*, pp. 224-244, Oxford: Oxford University Press.

Bowler, P. J. 1983. *Evolution: The History of an Idea*, Berkeley: University of California press.
Chiang, T. 2002. 'The evolution of human science', in his *Stories of Your Life and Others*, pp. 241-244, New York: Tor Books (originally published in 2000 as 'Catching crumbs from the table', *Nature* 405 (6786): 517).
Dawkins, R. 1986. *The Blind Watchmaker*, Harlow: Longman Scientific & Technical.
Dunbar, K. 1997. 'How scientist think: On-line creativity and conceptual change in science', in T. B. Ward, S. M. Smith & J. Vaid (eds.), *Creative thought: An investigation of conceptual structures and processes*, pp. 461-493, Washington, DC: American Psychological Association.
Dyson, F. J. 1988. 'Science and space', in his *Infinite in All Directions*, pp.158-179, New York: Harper & Row.
Frey, C. B. & Osborne, M. A. 2013. *The Future of Employment: How Susceptible Are Jobs to Computerisation?* [Retrieved February 4, 2020 from https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf]
Gaut, B. 2003. 'Creativity and imagination', in B. Gaut, & P. Livingston (eds.), *The Creation of Art: New Essays in Philosophical Aesthetics*, pp. 148-173, Cambridge: Cambridge University Press.
——— 2010. 'The philosophy of creativity', *Philosophy Compass* 5 (12): 1034-1046.
Holyoak, K. J. & Thagard, P. 1995. *Mental Leaps: Analogy in Creative Thought*, Cambridge, MA: MIT Press.
Kelly, K. 2016. *The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future*, New York: Viking Press.
King, R. D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L.N., Sparkes, A., Whelan, K.E., & Clare, A. 2009. 'The automation of science', *Science* 324 (5923): 85-89.
King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell D. B. & Oliver S. G. 2004. 'Functional genomic hypothesis generation and experimentation by a robot scientist', *Nature* 427 (6971): 247-252.
King, R. D. 2010. 'Rise of the robo scientists', *Scientific American* 304 (1): 72-77.
Kitano, H. 2016a. 'The day AI win the Nobel prize and the future of humanity:

An ultimate grand challenge in AI and scientific discovery', *Jinkou Tinou* 31 (2): 275-286. (Written in Japanese)
——— 2016b. 'Artificial intelligence to win the Nobel prize and beyond: Creating the engine for scientific discovery', *AI Magazine* 37 (1): 39-49.
Langley, P. 2000. 'The computational support of scientific discovery', *International Journal of Human-Computer Studies* 53 (3): 393-410.
Langley, P., Simon, H. A., Bradshaw, G. L. & Zytkow, J. M. 1987. *Scientific Discovery: Computational Explorations of the Creative Processes*, Cambridge, MA: MIT Press.
Lenat, D. B. 1977. 'The ubiquity of discovery', *Artificial Intelligence* 9 (3): 257-285.
Norman, D. 2007. *The Design of Future Things*, New York: Basic Books.
Novitz, D. 1999. 'Creativity and constraint', *Australasian Journal of Philosophy* 77 (1): 67-82.
Poincaré, H. 1908/2001. *Science and Method*, reprinted in his *The Value of Science: Essential Writings of Henri Poincaré*, pp.355-567, translated by F. Maitland, New York: Modern Library.
Science News Staff 2017. 'AI is changing how we do science. Get a glimpse', *Science* Jul. 5, 2017. [Retrieved February 4, 2020 from http://www.sciencemag.org/news/2017/07/ai-changing-how-we-do-science-get-glimpse]
Searle, J. 1980. 'Minds, brains, and programs', *Behavioral and Brain Sciences* 3 (3): 417-424.
Simon, H. 1996. *The Sciences of the Artificial (3rd edition)*, Cambridge, MA: MIT Press.
Simonton, D. K. 2004. *Creativity in Science: Chance, Logic Genius, and Zeitgeist*, Cambridge: Cambridge University Press.
Stix, G. 2005. 'Molecular treasure hunt', *Scientific American* 292 (5): 88-91.
Turing, A. 1950. 'Computing machinery and intelligence', *Mind* 59 (236): 433-460.
Weinberg, A. 1961. 'Impact of large-scale science on the United States', *Science* 134 (3473): 161-164.

Journal of Philosophy of Life Vol.13, No.1 (January 2023):64-96

Why Autonomous Agents Should Not Be Built for War

István Zoltán Zárdai*

1. Introduction

Several institutions are working on creating autonomous agents (the term covers both AIs and AIs controlling robots, and it will be shortened as AAs in the rest of the paper) for purposes of waging war.[1] AAs differ from usual automated machines because instead of their behavior, they are defined by the objectives they pursue. Modern AAs can learn and optimise their behavior to pursue their goal effectively in the given environment. Therefore, it is often difficult, even for their designers, to predict what behavior an AA will exhibit. This is not a real problem in the case of sophisticated industrial robots, or AIs trained to beat chess masters, since their behavior is geared towards achieving a handful of repetitive goals, so their behavior patterns are fixed, and the means at their disposal are limited. This paper addresses worries concerning genuinely autonomous AAs, especially ones that are supposed to make decisions in morally relevant situations. The paper argues that such AAs need to have sophisticated capacities, in some cases even human-like capacities, and this leads to specific risks that do not exist in the case of other machines.

AAs have the potential to be useful for military purposes, and there are some considerations that might make it tempting to deploy them in combat theatres: fewer human lives have to be sacrificed if only machines are destroyed; their deployment can deter war, since they might make engagement much more costly for enemies; and several tasks can be performed more efficiently by AAs because they are not subject to mood swings, are precise, do not become tired or lose focus,

* Visiting researcher, Department of Ethics, Faculty of Letters, Keio University. Office 20404 South Building Keio University, 2-15-45, Mita, Minato-ku, Tokyo 1088345, Japan. Email: zizistv[a]gmail.com

[1] Unmanned combat aerial vehicles (UCAVs) are one type of such agents. At the moment, most of the models deployed are under real-time human control or monitored by human controllers, but more and more autonomous models are being developed in China, India, Israel, Italy, Pakistan, Russia, Turkey, the USA, and other states. This development goes hand in hand with research by these and other states on increasingly sophisticated military AIs.

do not disobey orders, and are not afraid. AAs that complement existing military capacities and operate under constant human control and surveillance are already operative.[2] Such systems at the moment are not allowed to make lethal decisions; that is, they are not autonomous in the US military's sense that once they are activated, they can select and "engage" (i.e., intentionally fire at and kill) targets without further intervention by a human operator.[3]

At the same time, it is safe to assume that there are some developers and governments that agree that AAs should also be allowed to reason, decide, and act independently, without having to always involve human operators when deciding about a potentially lethal attack or an attack that is intended to be lethal.[4] If nothing else, the current trend of AA development makes this highly likely: there is no unilaterally accepted ban or constraint on the development of military AI and other robotics that would prevent this, and states are investing openly more and more into such research and encourage the integration of the latest technology into military tools.[5] This naturally leads to an "arms race" situation, as described by Chan (2019).

Those developing lethal autonomous weapon systems do not mean "autonomy" to imply that AAs should be able to choose their strategic objectives, switch which side they fight on, decide on what mission objectives they pursue, and so on. They mean that the AAs should be allowed to kill, or to use a popular military and marketing euphemism, to "engage targets" and "use lethal force." Such autonomous AAs, such as turrets, submarines, or bombers, could decide on their own when to fire missiles or torpedoes or to drop bombs.

I argue that the development and deployment of such AAs should be strictly prohibited because 1) if these systems do not possess very similar capacities to

[2] It is highly likely that the Israeli Defence Forces used UCAVs during several operations that killed civilians unlawfully, and the same is true of the US, while China has been reported to use such technology for massive-scale surveillance of its own citizens.

[3] Concerns have been raised that this definition is murky and has not been observed very closely when applying regulating procedures (see Gubrud 2015). It is known that there are deployed and currently developed weapon systems capable of performing various manoeuvring and offensive tasks without human control and oversight. For examples, see Fryer-Biggs (2019).

[4] Noel Sharkey (2010) discusses some plans, and the ethical worries about them, by the US military to transition from man-in-the-loop to full-autonomy systems.

[5] While no one admits that they want to develop "killer robots," research into offensive lethal autonomous weapon systems is known to be ongoing. Several AI researchers have lobbied the United Nations to help institute a ban on such developments, and the UN has discussed this, without much success or support from the major military technology-developing countries so far (see Glaser 2016).

humans, they are not able to make moral judgments in a way similar to humans and as such should not be allowed to act autonomously on a battlefield, and 2) in case they can be developed so as to possess very similar capacities to humans, it would be morally impermissible to create them for the sole purpose of sending them to war. In the former case, it would be unclear who is responsible for the ensuing deaths, and the AAs are not proper subjects for being held responsible and for being subjected to punishment.[6] In the latter case, we would create something that has the concept of value and can apply it in ways that humans can, and then send it out to kill and to get killed. Doing this to beings that we created, and which have capacities sufficiently similar to ours, would be unallowably cruel.[7]

In this essay, I first offer a brief discussion of what it means to be an agent and then of the capacities that ground moral responsibility in section 2. In section 3, I argue that AAs should not be allowed to act autonomously—in the above-mentioned sense, that once they are activated, they can select, fire at, and kill targets without further intervention by a human operator—because they are not the right kind of agents, since they cannot be held responsible. In section 4, I defend the claim that while human-like AAs could be held morally responsible, for moral reasons it cannot be allowed to create them. Finally, in section 5, I anticipate and address a number of possible objections, before summarizing the key points and concluding the paper.

[6] For an argument against the deployment of autonomous weapon systems on the grounds that the relations of responsibility between commanders, manufacturers, controllers, and other involved parties cannot be made sufficiently clear to warrant deploying such robots in war, see Sparrow (2007).

[7] There are two further important issues that will not be discussed here: 1) Sometimes it is argued that decision making about killing in war is somehow more mechanical or more straightforward than decision making about killing in other situations, and hence it is more permissible or even desirable to put AAs in charge of decisions than putting them in charge of important decisions in, say, politics or healthcare. There is, however, no good reason to suppose that decisions about killing are really much simpler in a war context than in other situations. They are surely of a different kind for several reasons. For example, they are unlike conflicts between civilians; they are in the contexts of political goals of communities like nations, groups, parties, countries, or alliances; they fall under specific regulations; etc. These differences do not mean that they are simpler, only that they are different. 2) Some philosophers have argued that killing in war can be understood analogously to the ethics of cases of self-defense in civilian life. This position is deeply oversimplifying and leads to morally troubling results, but since it is a reasoned and detailed position that deserves its own treatment, I will not attempt to address it here.

2. Autonomous Reasoning, Decision Making, and Acting

Agential capacities

I present in this section a broad view of agency. This view recognises artificial agents, such as AAs, as capable of acting. Then I argue that agents need to possess and exercise specific capacities in the right way. Beings that do not possess these capacities cannot be understood as choosing or as making decisions, and they cannot be held morally responsible. Hence, if AAs are to make life-and-death decisions, they need to possess these capacities. If this can be shown, the conclusion has both ethical and practical consequences. The ethical consequence is that there is a strong normative reason for not creating such AAs because it is immoral to do so, since it is like creating humans solely for the purposes of war. The practical consequence is that research aimed at creating such AAs might be reconfigured and the enormous funds that militaries, governments, and companies would spend on this could be spent on more useful research and practical activities vital for our societies.

My focus in general will be mostly on showing which capacities AAs would need to have in order to be human-like, and I will say less about why it would be wrong to manufacture such AAs for the purposes of war. This is so for two reasons: First, I think not enough has been published on the issue of exactly what kind of agents AAs are, should be, or could be. The literature on them has only a few discussions on how they compare to other agents.[8] At the same time, many arguments are available in the ethics literature that argue against manufacturing human-like AAs,[9] as well as broader ethical arguments that could be applied by analogy, such as from the literature on cloning, or Kantian arguments that could be employed to show that human-like agents are persons, and as persons they should be respected and not treated as means for ends like fighting wars.

Reasoning, decision making, acting, and moral responsibility are things an agent needs to possess and exhibit in order to act intentionally and/or voluntarily

[8] Some of the more detailed and higher-quality discussions are by Purves, Jenkins, and Strawser (2015) and Misselhorn (2018).
[9] For an argument that it is most likely impossible to programme AI to make moral decisions reliably, see Jenkins and Purves (2016).

and to be an appropriate target for being held morally accountable.[10] The role of these capacities and activities in this discussion is clear: reasoning, decision making, and intentional and/or voluntary action have to happen at some point in order to ground attributing moral responsibility to an agent and to treat it as a morally competent agent. Of course, both in everyday life and in law we hold agents responsible in cases when they are not active as agents, such as in cases of negligence (think, for example, of a security guard forgetting to lock a door) or omissions, but the cases that interest us in this essay are mainly those of active killing, and hence of intentional and/or voluntary action. Also, we only hold agents responsible who, we are more or less sure, possess such capacities.[11] This is also clearly indicated by the fact that we treat children, those living with serious mental disabilities, those in a permanent coma, and minors differently, as well as by the fact that we do not treat animals—even such highly intelligent and social species as dolphins, dogs, or chimpanzees—as responsible.[12]

Let me start the discussion by clarifying how this position relates to some of the issues that are related to but are not my main focus here. Sometimes it is claimed that if AAs have a small number of key features—like rational thinking or producing appropriate emotional reactions—in common with humans, then they are human-like. I call this way of thinking 'the smallest common denominator' approach. It is usually endorsed by people who are optimistic about the possibility of creating autonomous, human-like AAs. This does not mean that I presuppose that there is no free will or that moral responsibility and autonomy can be reduced to deterministic lower-level processes. What it means is that people who believe that creating autonomous human-like AAs is possible have to accept the premises proposed, since it is not intelligible to claim that something can be responsible if it does not meet at least these criteria (i.e., the conditions are sufficient, although one could argue that more is needed; that is, they are not

[10] It adds to the complications of creating AI that can reliably fulfil criteria of intentional and voluntary behaviour that it is an open question what exactly these concepts mean and which conception of them is the most relevant to law, morality, and war. I explore some of these difficulties in connection with interpreting the famous Knobe Effect in Zardai (2022). The literature on this effect shows what a dazzling array of potential explanations there is of how our practice of holding others responsible might work.

[11] The approach recommended here to what moral responsibility is, is compatible with a range of views of responsibility. It is particularly amenable to a Strawsonian or moral sentimentalist view. See, for example, Strawson (1963/2003: 78-81), as well as the work of Paul Russell and R. J. Wallace.

[12] Strawson 1963/2000: 78-81.

necessary). I want to show that if we accept that we can create AAs that could be autonomous and responsible, in the sense in which humans are, that possibility is an even stronger reason not to do so.

Free will

I also want to make clear that in the essay I do not discuss the free will debate, and what I say does not bear on it directly.[13] The kind of AAs that I describe as human-like are certainly determined and not free in the sense in which libertarians about free will would require them to be undetermined for them to count as free, and in virtue of having that kind of freedom, also morally responsible. Nevertheless, I think that such AAs can be morally responsible if they meet a number of conditions. My view does not claim that free will is mischaracterised by either libertarians or hard determinists about it, but it does claim that responsibility is possible without it, at least in some important cases. Hence, regarding the metaphysics of free will, my position is neutral, while regarding responsibility, it is closer to compatibilist positions. It can be accepted by libertarians and hard determinists, if they concede that something could have moral responsibility—maybe in a limited sense—even though it would not be free, or if they can plausibly claim that the moral responsibility I attribute to human-like AAs here is not the moral responsibility they are interested in because they are looking for a fuller, more demanding notion. I'm happy to accept either approach. And I think there are two good reasons why my own line is adequate. First, there are several serious and well worked out, even if not universally accepted, compatibilist positions in the literature. In this essay, I could hardly do better than those detailed discussions. At the same time, it is sensible to rely on the good work others have already done on the topic.

One of the best reasons to be a compatibilist is Harry Frankfurt's work on how to make sense of moral responsibility in cases when an agent cannot do otherwise than it does.[14] Frankfurt rejects what is called the Principle of Alternate Possibilities (PAP). The PAP says that a person is morally responsible for what he

[13] For short but excellent overviews of the major positions and debated points in the current literature on free will, see Watson (2003), Pink (2010), and chapter 1 of Mele (2017).
[14] Frankfurt (1969/1988), (1971/1988).

does only if he could have done otherwise.[15] That agents can be held responsible even in cases when they could not have done otherwise can be shown by a thought experiment proposed by Frankfurt. Imagine that Black is about to kill Jones by shooting him. Unbeknownst to Black, a team of mad scientists has planted a chip in his brain. They are monitoring Black's behavior and brain activities and are able to tell what Black wants (this is not possible at this stage of science, but it is helpful for understanding responsibility to imagine that it is). If Black for some reason decides not to shoot Jones, the mad scientists activate the chip in his brain, and this will change Black's neural activities so as to make him shoot Jones. That is, Black cannot do otherwise; he will shoot Jones. As it happens, Black sticks to intentions and, without any interference from the mad scientists, goes on to shoot Jones. He could not have done otherwise and is nevertheless responsible for shooting Jones.[16] What follows from this for the possibility of the responsibility of AI is that even if we look at AIs as machines that act according to mechanisms and are fully determined, that does not mean that they cannot be responsible for what they do. They can still be morally responsible, just as humans are.

A second reason to accept the approach proposed here is that it provides mutually agreeable grounds to libertarians and hard determinists. Since I claim that possessing the well-developed capacities that grown-up humans living without mental and emotional disabilities possess is necessary for the kind of agency that grounds moral responsibility, libertarians can say that such AAs would have free will just like humans if they would possess the same agential capacities and hence would be morally responsible. Also, hard determinists could claim that AAs lack the relevant kind of freedom, just like us, and therefore it does not make sense to attribute moral responsibility to them, the same way it does not make sense to attribute it to us. My further claim, that it would be impermissible to create free and responsible agents for the purposes of war, will still be something that they would have to address. Presumably, for libertarians, the question and how it is to be answered will not change much from cases of other free agents. For hard determinists, it will change in the same way as all other questions about how to deal with moral issues after giving up on moral responsibility are to be dealt with.

[15] Frankfurt (1969/1988: 95).
[16] For the original version of the example, see Frankfurt (1969/1988: 6-7).

To help understand my approach better, I use as a contrastive example a different, libertarian position with which I disagree: Helen Steward's thoughtful view worked out in her *A Metaphysics for Freedom*. Steward claims that

> (…) the falsity of universal determinism is a necessary condition of the possibility of any freedom or moral responsibility there might be. But the best reason for thinking that is so, in my view, is that the falsity of universal determinism is a necessary condition of *agency*—and agency is, in its turn, a necessary condition both of free action and of moral responsibility. (Steward 2012: 4)

That is, according to Steward, if something is not free, it is not an agent. This is a narrow view of what agency is—and perfectly fine for the purposes of Steward's investigations in her book—whereas I endorse a broader notion of agency, according to which several obviously non-free things are agents. Hence, for Steward, some criteria of agency will be part of the explanation of why agents are free and can be morally responsible. My approach requires the strategy that I have described earlier, namely, arguing that the kinds of agents who possess full moral responsibility—healthy adult humans—have specific capacities that distinguish them from other kinds of agents—like acids, wolves, or planets—and having these capacities is what explains why they can be morally responsible. Other agents also do things, just like humans do—acids dissolve other materials, wolves nurse their offspring, planets orbit suns—but nevertheless, they lack the capacities that would ground holding them responsible.

Consciousness

Another issue I address only tangentially is the important notion of consciousness. Every day as humans we experience the evaluative aspects of our existence: When we wake up, we feel somehow and in a kind of mood. We crave coffee, or want to go back to bed, or feel energetic and jump out from under the blanket with our minds already racing through what we need to do that day, and so on. All these cases come with their own specific experiential side that we can be aware of. That is, we have phenomenal consciousness. Due both to behavioural

evidence and neurological similarities, we also have good reasons to suppose that at least the animals closest to us in their social and mental features—dogs, dolphins, elephants, primates and whales (and most mammals really)—have this "what it feels like" aspect of mental life, even if it might be qualitatively and structurally different from ours. Experts on other life forms have argued extensively that birds, and even fish, have conscious experiences.[17]

It has been argued that this experiential facet of our mental life is important for being the kind of agent we are, including having the capacities required for being responsible and for having moral value.[18] This is significant because it supports my approach that humans are the kind of agents that can be morally responsible. Hence, an AA that is human-like has to have phenomenal consciousness too in order to be eligible to be deployed in a war where it has to make decisions that only morally responsible agents should be allowed to make. At the same time, this is also why a human-like AA should not be manufactured with the aim of deploying it in a war.

For the purposes of this essay, I will work with the assumption that if an AA can be manufactured that has all the other capacities necessary for being the kind of agent that can be morally responsible, then it also has phenomenal consciousness. The reason why it is safe to work with this assumption is that several of the distinctively human capacities involve experiential content in a range of ways, including the ability to consciously reflect on that kind of content; to take it as information and grasp it both as felt and experienced, such as when we commiserate with someone or share their joy; and to handle it in a way that enables meta-reasoning (thinking about it, rejecting it, questioning it, etc.) and taking it into consideration when deciding what to do. This is so due to the key role some of the capacities, making use of such experiential information, play in making decisions that render us morally responsible.

Since we are considering what kind of agents should be allowed to act autonomously in combat situations, it is relevant to discuss a case when human

[17] Victoria Braithwaite defends the view that fish experience pain in *Do Fish Feel Pain* (2010). More recently, it has also been argued that fish can even feel "emotional fever" (see Rey et al. 2015).
[18] For the idea that robots are not agents in the relevant sense unless they are conscious, see Talbot, Jenkins, and Purves (2017). For an extended argument for the importance of phenomenal consciousness for having moral status, see Shepherd (2018). I'm not using Shepherd's arguments here but drawing on them to motivate the thought that phenomenal consciousness is relevant for having some of the core capacities involved in actions for which agents can be morally responsible.

soldiers do not act in line with prescribed deliberative procedures. The case of a US patrol in Afghanistan in Taliban territory illustrates this well. The American soldiers were nearing a dwelling known to have given shelter earlier to Taliban fighters. The locals sent out one of their younger children to look after the livestock and meanwhile also take a look at where the American soldiers are. At this point, the American soldiers had intelligence that they will likely be ambushed, and hence could have regarded the child, who repeatedly looked in their direction in an obvious manner, as a combatant because she was gathering information that could endanger the soldiers. However, the locals' gamble paid off: the soldiers refrained from shooting a child and dispersed.

The decision making of the US soldiers in this case can be reconstructed in two ways: one is to imagine that it relied on an absolute ban on killing children, and one is to imagine that it involved some form of compassion or empathy. The first is unlikely, since the rules of war posit that children can be targeted in such highly specific cases when they are contributing to the war effort and their activity constitutes imminent danger. It is also known that militaries do kill children in such situations, the US not being an exception to this. Hence, the second line of interpretation is what is more likely. The soldiers' reasoning could have taken any of a large number of imaginable avenues; nevertheless, it is plausible that they relied to some extent on emotional experiences and empathy. If compassion played a factor, then such experiential features entered the reasoning because compassion is a feeling of sorrow or pity awoken by the suffering—real or imagined—of others. If pity for the child influenced the reasoning, then information provided by phenomenal consciousness was likewise employed, since pity is an emotion of sympathy, sorrow, or regret caused by the suffering of others. If the soldiers were reminded of their own children, then emotion played a role by influencing their reasoning through invoking impressions of the sadness that the shooting of the child would cause the child's family.

Agency

Considerations regarding the importance of phenomenal consciousness led us to discuss the capacities that are important for moral responsibility. However, by doing so, I have skipped ahead. Let me say more about what is needed for

something to be an agent in a wide sense. There are different views of what agency is, and what partly determines which notion of agency one wants to spell out is what work one wants the notion to do. In my case, the notion of agency serves the purpose of solving the problem of action.[19] The problem of action can be briefly introduced in the following way: The world is full of changes. Some of these changes can be attributed to agents, and such changes are actions or the results and consequences of actions. In some cases, the same change could happen to an agent while the agent is passive—say, if I fall on the bed after my dog jumps on me and pushes me over—and in an active way—say, if I fall on the bed because I'm pretending to fall after my niece shoots me with her imaginary gun while we are playing. While the changes are different in these two cases—someone's falling is not the same as someone's pretending to fall—the results occurring—my landing on the bed after suddenly leaning back—are identical. On the view of action that I defend, the distinction here can be understood in terms of the role the agent's capacities played in these two changes.[20] When I'm being pushed over and fall, none of my agential capacities are active in the sense that they are not bringing about this fall and their activity is not identical with my falling. In contrast, in the case when I pretend to fall, several of my agential capacities are actualised—my coordination, controlling my balance, and so on—and together with my body's landing on the bed, they constitute my pretend-falling (Zardai 2016, 2019).

This way of solving the problem of action has several attractions: thinking about actions as identical with or partially constituted by actualisations of agential capacities enables us to understand what actions are without relying on the concepts of reason, intention, will, plan, value, or other concepts involving the idea of higher-order mental abilities.[21] This, in turn, enables us to endorse a wide

[19] Harry Frankfurt (1978/1988) dubbed this issue so in his paper titled 'The Problem of Action.'

[20] The relevant notion of change is defined by Georg von Wright (1963) in his *Norm and Action*, chapter 3.

[21] For the most famous and influential narrow notion of agency, see the causal theory of action as presented in Donald Davidson's work, esp. Davidson (1980). The causal theory of action defines action as bodily movement caused by an intention (or a belief–desire pair that amounts to an intention), where the content of the intention is a judgment that the agent also has reason to perform the action that the movement is identical with. A current version of the causal theory is endorsed, for example, by Al Mele (2017, esp. chapters 1 and 3). I argued for abandoning the causal theory of action in detail (Zardai 2016, 2019). There are also other views of action and agency—like the new version of agent causation defended by Alvarez and Hyman—however, for the current purposes, the view I have introduced will do well in helping to capture what is needed for moral responsibility, that is, what

notion of agency. This notion of agency can be briefly presented as follows: anything is an agent that has specific capacities in virtue of its overall structure (i.e., in virtue of the relevant structure that makes it the kind of agent that it is). In the case of humans, our biological and psychological structure are both relevant to what we are and hence help highlight our agential capacities qua humans.[22] The view can also make sense of animal agency, and even of the agency of inanimate things, like chemicals, natural phenomena, and machines. As such, it works well for addressing issues about AI too, since it enables us to treat AAs as agents, without also automatically attributing them the ability to act intentionally or voluntarily, or the status of morally responsible agents. Narrower notions of agency, which define actions and agency with the help of intentionality, the will, goals, directedness, and other concepts that immediately invest a stake in normative issues, have a harder time providing an account of the behavior of robots and other artificial agents. Such agents all have a specific structure in virtue of which they are what they are, and this structure equips them with their agential capacities. According to the broad view of agency employed here, when these capacities are instantiated, the agents are acting.

So far in the essay, I have clarified the current discussion's relation to the free will debate and the role of phenomenal consciousness, and I have introduced working notions of action and agency. All this should enable us to see that in theory there is no obstacle to creating human-like AAs, since determinism allows for such agents as well as the relevant notion of moral responsibility. Also, we can propose a notion of agency and action that can explain why things like AAs are agents and how they can act. This notion of agency and action does not have any premises or presuppositions that are hard to accept: it is compatible with a broadly naturalist world view—one that accepts that science is the authority in its fields of study, while being open to the possibility of future changes and paradigm shifts in the sciences, and also being a realist about morality, values, and social entities. At this point, we can move on to discuss in which specific aspects AAs would need to be human-like to possess moral responsibility in a substantial sense. For this, we will first need to discuss what moral responsibility is and say a few words

kind of agents human-like AAs would need to be for them to make life-and-death decisions.

[22] This view of agency is inspired partly by Maria Alvarez's and John Hyman's work on action and agency. See Alvarez and Hyman (1998) and Hyman (2015, esp. chapter 2).

about autonomy.

In the introduction, I noted that the military uses "autonomous" to mean that after an agent is activated, it can select, fire at, and kill targets without further approval by a human operator. My essay is trying to show that for AAs to be allowed to do so would require AAs to be human-like. Otherwise, they would not be the right kind of agents to make a decision about the life and death of a human, or even to engage in any act of war that carries the risk of causing harm to humans, without human control and oversight. For example, if targeting would be carried out or get confirmed by a human operator, as is done currently in the case of cruise missiles when targeting other ships or planes in a battle, or the battlespace would be managed in a way that would ensure that non-combatants and allies could not be injured, the automatic operation of AAs would not be controversial. But AAs that are not human-like making decisions in such scenarios is not acceptable.

Autonomy

My point can also be understood as addressing the following issue: militaries might attempt to give different meanings to "autonomous" in order to claim that a specific machine should be allowed to operate without human oversight and control when deciding to kill. One way they might try to do this is to say that because the machine meets the required—weak—notion of autonomy, it does not need to be controlled. This strategy downplays the requirements of autonomy and tries to display many AAs as meeting them. Conversely, militaries might postulate a very strong notion of autonomy and claim that we already allow several machines to destroy targets without human intervention, such as in the case of air defense systems. Taking this strategy, militaries could say that AAs are similar to the highly automated systems that we already allow to target and kill humans without approval by a human operator. Hence, the operation of such AAs does not need to involve humans in the oversight and control of decisions to kill. My point applies to both of these attempts at loosening moral norms by watering down the notion of autonomy or making it unreasonably demanding: we need to resist such redefinitions of autonomy. Unless the specific machines in question, the AAs, can make decisions in a human-like way, *in the same way* a human operator would do, they should not operate without human oversight when making decisions about

taking potentially lethal actions. Hence, just because automatic missile defense systems or drones might currently be allowed to select their targets without the control of a human operator in some cases, that does not mean that they are not autonomous in the sense of autonomous that calls for control if they target someone for killing.[23]

According to the military's technical definition, something is autonomous if "after an agent is activated, it can select and fire at and kill targets without further intervention by a human operator." This shows that there are clear technical uses of the term. These technical notions might serve useful purposes when classifying weapon systems from a quality control, testing, budgeting, or safety perspective. However, they are irrelevant when we are trying to decide whether the construction of military AAs that could kill without explicit, real-time human orders should be allowed. It is irrelevant because such technical notions of autonomy can be met by several animals and also by automatic equipment like industrial assembly-line robots or statistical AIs analysing complex data sets and creating action recommendations for humans. Still, neither animals nor such robots qualify as responsible agents. The everyday moral notion of autonomy should override the technical notion of autonomy proposed by the military, since the actions of weapon systems affect everyone and have a moral dimension. When AAs perform an action that leads to harm or other negative consequences, the ones bearing responsibility are the humans interacting with these animals and robots and the producers and operators of the robots. This model should also apply to military AAs.

The notion of autonomy that is interesting for us when thinking about whether militaries should be permitted to develop, purchase, and deploy AAs is a robust notion. If an agent qualifies as autonomous according to it, then it can also qualify as morally responsible. We do not want to claim that something has autonomy in

[23] A suspected case of such over-definition of autonomy has been highlighted in the report of Drone Wars UK in their 2018 on the development of autonomous military drones (UVACs) in the UK. There is a tension between the verbal commitment of UK politicians not to fund the research, purchase, and deployment of autonomous weapon systems, and the military's repeated announcement that it is committed to go automated to a high degree. The possibility cannot be ignored that politicians and the military would attempt to define autonomous in such a demanding sense that they can claim of UVACs and other automated weapon systems, which would meet the US Army's criteria of autonomy, that they are simply automatic but not autonomous and hence can be used under more lax regulation and supervision that currently apply to automatic weapon systems. See Burt (2018: 54-55).

this sense if it cannot bear moral responsibility for its actions. This is so because something that lacks the capacities needed for bearing moral responsibility lacks the capacities for understanding the morally charged situations that it is operating in. Such agents would not understand features of key importance of the situations they would have to decide and act in. In such circumstances, we humans could not understand properly what they do, and we could not allocate moral responsibility effectively to their operators and manufacturers either, both of which are further key criteria that need to be met before any AAs can be deployed.

It seems then that the notion of autonomy is not the most helpful concept to get at the answer to the question of whether machines should be allowed to make decisions and execute potentially lethal actions without real-time human permission. It can create confusion, which might be exploited by vested interests, and its relevant requirements can be more helpfully discussed in more obviously moral terms. Autonomy is originally a political term that means that a nation can decide about its affairs without external interference, and in this way it is connected to sovereignty.[24] Politicians, militaries, and other industry use the term in a variety of technical senses and with different purposes. What I argued for here is that there is a sense of autonomy that is tied closely to moral responsibility, and this is the relevant notion to this discussion. We should resist attempts to redefine autonomy. And to properly understand the relevant sense of autonomy and find the answer to our question, what we really need to do is to think about moral responsibility.

3. Responsibility

There are different kinds of responsibility. It is plausible to distinguish at least three kinds: causal, role, and moral. An agent is causally responsible for something if it is part of the causal chain leading to that thing. For example, the typhoon can be responsible in this way for the fallen trees on the beach. Role responsibility depends on what duties an agent has in virtue of having a specific role. For example, parents in their role qua parents are responsible for the health and well-being of their children, including that they are fed and clothed appropriately, and that they attend school at least until they reach the legally

[24] Darwall (2006: 263-265).

required age of compulsory education. What we are interested in here, moral responsibility, means that an agent can be held accountable and be the target of specific normative responses. Agents—whether individual or group agents—are held responsible for particular items, such as choices, decisions, actions, omissions, results and consequences, character traits, and so on. These kinds of responsibility can come apart.[25]

The main focus in this essay is on moral responsibility, because this is the kind of responsibility that is connected to what is morally good and bad, right and wrong, and permissible or impermissible. The praise- and blameworthiness of persons is also connected to moral responsibility, since praise and blame are the relevant normative responses helping us to map the boundaries of attributing responsibility. What are the conditions for an agent to count as morally responsible? Exploring these criteria will help to show that AAs cannot meet such criteria currently, and they could only meet them if they would be sufficiently human-like. However, if they are human-like, then they should not be created and deployed to go to war.

Normally we attribute responsibility to persons. Views of responsibility work with the idea that the agent has a set of capacities and abilities, and these play, or could play, an essential role in their actions for which we judge them responsible. Such capacities and abilities include having values and preferences; being able to reason about these reflexively; being able to feel, to be vulnerable, know that they have something to lose by ceasing to exist or getting injured and incapacitated, to possess the ability to develop empathy and sympathy, to entertain plans and goals and rank these to choose between them or revise them; and a number of other capacities as well. If we judge that an agent acted in a way that its relevant capacities were or could have been involved, we can morally judge the agent.

Moral judgment usually goes together with blaming and praising. Blaming and praising can be understood as reactive attitudes, reactions that we have towards other people. Peter Strawson emphasised the importance of such attitudes

> We should think of the many different kinds of relationship which we can have with other people—as sharers of common interests; as members of the same family; as colleagues; as friends; as lovers; as chance parties to an

[25] Fischer (2010: 309-310).

> enormous range of transactions and encounters. Then we should think in each of these connections in turn, and in others, of the kind of importance we attach to the attitudes and intentions towards us of those who stand in these relationships to us, and of the kinds of *reactive* attitudes and feelings to which we ourselves are prone. (Strawson 1963/2003: 76)

Naturally, specific attitudes are appropriate under different circumstances, and to find the key to when blame and praise are appropriate, Strawson turns to considerations about when it is appropriate to feel resentment towards someone for what they have done and act on the basis of it towards them. We can say with Jonathan Glover, who is following Strawson here, that "To say that someone is morally responsible for what he does may be to say that he can legitimately be praised or blamed if either of these responses is appropriate to the action in question."[26] Considering when resentment, and more specifically blame, is appropriate and when it is not can help us to understand which capacities and abilities an agent must have and what external conditions must hold for blaming to be appropriate.[27] This can be our guide to understanding responsibility.

Since the publication of Strawson's original work, a variety of accounts of moral responsibility have been published. It is not crucial for us to focus on any of those here, and I will make use of ideas from more than one to help get a grasp of the capacities needed for moral responsibility. Another way to make sense of morality is to claim that "For an agent to be morally responsible for an item is (…) for that item to be attributable to the agent in a way that would make it in principle justifiable to react to the agent in certain distinctive ways."[28] Being responsible means a kind of attributability in this case, and the justified reactions would consist of judgments amounting to appraisals of the agent's moral virtues and vices, or to more inclusive judgments of the agent.[29] Whether we endorse this view of moral responsibility or Strawson's view, it is a common strand that being responsible depends on whether the agent meets the criteria of attributability.

As a quick aside before moving on to discussing the criteria of attributability, I want to summarise an important distinction between moral responsibility and

[26] Glover (1970: 19).
[27] Strawson (1963/2003: 75-79).
[28] Fischer (2010: 310).
[29] Fischer (2010: 311).

autonomy. In his overview of the two notions, Fischer argues that

> moral responsibility is a necessary but not sufficient condition for autonomy. (…) In order to be an autonomous agent (…), one must be a morally responsible agent (…). But some additional features must also be present; one can be morally responsible without being autonomous. Put metaphorically, the crucial additional ingredient is: 'listening to one's own voice' or 'being guided internally'. (Fischer 2010: 312)

I have to admit that while I find it plausible to make a distinction between the two notions, I am somewhat baffled by what exactly it would mean to listen to one's own voice in this context. The notion of autonomy is, as I mentioned earlier, historically connected to the notion of *political* self-determination. On the kind of compatibilist and naturalist approach that I have taken to agents and morality, autonomy seems not to add much to the personal dimensions of freedom or responsibility that it would be worth wanting. Hence, I will put the notion of autonomy—as I suggested earlier—aside and keep focusing on moral responsibility.[30]

It has sometimes been recognised that the conditions of praise and blame might be asymmetrical.[31] I will here work with the assumption, which to me is plausible, that meriting praise is harder than earning blame. In line with this, I also claim that to elicit a reaction of praise, one must do something both intentionally and voluntarily, which involves acting for a reason, and in case one merits moral praise, one must have acted for the right moral reason.[32] Blame comes to one

[30] Not having to rely on the notion of autonomy has further benefits too: Autonomy usually implies that agents can act against their own considered best judgments. In morally charged cases, this would mean that they can act against what they think they should do and what they think would be best. Whether humans actually have this kind of autonomy is debated (some find it too demanding a description, which humans cannot actually meet, since they cannot cut themselves loose from the motivations influencing their judgments); nevertheless, it could be possible to have AAs that can do so. Since such decisions are often irrational and go against the results of reasoning that the agent reached by using its capacities aimed at coherent, rational conclusions in line with their evaluations—emotions, desires, values—too, it seems exceedingly risky to construct such agents. For a detailed discussion of whether the structure of human agency requires agents to act in line with what they deem best—that is, to act under the guise of the good—see Kieran Setiya's 'Sympathy for the Devil' (2010/2017).

[31] See, for example, Nelkin (2011).

[32] Think of Kant's example of the merchant who, if he sets prices fairly because he wants to attract more customers, is not acting in a morally good way, whereas if he is doing so out of a sense of respect

much easier: one can be blamed when one acts intentionally but involuntarily, or unintentionally but voluntarily, and even if one acted neither intentionally nor voluntarily but neglected to do something one was reasonably expected to do.

To understand what this means for the capacities required to be a proper subject for moral responsibility, I spell out a broad notion of intentional and of voluntary. An agent acted intentionally when the agent pursued a goal in acting, wanted to achieve that goal, and this wanting (motivation) structured its behavior. This loose characterisation of intentional action captures that agents act intentionally if their action expresses the motivation—be that a desire, sense of duty, plan, or any other pro-attitude—that moves the agent to act and also specifies the agent's goal in its content. It also shows that intentionality is mainly a question of the agent's psychology and its links with the action. It also indicates that intentional actions have a means–end structure, and instrumental reasoning aimed at achieving a goal by means of doing something plays a role in them. Since agents receive moral praise or blame when they do morally good or bad things, or they exemplify their virtues or vices, agents also have to be competent in perceiving the facts that are morally relevant in a given case and treat those as considerations—as reasons—for and against acting. A wide variety of facts can provide a moral reason to do something; for example, my sister's being ill might be a reason for me to visit her, the political system's corruptness might be a reason for citizens to protest, someone's owing money to their bank might be a reason to pay it back, and so on. This is so due to the complexity of morality and requires agents who can bear moral responsibility and can judge others morally, to have the capacities and abilities to recognise all such facts and react to them adequately. In many cases, such as when my sister is ill, this will clearly require more than merely rational understanding that in such cases the right thing is to help her. Visiting my sister out of sheer adherence to a rule is problematic: I might be doing the right deed, but my doing so can hardly count as good, and doing so only out of duty highlights a questionable disposition. A certain level of emotional sensitivity and sophistication is needed to get by even in everyday life, and a great amount of experience and good judgment is required when it comes to public affairs. Serving public organisations like universities, NGOs, ministries, corporations, or political parties is a real test of one's capacities. Still, even

and duty, he is acting in such a way and deserves moral praise.

decisions made in such environments might be simpler than decisions made in war.

Regarding voluntariness, we can say that an agent acts voluntarily when the agent is free from duress and coercion while deciding to act and acting and the agent does not act in ignorance of what they are doing.[33] This notion of voluntariness captures that voluntariness is primarily defined negatively and requires freedom from specific types of social, political, violent, and other pressures and that the agent has to be knowledgeable about what they do to some extent. It also highlights that—differently from intention—voluntariness is partly about external factors rather than the agent's psychology, and nevertheless it requires that the agent have an adequate psychological life and abilities. Agents need to be able to understand whether they are subject to the relevant forms of external pressures negating voluntariness, and they have to be able to suspect that they might lack important information in a given situation to act rightly.[34] This notion of voluntariness highlights that agents need to be able to recognise their own interests, what they want, and when external pressure is applied to them, going against their own preferences. They also need to be able to gather knowledge regarding information that might be relevant and important in the contexts in which they act. This is why people receive training when they start a new job and why in several countries couples have to attend preparatory sessions when they are expecting a child.

Based on the discussion so far, AAs can only be regarded as acting intentionally and/or voluntarily if they have the right kind of perceptual, motivational, emotional, and reasoning capacities and these play a role in their activities. Certainly, AAs can be ignorant (lack relevant information). Beings that can act intentionally and voluntarily need to have capacities involved in recognising, choosing and revising goals, weighing up competing needs of others,

[33] Cf. Hyman (2015: 77). Note the similarity of this notion of the voluntary to Aristotle's, worked out in the *Nichomachean Ethics* III 1, 5, and V 8, and the *Eudaimonian Ethics* II 6-9. As Ursula Coope (2010: 439) summarises it, according to Aristotle, "1 an action is not voluntary if it is forced (1110a 1ff.); and 2 an action is not voluntary if it is done in ignorance of the particular circumstances of the action (1110b18ff)." However, this issue is tricky, since, as David Charles (2017: 13-4) argues, in many respects Aristotle's usage of voluntary resembles most closely contemporary ideas of intentionality.

[34] The characterisations I offer here capture the main aspects highlighted by Elizabeth Anscombe (1963), Davidson (1980), and Hyman (2015). Contra Hyman, I think that voluntariness is not entirely dependent on external circumstances. For his notion, see Hyman (2015, chapters 3-7).

reconciling clashing views, recognising sinister interests, and reasoning about how to achieve their goals. They also need capacities to have values, desire things, deliberate, plan, resolve inner conflicts, reflect on their own reasons, recognise facts as reasons, exercise empathy, and care for things.[35] They need to be able to know when to take the needs of others into account. At the moment, there exist no AAs with such abilities. Even the most intelligent AIs today are intelligent mostly in the sense of excelling at some domain-specific tasks in which they can rely on large amounts of energy, good-quality and precisely parsed data, and expert setup by engineers, programmers, mathematicians, and others. Nevertheless, there does not seem to be any conceptual or theoretical obstacle that would render it impossible that AAs having such capacities can be designed and deployed one day.

In order to be truly human-like, AAs would also need to understand praise and blame.[36] They must not only be appropriate targets for praise and blame, but they must be such that they themselves can blame and praise others. For this, they have to have participatory reactive attitudes themselves.[37] What does it take to blame someone? George Sher offers a helpful elucidation

> Given that anger, hostile gestures, reprimands, and the rest are so often associated with blame, we may reasonably suppose that anyone who blames someone must at least be *disposed* to react to him in each of these ways. This raises the possibility that what blaming someone adds to believing that he has acted badly or is a bad person may simply be the presence of the corresponding dispositions. (…) each such disposition is explicable in light of a single type of desire-belief pair—a pair whose components are, first, the familiar belief that the person in question has acted badly or has a bad character, but also, second, a corresponding desire that that person *not* have

[35] I'm not arguing here that such AAs would be able to solve the symbol-grounding problem. However, I assume that an AA that is human-like to the degree required to possess all the capacities and abilities needed to be morally responsible would likely be able to tackle this problem too. On some attempts to deal with the symbol-grounding problem, see Kukita (2014).
[36] Paul Russel argued that it is necessary to understand the moral sentiments and reactive attitudes, while Dana Nelkin has endorsed a similar but weaker approach to this issue, proposing that emotional capacities need to inform our reasoning, but in some rare or exceptional cases even without this it might be possible to understand responsibility. See Russell (2004) and Nelkin (2011, esp. chapter 2).
[37] Strawson (1963/2003: 81, section VI).

acted badly or *not* have a bad character. (Sher 2006: 14-15)

Sher's account is not universally accepted, but at the moment I'm not discussing the precise nature of blame; I'm looking for clues as to what kind of capacities it involves. Sher's view gives us some interesting clues, and while he offers a reductive analysis of the dispositions involved in blaming in terms of a desire and a belief, these attitudes themselves are such that having them requires further capacities. It requires the agent blaming someone else to correctly identify that person as appropriate for attributing blame to them, appropriate for holding them morally responsible, and this involves that one can identify that the other person has the relevant capacities. It also requires a good understanding and evaluation of the given situation, good moral judgment, which, in turn, relies on knowledge of values and, arguably, also on caring for them. This is not an exhaustive list of all the capacities that Sher's view requires the blaming agent to possess. What this list does is indicate that an AA that can bear moral responsibility will possess a large number of the core capacities and abilities that underlie our social life. Such AAs would then most likely be able to become embedded in social life and take a role in it.

4. The Problem of Too Close a Resemblance

So far, I have argued that AAs can only be agents, act, and be held morally responsible if they possess the right capacities. They need to be such that they can act intentionally, voluntarily, understand blame and praise, and themselves be blamed and praised. If an AA met all these criteria, it would be human-like. I have furthermore stipulated that in case we would like to have autonomous weapon systems that can operate without human operators managing them during targeting and shooting at targets, we need to construct AAs to be able to bear moral responsibility.

This line of thought highlights a problem for advocates of AAs for war. The problem is that there are no good reasons to create a being substantially similar to humans solely for the purpose of then commanding it to go to war. Since such AAs would have the capacities of valuing, caring, empathy, conceptual thought, and reflexivity and would be able to feel and have phenomenal consciousness, it

would be wrong to ignore their pains, suffering, and fear and to send them into extremely dangerous situations that often lead to death.

A connected concern is that while human soldiers can have several purposes in life—earning money to support their family, pursuing an ideal, serving their country, funding a business after their service, or pursuing their hobby—AAs that were created solely to be deployed in war would not have goals like these. To manufacture them and then sentence them to such a barren one-dimensional life would, again, be hard to justify. Furthermore, the high level of resemblance of AAs and humans would lead to the replication of some of the typical problems of waging war. If the psychology of AAs were morally responsible, then just like humans do, they might face issues of akrasia, moodiness, boredom, fear, obsessions, anger, urges to revolt or flee, temptation, corruption, and other debilitating, performance-lowering problems.

An example of how small differences in the volume of information handled and in the methods of processing that can make changes to the overall behavior of the system comes from the neuroscience of memory.[38] There have been experiments trying to supplement damaged areas of the brain that are responsible for creating memories with artificial parts to restore the capacity of humans to properly form long-term memories. Recently, this technology has been tested and shown to also improve the memory of participants by 37% in recall tests. One of the exciting aspects of thinking about such cases is that they highlight the riddles of information handling and volume. The right amount and coding of information are important for the workings of humans qua the kind of beings that they are. Otherwise, the memories will not be the kind of information that memories normally are or no performance improvement will take place. What we can take away from this for the case at hand is that if we just expect the outer behavior of AAs to be similar to human behavior, without paying attention to how the behavior was produced, then we might be ignoring the fact that there is no reasoning process behind their behavior—a very different process from our reasoning produces their behaviour—and they lack adequate grounds for being responsible. This is a problem because it cannot be guaranteed that different systems from ours will continue to behave like morally responsible agents normally do. One of the remarkable things about human behavior is the mix of

[38] Hampson 2018.

fixed and changeable dimensions of behavioural patterns and abilities. In learning and in becoming skilled at something—be that at painting, cooking, or fighting—human learning follows fixed patterns. This is why schools, trainers, and cooking schools work well. At the same time, individual creativity and ideas can give unique twists and introduce new methods to doing things. This combination of flexibility and capacity for improvement, and of fixed patterns and behaviours, is something that we all know well from our own lives. All our social practices are informed by these aspects of our being, including law, education, and caring practices like child rearing or caring for our parents when they are old. If the internal functioning of AAs is different to such a degree that they learn and adapt in different ways, that might again create problems for treating them as morally responsible. And even quite small differences might cause big divergences in behavior.

Building tools that process information more efficiently likely also means processing information in a different way than humans usually do. Robert Hampson's experiment was successful partly because he and his team could tailor the functioning of the neural prosthetic to the neural activity patterns of the subjects. They did not change the way the neural system of subjects worked; they simply helped to enhance the already existing processes. By analogy, if human-like AAs would be deployed autonomously in wars, they would not have a substantial edge over human fighters simply in virtue of being artificial. To actually make them more efficient would involve creating different information-processing modes from those that we rely on. AAs could have an advantage in how efficiently their capacities and abilities work, but their performance would still be in the human range.

We will almost certainly be able to develop in the future AAs that outperform humans in a wide range of activities. At the same time, if such AAs are not human-like, then we might simply not understand what they are doing and why. Such creatures might function in significantly different ways from us. Letting such creatures wage war is an unacceptable idea, since they cannot be held accountable. We cannot explain their behavior, which means that we do not understand their goals and reasoning. The risk in arming such AAs that are also efficient in war is enormous.

I will briefly mention one last concern raised by creating human-like AAs for

the purposes of war. Presumably, we would want to model them on the biological and psychological profiles of humans who have done well in wars. What kind of people do well in combat? In command? As a pilot? And are the people who do well usually also people who act morally well? By what standards do they do well? Do the bombing sprees of the more efficient pilots also systematically result in higher civilian casualties? Do effective sergeants do more lasting psychological damage to recruits? Are highly aggressive people better at certain types of military tasks?

Taking into account how difficult it is to pin down "do well in war" and to spell out how people doing well in wars do in society for the rest of their existence, the following emerge as the key questions: Would people with solid moral principles do well in military settings? Would virtuous agents carry out orders well? Are people who do well in the army happy? Are their well-being levels normal? It seems intuitive that it is another reason against creating human-like AAs that we should not create aggressive, morally conformist, or morally inferior agents. Also, we should not create human-like agents in case they are miserable and their quality of life is below the threshold deemed desirable. I do not offer answers to the questions I have posed in this paragraph. Most likely the human resources divisions of militaries have data on the correlations between biology, psychology, and performance. The point of posing these questions here is to highlight how many problems the human-like AAs might encounter in having to serve in the military.

One might counter that I made a shallow point, since human soldiers face the same issues, or arguably even worse, since they might have enjoyed higher life qualities during most of their previous lives than during their time serving in a war. While this is true, those humans do not end up in the military without any preliminaries leading them there, nor do most of them have to spend the rest of their lives in service. I will not touch here on the point whether it would make sense to manufacture AAs that would lead full human lives, with a period of military service. This question poses different issues and is a separate topic.

In this section, I provided some considerations against manufacturing and deploying human-like AAs for the purposes of war. I did not rehearse standard Kantian and other deontological arguments, nor the often-repeated utilitarian arguments. My goal was to tie in the points I made with the preceding discussion

regarding the capacities and abilities on which agents rely when they act intentionally or voluntarily. In the next section, I will address potential objections.

5. Objections Addressed

Two main types of objections could be raised against my position. Some objections try to show that it is actually good if the AAs deployed in war are different from humans. Other objections try to deny that human-like AAs are possible and claim that arguing against their possibility is not important. I reject both of these types of argument. I think the first type of argument gets something right: AAs that are not like humans can be beneficially employed in many roles in wars. But they cannot be allowed to make substantial morally charged decisions, including decisions to cause harm without the real-time management of a human. The second type of objection relies in most cases on too-demanding ideas about what counts as a morally responsible agent. Maybe some libertarians would claim that unless one can have a capacity for self-determination and hence for freedom, one cannot have moral responsibility. I addressed such worries in section 2. I would also want to add at this point that there are several convincing views of moral responsibility that do not depend on libertarian ideas. I tried to give a rough sketch of the capacities and abilities AAs need to count as morally responsible. If I was moderately successful, then the worry that they could be misused for purposes of war in morally bad ways is real.

Another objection would go like this: Using AAs in war has an advantage over deploying human soldiers exactly because AAs reason differently. Military law does not treat soldiers as persons in the same way we normally treat civilians as persons: soldiers are *not* supposed to make decisions on their own, follow their own values consistently, and so on. A good soldier—especially a lower ranking one—should follow orders. Employing humans as soldiers poses all kinds of issues: there are several historical examples of having to force soldiers to attack, soldiers fragging their officers to avoid having to follow them into a charge, and so on. In this sense, if AAs do not have normal reasoning processes, that is an advantage. They can follow orders more easily and more efficiently, without this causing them emotional distress and harm, and without causing trouble for their forces.

I do not claim that denying some autonomy for AAs is bad. As I said, AAs that are deployed for well-described tasks of limited scope—and if that scope includes potentially lethal engagements that are managed in real time by human operators—can be of much use to any military. What I claimed was that a human-like agential constitution—in line with the view of agency introduced in section 2—would be needed for AAs to be such that we can allow them to make decisions regarding killing. Support AAs and other systems could be fruitfully developed and deployed without exemplifying substantial resemblance to human psychology in case they do not take any direct harmful actions against humans.

Another objection would say that the concept of a person is tied to the notion of *birth* and through it to the notion of *life*. Something not born cannot hence be a person, and therefore AAs cannot be the kind of thing that would be morally autonomous anyway. So, no matter what their constitution is or how they function, they can never be allowed to make decisions about life and death. A different version of this objection would give up on the idea that AAs that can make decisions about life and death have to be exactly like human agents in their origin. It would instead say that the specific hormonal processes, childhood experiences, memories, and so on that humans typically have while growing up all play a fundamental role in our becoming persons, that is, becoming the kind of morally competent agents who can be allowed to make decisions about killing in war. The objection would go on claiming that something that has a different background—biological, psychological, social, cultural—cannot be a person, and as such, it cannot be a being that reasons sufficiently similarly to us to be allowed to make weighty moral decisions. Hence, moral responsibility cannot be attributed to AAs even if they have human-like capacities.

The position I advocated here can reply that what is demanded for something to be morally responsible is not that it should be exactly like a human person, including the complete historical background of typical healthy adult humans. Rather, what is demanded is that the reasoning and decision-making processes of such agents be like that of humans. This does not require them to be persons in the sense of having a typical human birth and childhood. Being human requires more than being morally responsible. For example, if we would want to understand what human persons are, we would need to go beyond moral responsibility and also offer an account of their autonomy. But as I have argued,

this is not essential for military AAs.

A further objection that could be made is that particular processes of reasoning, such as that of weighing options like when to kill in a combat situation, can be programmed well without having to provide an AA with a set of values and the ability to grasp that those are its values, and without the ability to recognise facts as moral reasons or to rationally reason to conclusions while taking emotional and moral facts into account. AAs that do not do these things—one could think of the automated machine gun turrets of the South Korean military in the border zone with North Korea—can actually be more efficient than humans in choosing whom to target. They are not swayed by emotion and other biases that humans are subject to.

To deflect this objection, it can be noted that it is unlikely that without representing the information that certain values are their own and that such values can be shared by other beings, AAs cannot treat values in a way similar to our reasoning processes. This makes it clear that they are not morally competent agents who are also more efficient than us. Rather, they lack any grasp of values. What guides their behavior then? Pre-programmed preferences and goals, which they cannot reflect on, unlike humans. I argued in section 4 that deploying agents that have non-human reasoning and permission to act without real-time management by humans is inherently dangerous. Taking such a risk is not justifiable.

There might be some exceptions, namely those cases in which a machine is reasoning in a non-human way, but its reasoning—whom it targets, when it fires, etc.—is clearly and completely understood by its operators, and it can be explained to others—law enforcement, international tribunals, expert committees—in terms that make the responsibility of its operators clear. This is easiest to achieve when such military systems are simple and set up with accountability in mind. My paper says nothing against the deployment of such weapon systems. Criteria of their ethical operation depend on the transparency of their operation and the precise calibration of their range of deployment. Such machines will inevitably be able to perform only much more primitive tasks than any human. AAs and other machines that are complex and at the same time reason in different ways from humans do not meet such criteria.

In this section, I answered a number of anticipated objections and questions

that the position I argued for could face. This concludes my discussion of the topic at hand, which is whether it is useful to develop robots that can kill in war without human supervision and management. The conclusion of my paper is that it is not permissible to do so. The reason is simply that such robots would need to be too human-like to make it permissible for them to operate without constant human supervision. Otherwise, they could not have moral responsibility, and nothing lacking moral responsibility should be created to make decisions about killing. I argue for this by showing that to be morally responsible—the right kind of agent to make life-and-death decisions—requires the possession of several high-level capacities. Nothing that is dissimilar from humans—more or less complex, or functioning substantially differently—has this status. The richness of the agential capacities and abilities needed to be morally responsible is such that it makes AAs that meet this criterion too human-like to manufacture for the sole purpose of deploying them in war. Doing so would amount to creating slaves doomed to miserable lives, which is clearly impermissible.

6. Conclusions

For AAs to be candidates for morally responsible agency in combat situations, they would need to resemble the functioning of healthy adult humans in great many respects than they currently do or as is a morally good thing to grant them. The deployment in wars of AAs that would be sufficiently human-like to be allowed to make autonomous choices, without the involvement of human supervision, would pose several issues for law and morality. Human rights issues would emerge, and their human-like functioning could lead to psychological and performance problems too, in the same way it does in the case of human soldiers deployed in war. Hence, their deployment might offer no moral or other benefit over deploying humans, while producing downsides, like the questionable employment of human-like beings for the sole purpose of war.

At the same time, we could not attribute moral responsibility and hence allow AAs that are not human-like to make decisions about killing if they are not relevantly similar to human agents, except for some simple agents, the workings of which can be fully grasped, operating in strictly restricted domains. Whether it would make sense to create and deploy such AAs for offensive purposes in wars

is not clear at all.

At the same time, AAs that would reason like humans would not do substantially better than humans on specialist tasks. This would undermine the purpose of using them in the first place. AAs are best not used at all without real-time supervision by humans in cases when moral decisions concerning harm and killing are involved. Where AAs can be made best use of is in well-defined, specific roles, where their dissimilarity from humans can be an advantage and can be utilized fully. AAs should then be mostly complementary systems for humans, compensating for their weaknesses. There are several roles in militaries where such AAs could be made good use of, such as in reconnaissance, where being able not to move and to tirelessly observe can be an enormous boon; in sweeping minefields, where no complex reasoning is needed; as medical assistants that are not vulnerable and might be able to carry large amounts of supplies and approach targets under fire; or in supplementary data analysis and advisory systems. If any uses of AAs would be encouraged, they should be that of non-lethal support AAs.

* The author is thankful for the support and funding he received during his tenure as a JSPS Postdoctoral Fellow at Keio University, Tokyo, Japan (ID No. P17783), and as a contributor to Dr Tamas Demeter's project "Values in Modern Science: Epistemic, Social, Moral" at the Hungarian Academy of Sciences.

References

Alvarez, M. and Hyman, J. (1998). 'Agents and their Actions.' *Philosophy* 73 (2): 219-245.

Anscombe, G. E. M. (1963). *Intention*. 2nd ed. Oxford: Blackwell.

Braithwaite, V. (2010). *Do Fish Feel Pain?* New York: Oxford University Press.

Burt, P. (2018). *Off the Leash. The Development of Autonomous Military Drones in the UK*. Drone Wars UK. Accessed: 13/11/2019. https://dronewarsuk.files.wordpress.com/2018/11/dw-leash-web.pdf

Chan, M. K. (2019). 'China and the U.S Are Fighting a Major Battle Over Killer Robots and the Future of AI.' *Time*. Accessed: 13/11/2019. https://time.com/5673240/china-killer-robots-weapons/

Charles, D. (2017). 'Aristotle on Agency.' *Oxford Handbooks Online*. Online publication date: May 2017. Accessed: 14/11/2019. DOI: 10.1093/oxfordhb/9780199935314.013.6

Coope, U. (2010). 'Aristotle.' In T. O'Connor and C. Sandis (eds.). *A Companion to the Philosophy of Action*. Singapore: Wiley-Blackwell.

Darwall, S. (2006). 'The Value of Autonomy.' *Ethics* 116: 263-284.

Davidson, D. (1980). *Action and Events*. Oxford: Oxford Clarendon Press.

Fischer, J. M. (2010). 'Responsibility and Autonomy.' In T. O'Connor and C. Sandis (eds.). *A Companion to the Philosophy of Action*. Singapore: Wiley-Blackwell.

Frankfurt, H. (1969/1988). 'Alternate Possibilities and Moral Responsibility.' In H. Frankfurt. *The Things We Care About.* Cambridge University Press.

Frankfurt, H. (1971/1988). 'Freedom of the Will and the Concept of a Person.' In H. Frankfurt. *The Things We Care About.* Cambridge University Press.

Frankfurt, H. (1978/1988). 'The Problem of Action.' H. Frankfurt. *The Things We Care About.* Cambridge University Press.

Fryer-Biggs, Z. (2019). 'Coming Soon to a Battlefield: Robots That Can Kill.' *The Atlantic*, published on: 03/09/2019. Accessed: 13/11/2019. https://www.realclearpolitics.com/2019/09/03/coming_soon_to_a_battlefield_robots_that_can_kill_485045.html

Glaser, A. (2016). 'The UN has decided to tackle the issue of killer robots in 2017.' *Vox*, published on 16/12/2016. Accessed: 13/11/2019. https://www.vox.com/2016/12/16/13988458/un-killer-robots-elon-musk-wozniak-hawking-ban

Glover, J. (1970). *Responsibility*. London: Routledge and Kegan Paul.

Gubrud, M. (2015). 'Semi-autonomous and on Their Own: Killer Robots in Plato's Cave.' *Bulletin of the Atomic Scientists*, published on 12/04/2015. Accessed: 13/11/2019. https://thebulletin.org/2015/04/semi-autonomous-and-on-their-own-killer-robots-in-platos-cave/

Hampson, R. E. (2018). 'Developing a Hippocampal Neural Prosthetic to Facilitate Human Memory Encoding and Recall.' *Journal of Neural Engineering* 15: 036014. DOI: 10.1088/1741-2552/aaaed7

Hyman, J. (2015). *Action, Knowledge, and Will.* Oxford: Oxford University Press.

Jenkins, R. and Purves, D. (2016). 'A Dilemma for Moral Deliberation in AI.' *International Journal of Applied Philosophy* 30 (2): 313-335.

Kukita, M. (2014). 'Can Robots Understand Values?: Artificial Morality and Ethical Symbol Grounding.' *Proceedings of 4th International Conference on Applied Ethics and Applied Philosophy in East Asia* Feb. 2014: 65-76.

Mele, A. R. (2017). *Aspects of Agency*. *Decisions, Abilities, Explanations, and Free Will*. New York: Oxford University Press.

Misselhorn, C. (2018). 'Artificial Morality. Concepts, Issues and Challenges.' *Social Science and Public Policy* 55: 161-169. DOI: 10.1007/s12115-018-0229-y

Nelkin, D. (2011). *Making Sense of Freedom and Responsibility*. New York: Oxford University Press.

Pink, T. (2010). 'Free Will and Determinism.' In T. O'Connor and C. Sandis (eds.). *A Companion to the Philosophy of Action*. Singapore: Wiley-Blackwell.

Purves, D., Jenkins, R. and Strawser, J. B. (2015). 'Autonomous Machines, Moral Judgment, and Acting for the Right Reasons.' *Ethical Theory and Moral Practice* 18: 851-872. DOI: 10.1007/s10677-015-9563-y

Rey, S. et al. (2015). 'Fish Can Show Emotional Fever: Stress-induced Hyperthermia in Zebrafish.' *Proceedings of the Royal Society B: Biological Sciences* 282 (1819): 20152266. DOI: 10.1098/rspb.2015.2266

Russell, P. (2004). 'Responsibility and the Condition of Moral Sense.' *Philosophical Topics* 32: 287-305.

Setiya, K. (2010/2017). 'Sympathy for the Devil.' In K. Setiya. *Practical Knowledge*. New York: Oxford University Press.

Sharkey, N. (2010). 'Saying 'No!' to Lethal Autonomous Targeting.' *Journal of Military Ethics* 9 (4): 369-383.

Shepherd, J. (2018). *Consciousness and Moral Status*. New York: Routledge.

Sher, G. 2006. *In Praise of Blame*. New York: Oxford University Press.

Sparrow, R. (2007). 'Killer Robots.' *Journal of Applied Philosophy* 24 (1): 62-77.

Steward, H. (2012). *A Metaphysics for Freedom*. Oxford: Oxford University Press.

Strawson, P. F. (1963/2003). 'Freedom and Resentment.' In G. Watson (ed.). *Free Will*. Oxford University Press.

Talbot, B., Jenkins, R. and Purves, D. (2017). 'When Robots Should Do the Wrong Thing.' In P. Lin, R. Jenkins and K. Abney (eds.). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press.

Von Wright, G. H. (1963). *Norm and Action*. London: Routledge, and Kegan Paul.

Watson, G. (2003). 'Introduction.' In G. Watson (ed.). *Free Will.* New York: Oxford University Press.

Zardai, I. Z. (2016). *What Are Actions*. PhD Thesis, defended at the University of Hertfordshire. https://uhra.herts.ac.uk/handle/2299/17222

Zardai, I. Z. (2019). 'Agents in Movement.' *Mita Philosophical Association Journal* 143: 61-84.

Zardai, I. Z. (2022). 'Making Sense of the Knobe-effect: Praise Demands Both Intention and Voluntariness.' *Journal of Applied Ethics and Philosophy* 13: 11-20.

Journal of Philosophy of Life Vol.13, No.1 (January 2023):97-111

Wheat and Pepper
Interactions Between Technology and Humans

Minao Kukita*

1. Introduction

Picture a barbed wire. It is essentially a wire with spikes, which was invented by an American businessman to keep livestock enclosed. However, the influence that this seemingly minor invention had on humans and the natural environment has been enormous to say the least. Many farmers in America used it because it was inexpensive and easy to install. Consequently, the enclosed animals ate up all the grass, thus leading to desertification of vast amounts of land. State power also used it to exclude, isolate, and repress certain groups of people. The barbed wire placed around the trenches of World War I made the battle all the more long and arduous [16].

Or think about the AK-47, a rifle invented in the former Soviet Union. The only significant difference between the AK-47 and other rifles was that it had a greater level of tolerance (i.e. an acceptable range of deviation of the size of each part from the standard). However, this improvement increased its reliability under harsh conditions, endurance, and ease of production. It was used not only in the former Soviet Union, but all over the world. Terrorists and criminals have also used it and it has come to be called a ‘small weapon of mass destruction’.

Could the inventors of the said objects or their early users have ever anticipated such results?

People’s opinions on artificial intelligence (AI) vary from extreme optimism (i.e. AI can address every human problem, including starvation, diseases, energy, climate change, etc.) to extreme pessimism (i.e. AI will dominate or exterminate all humanity). However, as for long-term influences at least, it would not be reasonable to bet on any one of them and disregard the rest. There is a moderately optimistic view that, given what current AI can actually do, it would be ridiculous

* Associate Professor, Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, Aichi Prefecture, 4648601 Japan.
Email: minao.kukita[a]i.nagoya-u.ac.jp
** This paper is an English translation of [15].

to worry about it. While I may accept the premise that the current level of existing AI is not so great, I am not convinced with the conclusion that it will not have any serious influence.

Society is complex. Introducing a great number and variety of autonomous AI devises into it may be like introducing unknown species into an ecosystem. It would be very difficult to predict its long-term effects. Therefore, every conceivable possibility must be considered in order to be prepared to promptly and adaptively deal with any unexpected results.

Although I may have previously used the ecosystem as a metaphor, its truth goes beyond being metaphoric. Humans and technology literally have a symbiotic relationship despite the latter playing a parasitic role at times. The prevailing belief that technology is nothing more than a tool that humans use to fulfil certain intentions or goals fails to capture this important aspect of technology.

This article examines the symbiotic and parasitic relationship between technology and humans; and considers the potential impact that AI could have on the latter in terms of social relationships, in particular.

2. Why Do Humans Use Tools?

'Why does a dog wag its tail? Because a dog is smarter than its tail. If the tail were smarter, the tail would wag the dog'. This was a joke made at the beginning of the film *Wag the Dog* directed by Barry Levinson. In light of the topic at hand, one could say `Why do humans use tools? Because humans are smarter than tools. If tools were smarter, the tools would use the humans'. However, this is not a joke by any means. Indeed, humans are often used by tools precisely because humans are not smart enough.

The obvious reason people use tools is to effectively accomplish certain tasks. Naive theories of technology presume a master-subject relationship between humans and technology, respectively; and the latter is reduced to a means to an end. However, the relationship between humans and technology is not that simple.

The phrase 'symbiosis with technology' is sometimes used and in most cases, such phrases are rhetorical expressions intended to mean 'utilizing technology well' or something to that effect. Nevertheless, these phrases convey deeper truths beyond being mere metaphors.

Richard Dawkins, an evolutionary biologist, believes that cultural products such as ideas, expressions, patterns of behaviour, etc. are also subject to the

processes of replication, mutation, and selection; and evolve and spread as a result. He coined the term 'memes', which refers to the units of cultural information that are propagated. Technological products, too, are considered as memes. Actually, individual artefacts are produced and used according to certain memes. For example, earthenware pots are produced according to the meme of 'kneading a certain kind of clay into a certain shape and heating it with fire'. Through memes, the relationship between humans and technologies may be viewed as a complicated symbiosis where biological genes and artificial memes make use of each other in order to increase the chances of spreading themselves.[1]

Generally, if humans find a particular type of technology useful, its meme tends to spread; whereas if it is found useless or harmful, its meme tends to vanish. Therefore, humans and technology can be viewed as reciprocating benefits to each other. However, the relationship between humans and technology is not always friendly or mutualistic. The promotion of every individual's well-being and the overall progress of the human race are the goals of the greatest priority. Some technological products, however, spread despite having little to no contribution—or even a negative influence—on these goals. Drugs such as heroin or LSD are typical examples of this. Although they can be beneficial, they have a strong bias towards abuse. Such technological products 'hack', so to speak, human physiology or psychology—take advantage of people's vulnerability in these areas—and drive humans to depend on them. In this way, they 'parasitise' humans and thereby succeed in survival and reproduction.

On the other hand, some technologies employ subtler strategies than drugs, which directly take advantage of human physiology. Such examples will be discussed in the following section.

3. Two Parasitic Technologies

3.1. Agriculture

The historian Yuval Noah Harari has made some interesting remarks on agriculture. He asserts that the Agricultural Revolution did not necessarily improve the standard of living of the ancient human beings who had been making a surviving through hunting and gathering. Hunter-gatherers lived on various

[1] Note that, in saying so, genes and memes are anthropomorphised. Furthermore, neither genes nor memes have any intentions or goals; only 'blind' and mechanical processes exist.

kinds of food such as nuts, fruits, animals, and fish. Agriculture, however, prompted them to depend on fewer kinds of food, particularly grains, and this led to an imbalance in their nutrient intake. They were also at a risk of serious starvation in the event of unfavourable weather. Labour for food became more tedious and time-consuming than before. Domestication of animals also brought about a variety of infectious diseases such as malaria, smallpox, tuberculosis, measles, and influenza. All things considered, agriculture has done far more harm to human beings than good, thus making the Agricultural Revolution 'history's biggest fraud' ([6], p. 79), according to Harari.

This then raises the question, why did ancient humans develop agriculture in the first place? The answer is simple. They did it because it increases the production of food per unit area. Agriculture created a large-scale society of settlers. Once such a society was built, it became increasingly difficult to return to hunting and gathering because the food required to support the population could no longer be acquired through this method. Hence, humans took on a more inconvenient life and grew crops in exchange for a larger population. Harari adds, 'We did not domesticate wheat. It domesticated us'. ([6], p. 81). From the point of view of genes, this is acceptable. Genes are only concerned about spreading their copies, regardless of whether the individual carrying them lives a happy or miserable life. The meme of agriculture took advantage of this interest of genes. Moreover, since population further increased alongside the increase in food production, people demanded more and more food. This spurred the endless race between food production and population growth'.

Harari does not claim that the life of civilised, agricultural people rooted in agriculture is more painful than the life of ancient hunter-gatherers. He simply compares the lives of early farmers to that of the hunter-gatherers before them. For those benefiting from civilisation today, the ancient people's decision to begin agriculture may appear to be a good one. However, as Harari points out, the famine and harsh labour that the ancient people had to go through must not be overlooked. The benefits that modern people enjoy today do not compensate for their suffering.

This could be applied to this day and age. Optimistic people would often say that although new technology may initially cause various difficulties, confusion, and pain, it will ultimately improve general human living standards in the long run. This may be true; however, it would not be logical for them to endure pain for the benefit of future humans.

3.2. Military

Another important by-product of the inception of agriculture is the emergence of the war industry. Needless to say, violence and looting also occurred between tribes of hunter-gatherers. Based on abundant archaeological, historical, and anthropological evidence, Steven Pinker concludes that the lives of ancient hunter-gatherers were much more violent than originally considered (Pinker, [11]). However, agriculture can be credited for turning warfare into an 'industry'. For if there was no surplus of food, then plundering others would not have been a good way to get the food needed to survive. As Pinker says, it was the fear that the opponents might attack themselves as well as revenge for past violence inflicted by the other party that drove the hunter-gatherers to violence.

Social classes specialising in the exploitation of others' labour were borne only after agriculture began and caused a food surplus. When people went beyond their communities and exploited other communities, industrialised wars ensued. The historian, William H. McNeill, characterises armed forces as `macroparasitism' (McNeill [9]). From another point of view, it can be said that the meme (or a system of various memes) of military came to stand in a parasitic relationship with the meme of agriculture. Here, the term 'military' refers to a wide range of war-related activities such as actual combat; strategies and tactics; logistics; development and manufacture of weapons; recruitment of soldiers; and military institutions.

Early troops literally devoured the agricultural areas. Violence against the settlements of agricultural people came in the form of either raids by nomadic people or expeditions by other settled people. For example, King Akkad of Sargon, who dominated the entire Mesopotamia at around 2250 BC, advanced with his army and devastated all the agricultural zones in his path. It took several years or even decades to undo the damage they had caused (*ibid*.).

Subsequently, the relationship between agriculture and war became friendlier. This was because the systems of administration developed as well as food supply and stock accompanied by military action were effectively carried out. For example, during the Greek invasion from 480 to 479 BC, Xerxes, the then Persian king of the Achaemenid dynasty, ordered his subordinates to collect food at reservoirs set up along the course of his army. In doing so, Xerxes successfully moved a larger army than that of Sargon's without devastating agricultural zones

along the way (*ibid*.).

War and agriculture, especially the livestock industry, gradually became even more active accomplices. When European countries competed with each other to build colonies all over the world, sheep and cattle were of great value. Wool was needed for the textile industry and cheap beef was necessary to keep workers' wages low. Therefore, countries often waged wars as a means to acquire more pastureland. On the other hand, the livestock industry directly contributed to the war effort by providing horses and salted meat etc. for military use. The sociologist, David A. Nibert, recounts the collaboration process between the war and the livestock industry during this time to promote ferocious capitalist colonial rule (Nibert [10]).[2]

After the two world wars, wars became obsolete as a means of expanding colonies or efficiently robbing other countries of territories and resources, in general—for most of the developed countries, at least. The law of war prohibits war based on such selfish motives. Going against international rules is too high of a risk, given the modern international political and economic structures, which mutually connect numerous countries in a complex relationship. Nevertheless, the world will never be free of wars; there will always be wars and conflicts somewhere. Countries heavily invest in their own militaries and, in turn, military technology makes remarkable progress on a daily basis.[3]

Just as the ancient people before the industrialisation of warfare, primitive fear and hatred continues to drive people to start wars (or prepare for them). According to recent psychological findings, human mechanisms of social cognition and emotion encourage empathy towards members belonging to the same group, as opposed to those outside the group (Bloom [1] and Greene [4]). These are most probably mechanisms that humans have developed throughout the course of evolution, when people were divided into small groups or tribes and being completely disconnected from any other tribes. During that time, such mechanisms were major factors in the survival of the said groups.

These mechanisms, however, hinder people, who realise the greater benefits of joining forces, from cooperating with people from other groups. They naturally lean towards discrimination, opposition, and conflict with different groups and

[2] Nibert, as well as Harari, thinks that agriculture made humans—and other animals involved in agriculture—much more miserable.

[3] The Stockholm International Peace Institute estimates that 2015 global military expenditure was approximately $1.6 trillion, which is roughly 2.2% of the world's GDP.

believe that it is ‘rational’ to alert the opponent; prepare for the latter’s betrayal and attack; and in some cases, engage in pre-emptive betrayal or attack. When a country is trapped in this mindset, the perverse logic of developing destructive weapons to overwhelm the opponent in order to maintain peace is validated. Taking advantage of such human psychology, military technology has advanced and expanded globally. It is obvious that if weapons are fully eradicated from the world, the world will be safer and more peaceful. However, it would be extremely difficult for any country to be the first to give up its weapons.

4. A Security Hole in Human Psychology

The previous section examined the histories of agricultural and military technology and saw that they did not necessarily contribute to human welfare. This chapter returns to the main subject, artificial intelligence, and explores the possibility of AI hacking humans.

There is a great variety of existing artificial intelligence with an equally wide range of technical details and applications. It would not make much sense here to discuss the self-driving car and the AI music composer at the same time. Instead, this chapter focuses on the so-called ‘social robots’, the robots and AIs that talk with humans and create ‘social relationships’ with them.

4.1. Psychological mechanisms for empathy and cooperation

As mentioned in the previous chapter, human psychology is biased towards empathising with people who belong to the same group. However, this bias is not completely immutable and the distinction between ‘inside’ and ‘outside’ can be rather open to interpretation. For example, one can easily empathise with another if a sense of trust is formed, even upon meeting each other for the first time.

Paul J. Zak, an advocate of neuroeconomics, conducted an interesting experiment using the ‘trust game’ (Zak [14]). The trust game is as follows: The players are given a certain amount of money ($10 in Zak’s experiment). A randomly selected player (Player A) is told to decide how much (possibly none) to invest in another randomly selected player (Player B). Player A determines the investment amount and Player B obtains three times the amount invested by the former. Player B may give some ‘returns’ (possibly none) to Player A. Player A may opt not to invest if he or she thinks that Player B will not repay, while Player

A may invest if he or she expects at least partial repayment from Player B. Therefore, Player A's investment represents the degree of trust with Player B.

Zak collected blood on the spot from the player invested on (Player B) and analysed it. Then, he found that in many cases, the oxytocin level in the extracted blood was higher than usual and that there was a correlation between the oxytocin level and the intention of repayment. Oxytocin is a hormone believed to be associated with sympathy, trust, tolerance, etc.[4] This experiment suggests that when an individual feels trusted by others, he or she will feel empathy for them. There are many other ways to induce empathy, Zak says. The oxytocin level can be raised through simple ways such as embracing, friendly interaction through social media, or being exposed to depictions of other's misery.

Human beings are social animals who cannot survive without cooperating with others. For individuals, finding people to cooperate with and the act of cooperating with these people are literally matters of life and death. Consequently, the philosopher Joshua Greene argues that in the course of evolution, humans have developed the ability to sensitively detect whether others are willing to cooperate or not detect whether others are willing to cooperate or not as well as the ability to act cooperatively against one's self-interest (Greene [4]).

Experiments show that such abilities already exist in young children even before they begin to speak. In an experiment, J. Kiley Hamlin and his colleagues showed a video to infants wherein certain objects made out of simple geometric shapes were behaving as if they had some intentions or purposes. Specifically, a round object behaved as if was trying to climb up a slope and a square object behaved as if it was interfering with the round one. Additionally, a triangular object behaved as if it was helping the round one climb the slope. After watching the video, the infants showed behaviour that suggested that they preferred the triangular object to the square object [5]. An experiment conducted by Yasuhiro Kanakogi et al. showed that individuals who demonstrate behaviour such as helping others under attack are preferred by pre-language children [7].

These experiments are interesting for two reasons. First, they suggest that human beings have an innate ability to identify cooperative behaviour from others and tend to positively evaluate it. Second, these experiments indicate that such ability works at a fairly abstract level. The 'individuals' that appeared in the animation used in these experiments were essentially triangles, circles, and

[4] However, the actual effects of oxytocin still remain uncertain. Moreover, recent results that refute previous experiments on the effects of oxytocin have been obtained.

squares with eyes attached to them. These characters were far from human beings and did not look like any other animal either. Since humans are very sensitive to the cooperative and non-cooperative behaviour of others, they also read the behaviour of non-human objects in the same way.

However, this is a vulnerability of human psychology, a security hole that could be taken advantage of.

4.2. How social robots hack humans

A video introducing the walking quadruped robot developed by Boston Dynamic called 'Spot' includes a demonstration wherein a human kicks Spot hard on the side and Spot withstands it. As it is kicked, it quickly moves its legs to maintain its balance just like a real animal would. The video ends with a disclaimer stating that 'no robots were harmed in the making of this video'. This was probably intended as a joke, but some people did not take it as such. Some of those who saw the video felt that it was unethical to kick the robot. Noel Sharkey, a computer scientist and one of the founders of Responsible Robotics[5], commented on the viewers' reactions, saying that the ethical treatment of robots must be considered only when they should be capable of feeling pain.[6] However, as Sharkey points out, there is also concern that waging violence against robots that closely resemble humans or animals may diminish the sense of aversion to such violence, eventually leading to violence towards actual human beings and animals. Nevertheless, apart from such concerns, if someone is asked whether anything ethically bad was done to the robot, the answer would probably be no. If a watchmaker's employee tramples on a new product to demonstrate its sturdiness, only a few people, if any, would be worried that something unethical was taking place. The different reactions simply result from the fact that the appearance and movement of the robot are similar to living animals, whereas those of watch are not, and this cannot be a significant ethical concern.

Robots that are capable of speaking out further complicate the issue. Just as the behaviour of triangles and squares are interpreted as cooperative, aggressive, etc. and pity for Spot is felt as if it were an actual living animal, most people

[5] Responsible Robotics is an NGO that focuses on ethical issues concerning robots.

[6] http://edition.cnn.com/2015/02/13/tech/spotrobot-dog-google/. Accessed on 21st May 2017. Sharkey's idea is based on the utilitarian principle that 'whether an action is ethically good or bad is evaluated in terms of the amount of pleasure and pain the action caused'.

would feel the heart behind the words from a machine. Psychologist Sherry Turkle says that the only spoken words ever heard throughout the course of history were those from other humans and that current humans are the first humans who need to distinguish human utterances from artificial ones (Turkle [13], p. 342).

In fact, there may not be a lot of people who need to make such a distinction. Turkle reports that many people now tend to avoid real-time, face-to-face conversations and prefer to communicate through text messaging. They think that real-time conversation is riskier because it is difficult to control (i.e. one might say something hurtful, offensive, or that would reveal one's weakness or faults). It might also require too much time and additional costs for others. For them, 'conversation' with machines would be ideal in the sense that the latter do not get hurt or angry; they are not disappointed in one's weakness or faults; and they satisfy people's desire for communication; while enabling them to avoid the 'risks' of communication.

Softbank's CEO, Son Masayoshi says that it is Softbank's goal 'to spread personal robots, thereby increasing happiness and reducing sadness'.[7] Softbank advertises Peppers as a 'robot understanding emotions', an 'emotional robot', or a 'loving robot'. In the promotional video of a home robot called Jibo, developer Cynthia Breazeal asks, 'What if technology actually treated you like a human being? What if technology helped you feel closer to the ones you love? What if technology helped you, like a partner, rather than simply being a tool? [...] And together, we can humanise technology'.[8]

Social robots are most likely to flourish in the field of healthcare or nursing. The seal-like robot, 'Paro', developed by Takanori Shibata, a roboticist at The National Institute of Advanced Industrial Science and Technology (AIST) in Japan, is used in nursing care and holds the Guinness World Record as the 'World's Most Therapeutic Robot'. Fuji Software's conversation robot, 'Palro', is used for recreational activities in elderly homes. In Japan's already aged and ever-aging society, the demand for such robots will only continue to grow. In addition, these robots will also be useful in education. In Jibo's promotional video, there was a scene where it tells a story to a child. On a separate note, some people even believe that human beings will have sexual relations with robots in the future. For example, David Levy, an artificial intelligence researcher, predicts that humans will have sexual relationships with robots in 2050 and that robots could

[7] http://logmi.jp/39604. Accessed on 14th September 2016.
[8] https://www.youtube.com/watch?v=H0h20jRA5M0. Accessed on 1st February 2019.

take the place of human sex workers (Levy [8]). Helen Driscoll, a psychologist, asserts that sexual technology changed social conventions and that sex with robots will be the norm by 2070—turning physical relationships between humans into a 'primitive' act.[9]

5. Pessimism and Optimism

Information and communications technology (ICT) and social media are changing the way people communicate with each other. Based on various social surveys and interviews, Turkle argues that smartphones and social media are also changing the practice and the norms of conversation. For example, she [13] mentions the 'rule of three' among American college students. This rule dictates that in social occasions such as dinners, 'at least three people have to participate in the conversation and not be immersed in their own smartphones or devices. This means that if three people are actively engaging in a conversation, then one is allowed to be absorbed in one's smartphone'. She is concerned that the current social norms allow people to retreat to their own smartphones even when they are in social situations.

According to Turkle, there are various consequences of the predominance of text-based communication via smartphones and people's reluctance to partake in real-time, face-to-face communication. Children will lose the opportunity to learn empathy for others. It will become virtually impossible to communicate in an integral personality. Conversations will become fragmentary and long, complex conversations will be difficult. People will suffer from the 'paradox of choice' after being confronted with too many opportunities for communication.[10] In the end, the multitasking enabled by smartphones is inefficient and reduces both creativity and learning effect. She insists that everyone should admit that they are 'vulnerable' to ICT and must therefore be careful to not allow the excessive use of ICT to impede real-time, face-to-face communication between humans.

While there are people such as Turkle who are concerned about the negative effects of ICT development, philosopher Luciano Floridi is more optimistic about ICT and the future of humanity. He pictures a world where more information is circulated between initially informational people and ICT products coexist with

[9] http://www.mirror.co.uk/news/uk-news/sexrobots-the-norm-50-6190575. Accessed 7th May 2016.
[10] People generally think that better choices can be made if there are more options, but in reality, too many options can lead to lower satisfaction with the result of the selection.

humans as new informational entities (Floridi [3]). Although ICT has various hurdles to overcome, it does not erode human-to-human communication; instead, it enriches human beings and the world. He believes that the bigger problem is the crisis in human self-recognition[11], which is brought about by the emergence of informational entities that can perform information processing better than humans. However, by claiming that machines cannot become more 'intelligent' than humans, he is (probably strategically) building a line of defence.

Andy Clark is another optimistic philosopher. While mentioning the potential ramifications of ICT, he also claims that humans have been extending their capabilities to technological artefacts since the dawn of civilisation and this is precisely human nature (Clark [2]). For Clark, there is no reason to regard the current development of ICT as exceptional because humans will adapt to it and establish a new way of coexisting with it, just as they always have.

There is a conflict of views among the aforementioned intellectuals with respect to the value of communication. Turkle highlights that direct communication that involves a small number of participants in which one invests one's whole personality, which is interactively and cooperatively constructed over time, ultimately leads to strong empathy. On the other hand, Floridi and Clark emphasise massive, frictionless information transmission among numerous informational entities, including both human beings and artefacts. They do not seem to think that the latter type of communication denies the former. In particular, Clark believes that it is possible for both to coexist and that ICT development will eventually enable communication that combines the merits of both. Conversely, Turkle thinks that people do not dare engage themselves in the former kind of communication if they get accustomed to the latter. This is what she regards as the 'vulnerability' of human beings to ICT.

ICT definitely has its disadvantages and some of them have already come to light. An increasing number of people have already fallen victim to 'phubbing'.[12] There was a child who actually wished to 'become a smartphone'.[13] A survey at the University of Michigan indicated that college students today are significantly

[11] He calls this the 'fourth revolution' following Copernicus's heliocentric theory, Darwin's theory of evolution, and Freud's psychoanalytic theory of personality.

[12] A combination of the words 'phone' and 'snubbing'; refers to the act of using one's mobile phone and leaving one's company unattended as a result.

[13] An anecdote circulated in Singapore about a primary school teacher that assigned her pupils to write down their wishes and one pupil wrote that he wanted to 'be a smartphone because Papa and Mama are watching smartphones all the time'.

less empathetic to others compared to those 20 to 30 years ago.[14] A researcher in this survey suggested that social media's easy friendship might be a factor in this, although further investigation is needed to identify the actual reason. More recently, a long-term comprehensive survey of Facebook users showed that people who 'like', click on links in other users' posts, or frequently update their profiles become less happy and their self-evaluation of mental and physical health declines. On the contrary, actual socialising has been shown to have a positive effect on the said variables. (Shakya [12]).

There will undoubtedly be some movements against ICT's development and use due to its adverse effects; however, its development and diffusion will never stop. People will increasingly use smartphones and social robots will only become more popular in the future. This is because society needs them, large companies like Google and Softbank want them to be used, and governments promote them as a growth strategy. This is a winning combination, and Floridi and Clark are probably aware of it. Given what they know, they prepare arguments for the technology, rather than objecting to the change, in order to ease the pain accompanying it. However, as social robots become more popular, more drastic changes will occur in communication and human relationships; the understanding of and expectation from them; and the customs and norms concerning them. No one can predict exactly what these changes will be like.

Conclusion

When vulnerabilities are found in software, one can apply patches to address them. However, vulnerabilities or security holes in human psychology and physiology cannot always be resolved as easily. The human mind and body has adapted to the environment for millennia and, as a result, it has become what it is today. It is complex and often cannot be easily fixed. After all, one cannot hate sugar even if one intends to do so. However, it is important to be aware that the craving for sugar is not for its health benefits, but because it is an evolutionary mechanism. People must keep this in mind and harness their desire for sugar.

In order to use technology as a tool to achieve personal happiness or the prosperity of mankind, humans must be smarter than technology. Humans must understand exactly what a tool is intended for, the mechanisms in which it

[14] http://ns.umich.edu/new/releases/7724-empathycollegestudents-don-t-have-as-much-as-theyused-to. Accessed on 16th September 2016.

operates, and how it will affect humanity, among other things. Otherwise, humans will not use tools, but tools will use humans.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP16H03341. The author would like to thank the Japanese Society for Artificial Intelligence for allowing the translation publication of the original paper published in its journal.

Reference

[1] P. Bloom. *Against Empathy: The Case for Rational Compassion*. Penguin Random House, 2016.
[2] A. Clark. *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press, New York, 2003.
[3] L. Floridi. *The 4th Revolution: How the Infosphere Is Reshaping Human Reality*. Oxford University Press, New York, 2014.
[4] J. Greene. *Moral Tribes: Emotion, Reason and the Gap Between Us and Them*. Penguin Books, 2014.
[5] J. K. Hamlin, K. Wynn, and P. Bloom. Social evaluation by preverbal infants, *Nature*, 450, pp. 557-559, 2007.
[6] Y. N. Harari. *Sapiens: A Brief History of Humankind*. Harper, 2015.
[7] Y. Kanakogi, Y. Inoue, G. Matsuda, D. Butler, K. Hiraki, and M. Myowa-Yamakoshi. Preverbal infants affirm third-party interventions that protect victims from aggressors. *Nature Human Behaviour* 1(2):0037, 2017.
[8] D. Levy. *Love and Sex with Robots: The Evolution of Human-Robot Relationships*. Harper, 2007.
[9] W. H. McNeill, editor. *The Pursuit of Power: Technology, Armed Force, and Society Since A.D.1000*. University of Chicago Press, 1982.
[10] D. A. Nibert. *Animal Oppression and Human Violence: Domesecration, Capitalism, and Global Conflict*. Series: Critical Perspectives on Animals: Theory, Culture, Science, and Law. Columbia University Press, 2013.
[11] S. Pinker. *The Better Angels of Our Nature: Why Violence Has Declined*. Penguin Books, New York, 2011.

[12] H. B. Shakya and N. A. Christakis. Association of Facebook use with compromised well-being: A longitudinal study. *American J. of Epidemiology*, 185(3):203–211, 2017.
[13] S. Turkle. *Reclaiming Conversation: The Power of Talk in a Digital Age*. Penguin Press, New York, 2015.
[14] P. J. Zak. *Moral Molecule: The Source of Love and Prosperity*. Dutton, 2012.
[15] 久木田水生. 麦とペッパー — テクノロジーと人間の相互作用. 人工知能, 32(5):653–659, 2017.
[16] 石弘之・石紀美子. 鉄条網の歴史 — 自然・人間・戦争を変貌させた負の大発明. 洋泉社, 2013.

Journal of Philosophy of Life Vol.13, No.1 (January 2023):112-124

Clockwork Courage
A Defense of Virtuous Robots

Shimpei Okamoto*

Introduction

With the development of artificial intelligence (AI) and robotics, robotic devices are expected to be available in various social situations. These devices may be able to perform various tasks previously done by human beings. Some of these tasks, of course, have a significant impact on human well-being. If this is the case, it is necessary to implement a mechanism in robots' internal systems to regulate the behaviour that can be activated in such situations. Such a mechanism must include content worthy of the name 'ethical constraints' or 'moral rules'.

Engineers (and some philosophers) of AI and robotics have been investigating methods of implementing 'morality' in robots' computational systems since the late 1990s. These studies are called 'Artificial Morality' (Danielson 1992) or 'Machine Ethics' (Anderson and Anderson 2007). Their goal is to develop agents that can act morally without human decision-making, that is, artificial moral agents (AMAs). The project of building AMAs was initially just a thought-experiment, but now it must be considered as a practical issue of actual engineering, for the use of robotic devices has become a real problem.

In moral philosophy since ancient times, various theories have been proposed to understand human moral standards or moral thinking; for example, utilitarianism, Kantianism, contractualism, and so on. These theories are rich resources for understanding morality. Therefore, it is necessary for engineers and moral philosophers to cooperate in order to successfully build moral machines. However, this is not an easy task. If the project of building AMAs is conducted with the approach of selecting the most desirable of the various moral theories, and rewriting it into computable formulations, the standard that determines the best theory is necessary. Although, where does this standard exist? Even moral philosophers hardly agree on which theory is best. Rather, the conclusions arrived

* Assistant Professor, School of Letters, Hiroshima University. 1-2-3 Kagamiyama, Higashi-Hiroshima City, 739-8522 Email: sokmt[a]hiroshima-u.ac.jp

at by philosophers continue to diversify these theories. If engineers expect to be advised by philosophers' opinions to implement morality in robots, they will not easily be able to determine which theory to use.

In previous studies, two of the various moral theories appear promising. One is the virtue ethics advocated by Aristotle (Wallach and Allen 2009). The other is contractualism, advocated by Rawls (Leben 2018). These two theories seem to have greater implementability to computational systems than other theories. However, many philosophers are sceptical of these attempts from the philosophical point of view. They argue that 'can' does not imply 'ought', so even if we can implement morality in a machine, this does not mean that we should do so.

This paper examines the theoretical problems associated with the implementation of moral theory in machines, particularly in favour of virtue ethics. Nevertheless, the issues addressed here will be important not only for AMAs imbued with virtue ethics but also for those created with other theoretical backgrounds, such as consequentialism, in mind. This is because virtues are equally important to Aristotelian, Rawlsian, or any number of other theorists. Section 1 outlines research that attempts to implement morality into AI systems and explains why virtue ethics seems to be important. Section 2 introduces two objections to the attempt to implement virtues into robots. In Section 3, it is considered that virtues for robots are different from those for humans. Section 4 responds to this objection.

1. Artificial Moral Agents and Moral Theories

Even if robots can behave freely in social situations, 'free behaviour' does not entail complete autonomy. Their behaviour must be limited to some extent, because if people are harmed by the robot's behaviour, the use of the robot itself will be blamed, regardless of who is responsible for it. In order to avoid such a situation, robots need to have internal constraints on their actions; however, it is difficult to design computational ethical constraints, not to mention Isaac Asimov's 'Three Laws of Robotics', which illustrated that even these three simple rules can cause serious ethical challenges.

We must implement morality in robots—in other words, we must teach robots right from wrong. This, of course, does not mean that robots need to have the ability to think, feel, and judge just as humans do. Robots will be designed for

specific purposes, such as medical practices or military operations. Even if autonomous robots do have 'autonomy', it is not complete autonomy, but autonomy only in a very limited sense. However, within this 'limited sense', the robots must be able to make decisions and act spontaneously, because the greatest benefit of using robotic devices is that decisions and actions can be made by machines alone, without human supervision. If this is the case, the internal systems of robots must include substantial moral considerations for those who may be affected by their actions, even though their process is completely different from human ethical thinking.

In moral philosophy, various principles and rules that justify our moral judgments and guide our actions have been studied. In order to implement ethical constraints in computational systems, many studies refer to various moral theories that have been considered in moral philosophy. If these theories are formulated computably, they may serve as a blueprint for the design of the internal constraints of robots. However, there are two difficulties to be aware of. First, even though moral theories play a guiding role, they are not intended to regulate the process of thought when we are executing a moral behaviour. Most moral theories are used as a basis for critically examining one's past actions or the actions of others, or when envisioning future actions. In other words, the grounds provided by moral theories are used for the justification of actions, not for performing actions.

If a theory is to be worthy of being called 'moral', of course, it must not stray too far from our intuitive reactions. However, even a hard-core utilitarian, for example, does not calculate utilitarian consequences in their everyday actions. Rather, those who always refer to moral theory in their every decision may even be considered a morally flawed person. For example, Michael Stocker spoke of such people in terms of a kind of 'Schizophrenia of Modern Ethical Theories'. Agents who ignore their intuitions or emotional reactions, and follow only their committed principles, are closer to psychiatric patients than moral saints (Stocker 1976). Perhaps this objection also applies to robots. Indeed, we sometimes accuse humans who place more emphasis on principles and rules than personal commitments of 'acting like a robot'. For the same reasons, we may not want social robots to act 'like robots'.

Secondly, traditional moral theory is required as a reason to justify 'human actions', not 'robot actions'. Furthermore, the ethical hurdles for robots will be much higher than for humans (Allen, Varner, and Zinzer 2000). Colin Allen and his colleagues devised the 'Moral Turing Test' as a complete AMA standard.

However, unlike the original Turing test, it was found that the criterion for passing, which was if the robot and the human being could not be differentiated, is not adequate to determine the morality of the robot. This is because we often evaluate that an action is acceptable if it is done by humans, but if it is done by a robot, it will be considered unacceptable. If the standards of human morality and robot morality are different in the first place, undoubtedly, applying a theory made for humans to a robot without modification will at best provide an inappropriate standard.

Among the various moral theories, one of the theories that seems to avoid these difficulties is virtue ethics, a theory proposed by the ancient Greek philosopher Aristotle. This theory treats the ultimate standard as the agent's 'character traits' represented by their actions, rather than the results or the motives of their actions. In Aristotelian virtue ethics, for example, the reasons not to tell a lie is neither bad results (e.g., pain) nor bad motives (e.g., treating others as a means), but vicious character traits, in this case, the dishonesty represented by telling a lie. This theory avoids the difficulties described above. First, desirable character traits in virtue ethics imply respect for personal commitment and appropriate emotional responses. Second, the ideal moral agent in virtue ethics requires higher standards than ordinary people. In this theory, ordinary people can be moral agents, but they are actually imperfect moral agents, and are training to acquire virtues. If so, AMAs, which embody the ideals of virtue ethics, are more moral than human agents. In this case, even if the ethical hurdle of the robot is higher than that of a human, this will not present a problem.

Computability is also an advantage of virtue ethics. For example, Wallach and Allen (2009) distinguished the ways to implement moral theory in AMAs in both top-down and bottom-up approaches. On the one hand, a top-down approach is a way to rewrite an existing moral theory into a computable formulation. However, to calculate the consequences, for example, a large amount of simulation is required; therefore, it is not feasible in real-time moral thinking, where a conclusion must be drawn in an instant. Furthermore, this approach is difficult to deal with when AMAs face a case that the designers could not have predicted in advance. On the other hand, the bottom-up approach is the method of generating ethical rules by AI itself using tactics such as machine-learning or multi-agent simulation. However, by bypassing ethical hurdles through learning, AI may behave inappropriately. Moreover, if the data on which learning is based include prejudices, the tendency that AI learns may become an ethically bad habit.

Virtue ethics can be understood as a hybrid approach that includes both top-down elements that regulate actions through virtue, and bottom-up elements that learn virtuous actions through habits. In order to teach robots what kind of actions are virtuous, it is necessary to incorporate abstract virtues as ideals from the top-down approach. However, in order to have the ability to actually perform moral actions, they must learn virtuous one through daily training from the bottom-up approach. This ethical framework, as Aristotle considered, is appropriate not only for human moral education, but also for generating moral standards for robots.

Furthermore, virtue ethics has the advantage of having a high affinity with connectionist theory in the philosophy of mind. In neural-network theory, which is the basis of machine-learning, intelligence is understood as a combination of certain modules with various roles. According to some philosophers, the human thought-process assumed in virtue ethics is close to the cooperation of such modules, and human moral psychology is completely different from the top-down model that, for example, Kant had envisioned (Wallach and Allen 2009). If the intelligence assumed by virtue ethics is similar to AI as well as human intelligence, virtue ethics would be the best moral theory for AMAs in this sense.

2. Some Objections to Virtuous AMAs

As seen in the previous section, virtue ethics has many advantages as a model of morality implemented in AMAs. However, there are some objections to these attempts. The first objection occurs on the empirical level. For example, at present, it is difficult to even design an AI with sufficient ability to speak with humans. For this reason, it is even more difficult to build robots that embody moral ideals by making better moral judgments than humans. This is a distant dream, with no clear path to achieving it at present. Furthermore, even if robots could have virtue, there would be no way to determine definitively whether they really do have virtue.

Unfortunately, we will have to make a futuristic response to such objections. It may not be possible to build ‘perfect’ AMAs with current technology, but there may come a day when it is possible to do so in the future. However, it can be said that the virtue ethics model would be more appropriate for current artificial intelligence than certain other theories. For the question of whether virtue is possessed or not, it may be sufficient for human evaluators to accept that a given AI appears to have virtue. In any case, it will be desirable to respond to these empirical objections at the empirical level. For the purposes of this paper, it is

only to say that virtue ethics has a certain advantage, at least when considered as a model for ethical AI.

There are more important problems beyond the empirical level. In principle, and not only with reference to current robots, is it possible for robots in general to possess virtues? If so, should a virtuous robot be built? There are two types of questions here. One is the metaphysical question of whether a robot is an entity that can acquire virtue. The negative response to this question is that virtue is for humans and, by definition, no robot can have virtue. The other is a normative question of whether the attempt to implement virtue in AI is a desirable design from an ethical point of view. If the attempt to build virtuous robots is itself an ethically wrong project, we should not aim for it.

Consider the first question. In Aristotelian virtue ethics, virtues have a teleological structure. Every being has a purpose for existing. Achieving that goal well means excellent being. The purpose of a tool such as a hammer is, for example, to strike the nail well. If it has the ability, it is an excellent hammer, because the purpose of being a hammer is to hit a nail. Therefore, what is the purpose of being a human? In Aristotle's opinion, the purpose of human beings is happiness.

According to contemporary Aristotelian philosopher Rosalind Hursthouse, 'a virtue is a character trait a human being needs for *eudaimonia*, to flourish or live well' (1999: 167); that is, we have to acquire virtue because it is necessary for our happiness. Hursthouse understands happiness in the Aristotelian sense, as a flourishing of human ability. This is sometimes called *Eudaimonism* or *Perfectionism* with regard to well-being. In short, virtues in the Aristotelian framework are inseparable from perfection 'as human being'. We cannot live a happy life without having virtues; therefore, we need to acquire virtue to live a happy life. Individual moral acts are like by-products derived from this purpose.

Assume that virtues for humans are like those that Aristotle considered. Is a robot's purpose to live happily? Probably not. Most robots are designed, produced, and deployed as a means to achieve some purpose. If virtue means the perfection for achieving the purpose of being, then virtues for robots will be completely unrelated to morality. In other words, no matter how excellent a hammer is, it is not an ethically excellent hammer. Similarly, no matter how excellent robots are, it would be wrong to evaluate them as ethically excellent robots. On the other hand, because robots look as if they have morally excellent character traits and actually do not, they cannot be evaluated as virtuous. In the worst case, robots that

appears to have virtue may be called deceptive.

Let us proceed to the second objection. If we can call the virtues of robots ethical, this means that robots can aim for their own happiness—for virtue means the perfection of the purpose of being. If so, the attempt to design such a robot for a specific purpose would be a manifestation of vices. This is because, if robots have autonomy that is worthy of virtue, it would be wrong to ignore their autonomy and bind them to a specific purpose, just as it is wrong for parents to pre-determine how their child should live.

For example, Ryan Tonkens explains this by giving an example of a 'robotic clown'. A virtuous robotic clown may perform a variety of acrobatics with autonomy and, in some cases, even invent new arts. It will behave kindly to the audience and be generous to human colleagues. Perhaps the people around the robot may treat it as morally considerable, though, of course, not as morally considerable as humans. However, here is the problem: even if the robot clown has a high degree of spontaneity and has virtuous character traits, we cannot admit that it has the freedom to resign from being the clown. For, if this is allowed, there is no point in building such a robot. We will develop and use robots for specific purposes. Consider the military case. A virtuous robot soldier will fight 'with courage' in various military operations. The robot may have autonomy to shoot enemies and protect non-combatants, but we will not grant the robot soldier the freedom to retire from the army. This is because the robot was made for the purpose of being a soldier.

If this is the case, using AMAs is an activity very close to slavery. Although robots could have rights, they are not allowed to exercise them. It is hard to say that creating such slave agents is something that virtuous engineers and business owners should do. However, it is correct to argue that if AMAs can only become slaves, irrespective of how well they work, the autonomy or free-will that be the title of moral rights must not be implemented in their systems as they are tools and should be tools. Joanna Bryson (2010), for example, makes such a claim. If Bryson is correct, however, we should not make robots virtuous in the first place.

These two questions are parallel to the ones on moral patiency that David J. Gunkel proposed in *Robot Rights*. According to Gunkel, the following two questions about robot patiency (or moral rights) are often confused: (S1) 'Can robots have rights?' and (S2) 'Should robots have rights?' (2018: 5-6). Many argue that S1 and S2 are equivalent, that is, either if robots can have rights, then robots should have rights, or if robots cannot have rights, then robots should not

have rights. However, Gunkel says that S1 and S2 are different types of questions and can be answered separately. As with Gunkel's questions about rights, those about virtues need to be distinguished ontologically from normative questions. Further, normative conclusions are not always derived directly from ontological positions. In this paper, I propose that robots can acquire virtues even at the ontological level and should also acquire virtues at the normative level.

3. Virtues and Virtual Virtues

If the robot can have virtues, it should not be built. If the robots should be built, it should not have virtues. Both of these claims seem to be correct, but the current and following sections will attempt to refute both of these propositions. In other words, the following will argue that robots can have virtues, and should also have virtues. However, this does not mean that virtues for robots are the same as those for humans, but they are still worthy to be called virtues. As some researchers have suggested, virtues for AMAs are 'virtual virtues' (e.g., Wallach and Allen 2009; Coeckelbergh 2012; DeBeats 2014). Sceptics who argue against virtuous robots are wrong in thinking that virtues for AMAs are the same as those for humans.

Indeed, Wallach and Allen (2009) set the following criteria in order to find a moral theory suitable for AMAs.

> Given the range of perspectives regarding the morality of specific values, behaviors, and lifestyles, perhaps there is no single answer to the question of whose morality of what morality should be implemented in AI. Just as people have different moral standards, there is no reason why all computational systems must conform to the same code of behavior. (78-79)

They chose the best theory from the engineering point of view (e.g., 'Which theory is the easiest one to implement in the system?'), rather than from the philosophical point of view that cannot expect consensus (e.g., 'Which theory is the most ethically desirable?'). In their argument, the desirability of virtue ethics is ensured by having two sides, bottom-up and top-down, and its affinity with connectionism. For these and additional reasons, virtue ethics can be said to be ethically appropriate for AMAs.

Human virtues and virtual virtues are similar in some ways, and different in

several ways. First, we will consider the similarities. According to Aristotelian virtue ethics, various virtuous actions are more than individual character formation, such as courage and honesty, for example. They are more comprehensive and necessary traits for unconditional 'excellent agents'. If someone has virtue, she can perform virtuous acts in almost all situations. It was Aristotle who originally claimed this feature, but has been called 'the unity of virtues' by philosophers after him. Virtue for AMAs and humans alike is arguably the trait necessary for 'excellent robots'. This is because, from the viewpoint of engineering ease, it is unreasonable to design and implement individual virtues corresponding to each very complex social aspect. Just as human virtues are a kind of rationality and are necessary for 'excellent judgments' in general, virtual virtue is also a kind of rationality that regulates various actions in general. Therefore, virtual virtues also have their 'unity'.

The second similarity is their teleological structures. Virtues for humans have the purpose of the possessor's happiness. For Aristotle, happiness does not mean mere maximization of pleasure or satisfaction of preference, but perfection as a human being. Aristotle argued that happiness in this sense is an overly complicated and vague purpose, so that human beings cannot acquire happiness unless they have genuine virtue. Virtual virtues have a teleological structure as well. Just as virtues for humans are character traits useful for perfection as human beings, virtual virtues are character-traits useful for AMAs' perfection.

However, the similarity ends here. Virtues for humans are aimed at the happiness of their possessor, but virtual virtues are not. Rather, their purpose is not the AMAs themselves, but the well-being of those who are affected by AMAs' actions. Why do we need to build AMAs? Because robots are likely to be used in social situations where their behaviour significantly affects the well-being of humans. A situation wherein robots make such decisions without ethical constraints is undesirable; therefore, AMAs should be built. It is not because we want to increase new moral patients, nor is it because we want to increase new 'persons'. What we want is for the system to have ethical constraints. For this reason, AMAs need to establish 'safe interaction with humans', but it is not necessary for the robot to have a happy life. Recall Asimov's third law. Robots certainly must protect themselves or their happiness. However, this is only conditional.

4. Reply to Objections

Given the above, this section will respond to the two objections in the previous section. The first objection states that robots cannot have virtues, because robots cannot aim for their happiness. This is correct in one sense. However, for AMAs, aiming for a happy life as robots is not necessary for acquiring virtues. Sceptics seem to have a very narrow understanding of the concept of virtues. Certainly, it is misleading to refer to the excellence of non-human beings as a virtue in an ethical sense and many cases, even a mistake. For example, the hammer suitable for hitting nails has its excellence; however, this does not mean that it is an ethically excellent hammer, but that its ability to hit a nail is excellent. In the case of robots, however, being excellent as robots can mean being ethically desirable at the same time. Just as humans are not hammers, AMAs are not just hammers.

Further objections are anticipated in this regard. If virtual virtues are needed to make robots do something that makes them appear ethical , it is wrong to call them virtues. After all, even though robots cannot have human virtues, designers act as if robots actually encompass them. This is deceptive, because it is not true; as part of the robot's purpose includes interaction with humans, it simply appears as though human and virtual virtues are the same. In personal communication, robots must imitate humans, and this imitation is a goal for the robots. Even if it is only apparent, it is sufficient for robots. Even if a robotic soldier's courage is a clockwork courage, in other words, we may consider it virtual courage as long as it is a useful character trait for the robot's colleagues.

However, if virtual virtues are to be considered as described above, they should not be called virtues, because they are excellence relative to a specific purpose. For example, even if there is a soldier with excellent shooting skills, with the ability to snipe well, it does not mean that he is a virtuous soldier. However, remember that virtual virtues have the unity of virtues. Robots put into military operations will have the virtue of 'courage' to confront difficulties. Robots used for patient recreation will have the virtue of 'humour'. A multi-purpose communication robot may be worthy of having the virtue of 'kindness' or 'honesty'. Like human virtues, AMA's virtues cannot be bound to be certain lists. It is important to have a mechanism for making appropriate judgments in any situation. Virtual virtues, like human excellence, require AMAs to work well in every situation, even if the robot is only used for a specific purpose. In this respect, virtue for robots is still worthy of being called virtue.

Perhaps such an understanding of virtue may seem to face a second objection as this proposal means that robots are not treated appropriately, even though they have virtual virtues and autonomy. Building a robot for a specific purpose means that the robot will not be allowed to rewrite its role for any other purpose. If such an agent is a human being instead of a robot, they would be considered a slave. If this slavery metaphor is accurate, what an AMA designer attempts to do is nothing but a slave merchant's job. Slavery implies a wrong norm. The project implying wrong norms will be itself a morally wrong activity, irrespective of its economic advantages.

Tonkens (2012) argues that many AMA advocates are unaware of the inconsistency between the norms they follow and the norms they are trying to give robots. Certainly, as he says, AMA advocates do not believe that the norms they follow and the norms they attempt to implement to robots must necessarily be the same norms. However, this objection is valid only when the virtuous robots are also moral patients. Since it is assumed that virtuous robots are subject to moral consideration just as humans are, it seems as if it is morally wrong to constrain such agents for a specific purpose. If robots have virtue in the same sense as humans, this objection is valid, because the purpose of virtue is the happiness of its possessor, and agents who can be happy are also moral patients. Given that the character traits to be implemented in the robot are virtual virtues, however, this objection can be avoided. Even if the purpose of the robot is a kind of perfection as robots, its aim is not to contribute to one's own happiness, but to contribute to the happiness of the people around it.

Individual engineers do not need to have good intentions when trying to design AMAs. Furthermore, there is even no need for planners and operators to have virtue. The second objection is aimed at sceptics of AMA designing. According to T. M. Scanlon, the moral permissibility of an action is independent from the agent's intent (2008: Chapter 1). What is important for an action to be morally permissible is that it does not violate the principles on which it is premised, or that the agent does not perform it for the wrong kind of reasons. Whatever its intent is, an action from the right kind of normative reasons and moral considerations is morally acceptable. Thus, it can be said that the development of AMAs would be one such morally acceptable project.

Conclusions

This paper has discussed the importance of virtue in the development of AMA, but this does not mean that it necessitates Aristotelian virtue ethics as a normative theory. Irrespective of which moral theory is adopted, it is possible to adopt the theory of virtue as a guideline for actual actions. For example, it is reasonable for consequentialists and Kantians to recognise the practical usefulness of the concept of virtue, and to argue that agents have to possess some virtues to achieve the greatest amount of the greatest happiness or to obey the categorical imperative. Coeckelbergh (2012) says that we must admit that virtue is important for AMAs, but his rationale seems a consequentialist one, leading to human well-being.

This is very suggestive. As mentioned earlier, the main purpose of normative moral theories is the justification of action, not the regulation of the actual thought process. If this is the case, while adopting the theory of virtue as a thought process, it is not necessary to appeal to virtue as a normative theory. For example, Julia Driver or Roger Crisp's theory of virtue can be understood as such positions, because they appeal to the importance of virtue from a consequentialist point of view (e.g., Driver 2005; Crisp 1992). The approach that they take to virtue is called 'virtue consequentialism'. In their view, moral virtue is 'a character trait that would systematically produce actual good under normal circumstances' (Driver 2005: 78). This view of virtue is not Aristotelian, but it is still a theory of virtue. In other words, again, robots can be courageous, even if it is clockwork courage.

Even adopting their consequentialist point of view, we can say that it is not the ultimate moral theory to which we ought to be committed, although we should implement virtues in the robots' internal mechanisms. Regardless of which moral theory is adopted, virtues can be recognised as a kind of secondary rule of the principle. This is true both for ourselves and for robots. However, the Aristotelian framework still has two advantages: as a sophisticated theory of the unity of virtues, and as a teleological structure of happiness-seeking (though, in the case of robots, it is not the possessor's own happiness that is being sought).

Reference

Anderson, M. and Anderson, Susan L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent, *AI Magazine*, 28(4), 15-26.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence* 12, 251–261.

Aristotle (translated by W. D. Ross). *The Nicomachean Ethics: Translated with an Introduction*. Oxford University Press.

Bryson, Joanna J. (2010). "Robots Should be Slaves," in Yorick Wilks (ed.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, John Benjamins, pp. 63-74.

Coeckelbergh, M. (2012). "Care Robots, Virtual Virtue, and the Best Possible Life," in P. Brey, A. Briggle, and E. Spence (eds.), *The Good Life in a Technological Age.* Routledge. pp. 281-293.

Crisp, Roger (1992). Utilitarianism and the Life of Virtue, *The Philosophical Quarterly* 42(167), 139-160.

Danielson, Peter (1992). *Artificial Morality: Virtuous Robots for Virtual Games*, Routledge.

DeBaets, Amy Michelle (2014). Can a Robot Pursue the Good?: Exploring Artificial Moral Agency, *Journal of Evolution and Technology* 24(3), 76-86.

Driver, Julia (2005). *Uneasy Virtue*, Cambridge University Press.

Gunkel, David J. (2018). *Robot Rights*, The MIT Press.

Hursthouse, Rosalind (1999). *On Virtue Ethics*, Oxford University Press.

Leben, Derek (2018). *Ethics for Robots: How to Design a Moral Algorithm*, Routledge.

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*, Harvard University Press.

Stocker, Michael (1976). The Schizophrenia of Modern Ethical Theories, *The Journal of Philosophy* 73(14), 453-466.

Tonkens, Ryan (2012). Out of Character: On the Creation of Virtuous Machines, *Ethics and Information Technology* 14 (2), 137-149.

Wallach, Wendell and Allen, Colin (2009). *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press.

Journal of Philosophy of Life Vol.13, No.1 (January 2023):125-134

Reconstructing Agency from Choice

Yuko Murakami*

1. Introduction

Research in artificial intelligence (AI) seems to cast doubts on the presuppositions of mind-body connections. In fact, AI by itself does not reveal the inadequacy of the mind-body problem setting. AI mimics an accelerated cognitive process that amplifies the dissonance of our models of humanity. Most of us tend to feel uncomfortable when faced with information technology, even without knowing what philosophers have discussed.

Traditional cognitive science is one field of dissonance and it assumes that the brain plays an essential role in performing mental activities. It is also a common belief that the mind supervenes the body. Researchers on AI focus on brain science because they believe that the mind is reduced to, or at least corresponds to, the activity of the body, particularly the brain. Observational results of vital signals, such as cerebral blood circulation, are considered physical counterparts of mental activities. For example, Kamitani [1] decodes fMRI results to display what experimental subjects see or even dream. This shows that there seems to be a direct correspondence between the mind and the brain. Such traditional approaches toward mind and person involve tacit assumptions that (1) each individual human body is connected; (2) the human mind is associated with a human body in the sense that mental activities correspond to bodily states and cannot exist without these bodily states; and (3) physical bodies outside of one's own human body are not construed as necessary factors of her mental activities. AI undermines the intuitive justification of these assumptions.

In contrast, some AI research can assume another set of assumptions. For example, the transhumanist notion of upload human presupposes the possibility of total separation of the mind and its original body, particularly when they insist on "copying information to media other than the body." The concept of the

* Professor, Graduate School of Artificial Intelligence and Science, Rikkyo University. Ikebukuro Campus 3-34-1 Nishi-Ikebukuro, Toshima-ku, Tokyo Japan 1718501.
Email: yukoim@rikkyo.ac.jp

individual person in the transhumanist picture is like an iPhone account. Your account information is simply saved on a cloud system; when the hardware of your iPhone is damaged or outdated, you can upgrade it to a new one by downloading all of the data from the cloud. Thus, you can replace your iPhone without stress.

Both brain scientists and transhumanists share in the assumption that each personal mind is associated with a single physical body, while the possibility of replacing body parts remains controversial.

The disconnection between mental activities and human bodies has been noted in history, even before AI began to flourish. It is seen in the series of externalization of mental contents and extension out of the human body, as will be discussed in the following section. Moreover, a wide range of scientific research suggests that it not only humans have the capacity to think, collect information, use tools, and use language to communicate, judge, and reason. Intellectual superiority to other creatures is no longer regarded as the basis of "humanity."

What then should serve as the conceptual basis of humanity? We claim that it is the concept of agency, since human beings are social beings. We must act interpersonally to be counted as humans. In this paper, we methodologically eliminate the distinction between human and non-human to reconstruct the notion of agency without relying on reduction to individual agents associated with each individual human body. According to the "extended notion of agency," an agency is a dynamic system of continuously choosing the world where the agent resides by describing a set of possible choices and selecting priorities among those choices. Such a system can have both human and non-human components. A historical series of choices by an agent formulates or defines the personality of the agent. If such conceptual model is right, (1) the concept of agency will be reduced to a temporal model of choice; (2) an agent is a system of dependence among component systems; (3) the dependence relation of an agent can be non-well-founded; (4) the identity of two agents will be defined as an existence of a bisimulation relation between dependent components; and (5) the concept of rationality of agents will likewise be non-well-founded.

2. From Extended Mind to Extended Agency

2-1 Precursor: Externalism of meaning

Since the mid-twentieth century, there have already been trends toward externalization in the history of philosophy [1]. Before Putnam [2], the meaning of language, which is the basis of language use regarded as activity in the mind, was considered mental content. However, he argues that meanings are not found in the brain; rather, it is external conditions that determine meanings (externalism of meaning). He established this through the twin earth thought experiment. Being a thought experiment, its physical feasibility is irrelevant.

Now, let us think about “Twin Earth,” which has almost the same structure as the earth we live in, with the same people who have the same things. The same person is the same in all properties, especially in the substances that make up the brain. The only difference is that the liquid (that is, water) indicated by “H_2O” in the language used by this side is replaced by “XYZ”, which has another chemical composition but shows the very same property as water. Residents of the Twin Earth call “XYZ” water. It is assumed that Alice, who is a resident of this earth, and Alice', a resident of twin earth corresponding to Alice in this earth, are talking about a statement that makes some claim about water. The question is: are Alice and Alice' talking about the same thing? In other words, do their sentences have the same meaning? We answer in the negative. Given that Alice talks about H_2O and Alice' talks about XYZ, these two sentences must have different meanings (It is as though talking about two different people with the same name). Since we have assumed that the brain structures of Alice and Alice' are the same, this example shows that relying on brain structure alone is not enough to determine whether or not the meaning is the same. Simply put, brain structure is not enough to determine meaning.

2-2 Extended Mind Thesis (EMT)

The next externalization was memory, which has also been considered to be “mind-related.”

Memory, as well as language use, has been regarded as human heart activity. Clark [3] claimed that the Extended Mind Thesis (EMT), inspired by the example of a note written in a notebook, could extend one’s memory. He thus formulated

the Parity Principle [4].

> [Parity Principle] When we work on something and accept it without hesitation as a part of the cognitive process and as a part of the world, it enters our mind that part of the world is part of our cognitive process.

As regards this issue, it was previously claimed that the way of searching for information differs between inside and outside the mind [5]; however, several more digital devices are in use now than when the extended thesis was claimed. Despite the habit of using memos for memorandums, we are now experiencing lifestyle changes that lessen the need for setting concrete meeting places and instead encourage the use of mobile communications.

2-3 Extended Agency Thesis

Technologies have ubiquitously penetrated our life, such that the earth is no longer resilient; human beings can now go and act further than ever. In particular, on the basis of AI-related technology, the notion of agency should be extended from agency of an individual to agency in general.

Consider the case of when you start learning to play a musical instrument. You are gradually taught how to hold the instrument and how to move your body; it takes a lot of practice before you can start playing the correct sounds. It entails moving your body in an extremely unnatural manner and using muscles that you never intentionally moved until you started using the instrument. It is only when these unnatural movements can be performed easily, without hesitation or effort, that you can finally claim to be a good instrumentalist. Well-trained players reshape their bodies according to the instruments and play music as they think, as though their reshaped bodies and the instruments are united.

Similar to music instruments, we extend our own body to involve our social systems. In addition to driving a car and riding a public transport system, using information devices such as smartphones and tablet computers require training before you are able to use them. Once you have access to these systems and devices, using them to achieve what you set out to do becomes a trivial matter. For example, a public transport system is not owned by you (unless you are the owner of the transportation company). It is merely available for use for a fare. Its operation may not be exactly commensurate with your wishes, but you undertake

trips with it nonetheless.

All parts of such a system are often interdependent. Adjusting one part affects other parts, and changes in other parts also bounce back to the original part. By putting dependencies, access relations, and reference relations together and describing them as a subsystem, there is no guarantee that atomic elements exist in the "part" of the agents. They form a network. Furthermore, it is necessary to consider actions separately from human physical activity even for agents considered to have intentions. The mainstream version of the notion of behavior is that an event in a causal sequence with intention is identified as an activity. The presence or absence of the agent's intention explains why the agent caused the event. It is considered an ability. Although there are variations in terms of whether cause and reason are considered to be the same, recent tendency to seek explanatory ability in autonomous AI assumes this type of action theory.

However, there are actions that cannot be handled by the type of action theory discussed in the previous section. In other words, it is not always appropriate to place an action in a causal sequence. Note that human actions do not necessarily involve the physical acts of the agents of the acts [6]. For example, an act of omission, or "I have a duty to do that and I can do it. But I do not," may not be associated with any physical movement. It is an omission of the organization that the responsible agency knows the possibility of drug damage and does not cancel the drug approval. Moreover, there may be omissions with regard to actions involving language. Being silent in a conversation is an act of omission with no physical movement. In other words, the premise that actions always have physical basis is not always true.

Furthermore, if physical entities can be considered separately from the cognitive process, the claim of independence between the act and the physical basis does not contradict the parity principle.

Of course, we can also think that the results of other people's actions have causally contributed to the realization of our own actions. However, what I would like to emphasize here is that rather than taking into account the causality of the consequences of these people's actions, it is important to note that these actions are carried out and that the credibility of the equipment is almost self-evident. It is said that this has a structure similar to that of Clarke's extended thesis example. Thus, the idea of the agent itself should be expanded.

Is it possible to take agents as primitive concepts while denying physical reductionism of agency? I would like to think so. Nakayama [7] [8] set forward

Clark's EMT as a criticism of the idea based on the assumption that the mind is in the brain and also in contexts including body extensions such as cyborgs. On the basis of mereology, he formulates the extended agent as follows.

(a) [Atomic Agent] An atomic agent is an agent. Any spatial part of an atomic agent is not an agent. Here, we simply presuppose that there are atomic agents. An atomic agent constitutes the core (or one of the cores) of any extended agent.
(b) [Agents and Tools] Let temporal-part (x, t) denote the temporal part of object x in time t. Let A be an agent that uses (tool) B in time t to perform an action. Then, the (four-dimensional) mereological sum, temporal-part (A, t) + temporal-part (B, t), is an agent. We can easily prove within the four-dimensional mereology that temporal-part (A+B, t) = temporal-part (A, t) + temporal-part (B, t).
(c) [Collective Agent] If agents A1An perform a joint action, A1An is an agent (for the notion of joint action, see [21]).
(d) If an object satisfies neither (2a) nor (2b) nor (2c), it is not an agent.
(e) [Extended Agent] An agent that is not atomic is called an extended agent.

Nakayama, by this formulation, aims to characterize the mind as situations and processes associated with agents. The notion of agency—not of the mind—is primitive.

Contrary to Nakayama's claim, agents may not be reduced to atomic agents because agency is a non-well-founded concept that may include cycles in the mereological structure of agents.

We seem to believe that our own body, without doubting that it is given to us, is our property, and controllable to us. However, we assume physical systems including musical instruments and social systems as an extension of our body—the apparatus of our mind in the causal sphere. Consciousness on one's own body is vague. The body can be moved as usual unconsciously. However, when I think about my illness and pregnancy, I notice that my physical condition changed even before I became aware of it; my control over this change is very limited. In addition, awareness may be directed to a part of one's body only through an abnormal situation such as pain. In other words, it is only a belief that you can control your own body.

The boundaries within the human body are not clear. The external boundary

is physically visible with the naked eye and is considered to be the boundary as a human individual. It is a cylindrical mass with irregularities and a large number of microorganisms inside of it. The body is attached to the inside of the mass as if there are various things floating inside the liquid. Since this microbe is outside the tube, it is outside of its own body; but because the tube itself is regarded as an individual, it is thought to be inside the (digestive) individual. However, in terms of the behavior of these microorganisms, they are not directly controllable. It is hard to try to change the internal environment by pouring some liquid or solid inside the cylinder. In addition, intentional control of the behavior of the cells that make up the body is almost impossible with willpower alone; thus, people try to control it through nutrition intake, the use of drugs, or by performing genetic manipulation.

> [Parity Principle of Agency] When we act and accept without hesitation as part of the cognitive process, as part of the world functions as if it gets in our mind (at that time) part of the world is part of the process of our agency.

3. Formalization: Choice and agency as primitive concepts

Now, we will delve into the main proposal of the paper. With choice being a primitive concept, we can define the notion of agent/ person on the indeterministic temporal frame proposed by Prior and Thomason [12] [13]. Here, the notions of agency and choice are extended versions of the STIT theory [7] [8]. In STIT, the formulation focuses on an evaluation of an action at the moment of the choice toward the action or at the moment strictly after the choice. Our idea is based on the first formulation of evaluation of action at the moment of choice, while modifying the definition of each "agent" on semantics. The following is a sketch of our formulation.

A tree structure (*branching-time frame*, BT) is a pair F = 〈T,<, where T is a nonempty set, whose members are called *moments*; < is a tree relation on T (a partial order on T being (1) linear to the past and (2) connected). A *history in F* is a maximal linear subset of T. A history h is said to *go through a moment* m when m ∈ h. A *moment-history pair* is a pair of a moment m and a history that goes through m. H_m denotes the set of histories that go through m; HF denotes the set of all the histories in F; and Moment-History denotes the set of all moment-history pairs in F.

We now extend branching frames of time with "choice." Consider a branching-time frame F = 〈T,<〉 . A *choice at a moment* m ∈ T is a partition of the set of moment-history pairs 〈m,h〉 . A *choice set at* m is a set of choices at m. Note that it may be the empty set or the powerset of the set of all choices at m; an arbitrary set of choices at m may play a choice set at m. C is called *the choice set on F* if C is the set of choice sets at m where m is in T. A *choice frame* is F_C = 〈T,<,C〉 , where F = 〈T,<〉 is a branching frame and C is the set of choice sets on F.

An *agent* in a choice frame F_C is a series of choices along a history. A *fragment of an agent* a in F_C is a series of choices along a consecutive fragment of a history in F_C.

The choice frame seems similar to the model that Parfit describes when he argues partial survival [Parfit 10: 298-302] and successive selves [Parfit 10: 302-305]. Our proposal differs from Parfit's in that each successive self is formed by itself with the series of choices on the branch. Choice is the primitive notion that defines the notion of person. With the choice frame version of the notion of agents, Parfit's division of me is just creating two agents with an identical fragment of agents in the initial part of history.

The choice-based notion of agents may seem strange as it is meaningless to say, for example, "I regret not doing X" as "I" refers to the agent on the very history where "I" resides. The agent who might have done X is not the same agent of "I." Such failure of transworld personal identity makes counterfactual statements about any person meaningless. For example, the counterfactual sentence "Hanna should have left the heavy shoes home" seems to require transworld identity among Hanna in the world where the sentence is stated ($Hanna_1$) and Hanna in the world where she wisely leaves the heavy shoes home ($Hanna_2$).

The solution to the problem is rather simple. There are two ways of identifying agents/person: one is in the causal structure of choice and the other is in the world of reason. Physical continuity is in the causal world and may not be reflected in the world of reason. The justification of actions is made in the world of reason, which requires transworld identity of the person to consider counterfactual cases, while also being based on an integrated series of choices among those agents in the causal world, which are associated to be identical on the world of reason.

It is the generalization of the notion of person. Physical continuity in Parfit requires the determination of a person to consider only the agents that share

fragments of agents. Identification of "I" can be formulated as determination of the fragments of agents. Such a determination is to define one's own life, which is an essential component of human dignity. In fact, it is the privilege of human beings to define the world he selectively lives in and the meaning of the language he speaks. He is both a component of the world and an agent who gives meanings and conventions in the world. The determination must be socially shared to make the notion of person significant as any person cannot be isolated from the rest of the community. Transhumanists deny such physical continuity but claim that psychological continuity is enough. However, it lacks social determination, and consequently such an identification of self is socially meaningless.

References

[1] Lau, J & Deutsch, M. "Externalism About Mental Content", *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/content-externalism/>.

[2] Kamitani, Y., & Tong, F. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5), 679–685. 2005. doi:10.1038/nn1444

[3] Putnam, H. "Meaning and Reference." *Journal of Philosophy* 70(19), pp. 699–711, 1973. doi: 10.2307/2025079

[4] Clark A. & Chalmers, D. "The Extended Mind." *Analysis* 58(1), pp. 7–19, 1998.

[5] Clark, A. *Supersizing the Mind*. Oxford University Press, Oxford, 2008.

[6] Simon, H. *The Science of Artificial*. MIT Press, 1969. 3rd Ed. 1996.

[7] Belnap, N. "Backward and Forward in the Modal Logic of Agency." *Philosophy and Phenomenological Research*, 51(4), pp. 777–807, 1991.

[8] Belnap, N. "Before Refraining: Concepts for Agency." *Erkenntnis*, 34(2), pp. 137–169, 1991.

[9] Nakayama, Y. "The extended mind and the extended agent." *Social and Behavioral Sciences* 97, pp. 503–510, 2013.

[10] Parfit, D. *Reasons and Persons.* Oxford University Press, 1984.

[11] Parfit, D. Divided Minds and the Nature of Persons. In Colin Blakemore & Susan A. Greenfield (eds.), *Mindwaves*. Blackwell, pp.19–28, 1987.

[12] Prior, A. *Past, Present and Future*. Oxford University Press, Oxford, 1967.

[13] Thomason, R.H. Indeterminist time and truth-value gaps. *Theoria*, 36, pp. 264–281, 1970.

Journal of Philosophy of Life Vol.13, No.1 (January 2023):135-146

Gushing Prose
Will Machines Ever be Able to Translate as Badly as Humans?

Rossa Ó Muireartaigh*

Flow without Flaw

Type in the word “bowl” into a software translation program, such as Google Translate, and watch as the equivalent Japanese “ボウル” appears almost the moment you press the keys. It is a translation provided instantly, flawlessly, without any of the tortured hesitations and indecision a human translator would experience. Hand the same text “bowl” to the human and watch the paroxysm of self-doubt as they fluster and fret, maybe even grab their heads, and wrack their brains over whether to say “ボウル”, “茶碗” or something else. When finished, the human translator will never really know, heart of hearts, whether their translation was really the best one possible, or whether they mistakenly chose the less best option. The machine has done what was asked of it with the resources it was given. It has made no mistake. When the machine gives us an answer we do not want, (such as when I type in “bowl” to mean throwing a ball at a wicket but get uncricket “ボウル”) the machine has not made a mistake. It did what I asked. My question was simply wrong. If I build a watermill but flow in sludge rather than water and block the wheel, it was not the watermill’s fault, but my own for feeding it the wrong inputs. Not so human translation which is like a watermill that will stop and pause and groan and reverse and spin, once fast and then slow, creaking and squeaking with panic, even when fed a perfect flow of water.

The Wheels of Cognition

What is the difference between a machine and a human translating? Imagine we could watch the process as an outside observer, the same way we can watch water pushing a watermill to grind corn, with each step in the process following

* Associate Professor, Department of British and American Studies, School of Foreign Studies, Aichi Prefectural University, 1522-3 Ibaragabasama, Nagakute-shi, Aichi, 480-1198 Japan. Email: omuireartaigh[a]gmail.com

the next. We see the inputted text (the source text) flowing in. Its arrival pushes out a target text. In between, what has happened is that the source text has been converted into some kind of code which is matched with equivalent codes from which the target text will be assembled. There is a neat inevitable flow, each effect has had its singular cause, as the whole translating machine cranks into a symphonic whirl of texts and codes, one pushing the other as the target text is cranked out with all the mechanical reliability of cogs turning cogs. The whole process is automatic. One step producing the next step with the same blind will by which in a watermill water pushes one flap making it rise slightly, to present another flap for water to push, launching the mill into full rotation to turn the connected shaft, that will drive the runner stone to grind the wheat. Nothing in the water or the mill or the runner stone actually knows what it is doing. It is simply parts moving other parts according to the laws of physics. So too with machine translation. Nothing in the machine or the source text or the target text knows what is going on. It is all the action of no-mind, one could say, like a zombified samurai swishing his sword when doing his bushido thing. This is what John Searle's famous thought experiment, entitled the "Chinese Room", so deftly illustrates.[1] Let us imagine someone sitting in a room whose job it is to look at cards with certain codes written on them and to output corresponding cards according to an instruction manual detailing which output card goes with which input card. The cards show Chinese characters which the person reading the manual does not understand. Even so, if the instruction manual is correct and the job properly done, anyone inputting cards into the room would be under the impression that the Room is communicating with them and has, as Alan Turing would argue, a conscious mind.[2] And yet, as Searle's description shows, this is not the case. By pure unconscious mechanical operations language has been used. My point here, of course, being that machine translation is working the same way, outputting in response to what is inputted without ever knowing what it is doing. It is as flawless and devoid of mistakes in its proper operation as a watermill grinding wheat, but equally unknowing of what it is doing.

Blackbox Beetles

With humans it could not be more different. For no matter how much we try

[1] John Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences* 3 (3) (1980): 417-457.
[2] Alan Turing, "Computing Machinery and Intelligence," *Mind* 59 (236) (1950): 433–60.

to observe each step of the translating operation when done by a human, there will be a certain blackbox moment where we cannot peer in any further. This is the moment when the translator is consciously choosing which words to use, or to borrow Searle's image again, which cards to output. For in the case of the human, there is no instruction manual, or rather the instruction manual lies locked and hidden within the blackbox of the conscious mind. Maybe it is not even there at all. This blackbox I speak of can be thought of as the beetlebox in yet another famous thought experiment, this time by Wittgenstein.[3] The idea of the beetlebox is to imagine that I hold in my hands a box which may or may not contain a beetle. You can never open the box to find out. And yet if I keep using the word "beetle" correctly in our shared language, the box might as well have a beetle in it. In other words, when we use language with another person we can only judge what is going on in their minds by how they use language. This secondary layer of evidence, the visible, surface presence of language, reflects, we can only assume, an inner consciousness. It does not grant us full direct sight of it. Let us take, for example, the word '茶碗' [chawan]. When I use this word you have absolutely no idea if my understanding of it is the same as yours. So long as I continue to use the word correctly, to play the language game correctly (or let's call it "translation game"), such as when I utter the sentence "this *chawan* evokes the asesthetics of *sabi-wabi*", you cannot accuse me of not understanding what "*chawan*" means. (On the contrary, if I say, for example, "my *chawan* is up by 5 degrees since I went jogging this morning", you are entitled to wonder if our meanings are the same.) Which means then that if I translate "茶碗" as "cup" you cannot accuse me of not understanding the word, even if you think "*chawan*" or "tea bowl" is the better translation. When we dispute the best possible translations for a source text we are still not looking into each other's blackbox of inner consciousness, we are merely negotiating and contending moves in the translation game. There is no telepathy, only translation games. As an aside, may I mention the fact that I have in my time met orientalists who do, implicitly, claim to have telepathic powers and to be able to look into the inner conscious minds of non-Japanese and non-Asians and convince themselves that the understandings such people have of words such as "茶碗" [chawan], "さび" [sabi] and "わび" [wabi] are not the same as theirs. They have a remarkable ability to see the ignorance in the minds of others, but, alas, not in their own.

[3] Ludwig Wittgenstein, *Philosophical Investigations* (Oxford: Blackwell, 1958).

Seeing Consciousness

The essential difference between the machine translator and the human one is the existence of this consciousness at the core of the human mind—which cannot be seen or accessed. In other words, from the third person point of view, the human translator and the machine translator look exactly the same. The only giveaway, maybe, is the fact that the human translator isn't as good. I can only assume the human translator has a consciousness on the grounds that I have one myself, and because the human is behaving as though they had a consciousness, something I judge on the basis of whether or not they replicate my own behavior. When I translate, I do not translate with the automaticity of a machine, as much as I may try, but instead experience constant indecisions, aporias, and dissatisfaction as I, with conscious deliberation, abduce interpretations of the source text. Other human translators seem, on the surface, to be doing the same thing. On these grounds alone, I assume that they are conscious (and the translating machine is not).

To express this further let me reach for another famous thought experiment, David Chalmers's zombies.[4] Here we are asked to consider the difference between conscious humans and zombies (who can perfectly replicate the behavior of said conscious humans). It would seem that for third party observers telling this difference would be impossible. A zombie that behaved perfectly like a human would be, for us, a human. And we would falsely assume conscious experience where there was none. Similarly, a machine translation could, theoretically, replicate perfectly a human translator, perhaps by being programmed to behave with slow self-torturing hesitancy, shifting between flashes of creative brilliance and dull ploding, all the while demonstrating stressed-out behavior, worrying about impossible deadlines imposed by narky underpaying translation agencies spouting sanctimonious gunk about 'quality control'. In such a case we would then falsely ascribe conscious experience to the machine, apropos again Turing's *gedanken*. Both zombies and translating machines lack something that cannot be seen. And that is, conscious experience. And this can pose a problem for the consciousness-seeking outside observer who ends up having to try to prove a negative, looking for what is not meant to be seen.

[4] David Chalmers, *The Conscious Mind* (New York; Oxford: Oxford University Press, 1996).

All or Nothing

Slavoj Žižek in *Organs without Bodies* refers to Chalmers and makes the argument that trying to attain a third-person perspective on consciousness's place in the universe is both an epistemological and ontological impossibility (it isn't to be known and it is not there). Žižek summarizes his view in a paragraph that is worth reading in full:

> When Chalmers writes in his argument against the reductive explanation of consciousness that "even if we knew very last detail about the physics of the universe—the configuration, causation, and evolution among all the fields and particles in the spatio-temporal manifold—*that* information would not lead us to postulate the existence of conscious experience," he commits the standard Kantian mistake: such a total knowledge is strictly nonsensical, epistemologically *and* ontologically. His reasoning is the obverse of the vulgar determinist notion articulated by in Marxism by Bukharin, who wrote that, if we were to know the entirety of physical reality, we would also be able to predict precisely the emergence of a revolution. More generally, this line of reasoning—consciousness as an excess/surplus over physical totality—is misleading since it has to evoke a meaningless hyperbole. When we imagine the Whole of reality, there is no longer any place for consciousness (and subjectivity). There are, as we have already seen, only two options left open here: either subjectivity is an illusion or reality is *in itself* (not only epistemologically) not-All.[5]

Žižek, of course, is the great *pratyekabuddha* of our times. And so, his views dovetail nicely with that of much of the Kyoto School.[6] Reality can never be seen

[5] Slavoj Žižek, *Organs without Bodies* (New York; London: Routledge, 2004), 115.

[6] Nishida Kitaro clearly places links world to consciousness in the now: "The world thus begins in the dimension that the present determines itself as the self-determination of the absolute present. The self begins in the dimension that knowledge and will unite, in the sense that that which reflects and that which is reflected are one. The self arises in the place that the world arises; the world arises in the place that the self arises. Therefore, our self-consciousness is the self-expression of the world, and the self-expression of the world is our self-consciousness." ("Toward a Philosophy of Religion with the Concept of Preestablished Harmony as a Guide," *The Eastern Buddhist* 3(1) (1970), 36-37) This placing of consciousness in the now from the which the world is emergent leads to the assumption of an emptiness where the All that can only be the not-All can be. As Nishitani Keiji describes it: "In short, it is only on a field where the being of all things is a being at one with emptiness that it is possible for all things to

in a totality that could be observed by an omniscient third party and the proof of this impossibility is the unplaceability of consciousness when one does adopt such a view. This is an opinion that Nishida Kitaro and Nishitani Keiji, along with Žižek, judging from his take on Chalmers, would hold to. Why is full knowledge of everything not possible? To have full knowledge of everything would be the crudest of pantheism. It would mean that the thoughts in your head would be the same as the thoughts of God. There could be no separation between you and God. And yet you experience yourself as being separate to God (or at least separate to the rest of the world that is not you). This experience, which by definition means that we have a *partial* knowledge of things, would have to be included in the *full* knowledge of things. Not knowing everything would be part of the absolute condition of knowing all. It is an impossibility. Your consciousness, with its partial knowledge, is a constant knowledge gap within the All of existence, of the world, of the universe.

Instead we need to lose the idea that knowledge is the mapping of all that is there onto all that we can consciously make out to be all that is there. Instead of seeing the world out there as a stable body of knowledge on which we flash the torches of our inner consciousness, attaining knowledge of one piece of a giant picture that someone with a bigger flash lamp (like God or a cosmic machine) could see overall, we should see the world out there as not prior known. The world out there, in fact, shapes itself into a stable body of knowledge to suit the flash lamp we shine upon it. Otherwise, without our conscious attention, it is not, in effect, "there". It is empty.

In this picture, knowledge must be relocated at the point of consciousness on the field of emptiness. Only a consciousness, such as that of a human, can be there at such a point. A translation machine, crunching away the bits and signals and codes of a fed-in text is not there. It can only be put there when it is the object of our attention. In other words, when a machine translates, it is because we have let ourselves believe that it has translated. The machine translates and the watermill turns because our minds translate and our minds turn.

The Mystery of Translation

To consider the special place (or non-place) of the conscious mind, let us

gather into one, even while each retains its reality as an absolutely unique being." (*Religion and Nothingness*, 1982, 148).

consider further how the act of translation actually works. When we translate we produce a new text that does not look in anyway like the old text. And yet there is a sameness about them, otherwise they would not be translations, but merely two different texts in two different languages. To illustrate, let us take the following Japanese source text:

[source text] 青磁輪茶碗[7]

This can be rendered into English as:

[target text] Celadon bowl with foliated rim.

Other than being black lines on a white page, there is absolutely nothing in the source text that in any way resembles the target text to hint at the fact that they are the same thing. They are no more connected than the following three lines "///" are connected to the word "flower".

And yet a translation has occurred. I, the translator (with the help of an uncredited translator of a museum brochure), "saw" meaning in the source text and was able to create a target text with the same meaning. Where am I finding the meaning that links the two texts? It is not in the surface, visible depictions. Where can it be? It is always tempting to place meaning in translation as residing in-between the source and target texts. We can make this argument in two ways: by focusing on how the world is divided into universal ideas that can then be represented by words in any language (the semantic path), or by focusing on how each language is carved from a finite set of parametric choices that are common to all languages (the syntactic path). Indeed, if we were to build a translating machine, we would look to access this in-between language zone and automate the process by which source and target texts move between this zone. Following my distinction above, we could create two possible programs, a semantic one and a syntactic one.

[7] *The Lee Kilsoo Collection—Ceramics of Korea, Japan, China and Southeast Asia/委吉秀コレクション—韓国・日本・中国・東南アジアのやきもの* (Aichi Prefectural Ceramic Museum, 2008), 12. The catalog has rendered it as "Bowl with foliated rim, caledon". I moved "caledon" to the front because I feel that it can be used as a modifying adjective. I grieve that my knowledge will never be enough to know if this was the correct thing to do.

A Semantic Machine

Let us look at the semantic program first. We know, from our everyday experience, that the world is divided into different objects. Only in the realm of magic and fantasy would this not be the case. Our task, then, in programming our machine would be to identify each of these objects. For instance, let us take a small round handle-less piece of porcelain as one such object. I could, let's imagine, give it the label "1.1". I can then attach different labels to Object 1.1., such as "bowl" in English. In fact, I could use words in any other language I wanted (for example, "*mias*" in Irish, "*bovlo*" in Esperanto, or "*bol*" in French). And so when I feed the word "bowl" into my machine, it will automatically connect it to "1.1" and, if doing English to Japanese, will output "鉢", providing a successful translation. The problem, though, with such perfect in-between languages is that ideas do not attach themselves to the world of objects so easily.[8] Or rather, the problem is that the objects of the world do not attach themselves so easily to our ideas.

I can see before me, undeniably and unmistakably, an object that I know to be, without any controversy, a "bowl". It is a coherent universally recognizable physical object transcendent of all cultural perspectives. One could pick it up and throw it to or at anyone who claimed that my seeing it was merely a product of western patriarchal socio-cultural discourses of power. And yet the whole concept of "bowl" starts to get fuzzy and unclear as soon as I start to relate it to other related objects. When is a bowl not a bowl? When it is a basin? Is it defined by shape? If I make a bowl from very thin paper is it still a bowl, even if it cannot really function as a bowl? If a bowl is small and I drink tea out of it, is it still a bowl? Or is it a type of cup?

It seems, then, that the concept of "bowl" is not as solid and indisputable as the hard ceramic object sitting before me but is highly dependent on the particular local context in which it is being used. A *bowl* can become a *basin* or *cup* or a *paper craft* depending on the situation in which we are using the concept. Indeed, the more common the object, the more slippery the concept. The irony is that the more obscure and domain-specific a word is the easier it is to translate because the less chance there is of the concept having slid naturally into other concepts. For instance, when I input the words "青磁輪茶碗" into the machine translation

[8] See Eco, *The Search for the Perfect Language* (Oxford; London: Blackwell, 1995), for a fuller history of the centuries-old unsuccessful quest for a perfect pure in-between language.

program SYSTRAN (an historic leader in the field) I get "celadon tea bowl". (Which is very impressive: Google Translate gave me "celadon wheel tea bowl"). However, when I type in simply "茶碗" I get the rarified "*chawan*." But if truth be told, if I were to pick up a 茶碗 and speak of it in English, I would say "cup", "little cup", "saucer", "porcelain piece," "tea cup," "tea bowl," "thing," "this thing," "small mug," "yoke," "tankard type thing" and a dozen other words. However, I don't think I would ever say "*chawan*." (I lack the *savoir faire* for italicized words). Could a computer ever be programmed to replicate the messiness, inconsistencies, vagueness, and borderline inarticulateness of my language usage? Universal interlanguage fails because we humans are too confusing in our descriptions of the world we live in and too unpredictable in how we go about dividing it and categorizing it. We never really are knowing what we are going to say and make it up as we go along. Unlike machines, which often seem to have already made up what we never even knew we didn't want to say. Meaning is local and hinged to present usage. For a language to work as a language it does have to present itself as being absolute in its semantics (a bowl is a bowl) but this absoluteness works only in so far as the contingent use of language allows it. Machines can never get at these contingencies, and processing is always handicapped and limited by the machine's dependence on absolute semantics. To do meaning one must be conscious in a present place where all the world converges in one's interpretation of the now moment. With this, it can be said that, quite simply, machines don't do meaning.[9]

A Syntactic Machine

Another way to build our translation machine would be to separate the meaning from the syntax (so the world out there, the semantics, is not a problem) and just let computers work with and replicate the universal parameters from

[9] This seems to be the argument Hurbert Dreyfus is making in his book *What Computers Can't Do: A Critique of Artificial Reason* (New York: Harper and Row, 1972), 200: "A phenomenological description of our experience of being-in-a-situation suggests that we are always already in a context or situation which we carry over form the immediate past and update in terms of events that in the light of this past situation are seen to be significant. We never encounter meaningless bits in terms of which we have to identify the context, but only facts which are already interpreted and which reciprocally define the situation we are in. Human experience is only intelligible when organized in terms of a situation in which relevance and significance are already given. This need for prior organization reappears in AI as the need for a hierarchy of contexts in which a higher or broader context is used to determine the relevance and significance of elements in a narrower or lower context."

which, as the Chomskyanists tell us (and nobody has as yet proved them wrong) all languages are carved. A sentence in a language is arranged according to an underlying syntactic structure which is systematic and runs regardless of the words being fed into it. “Colorless celadon bowls with foliated rim sleep furiously” is semantically meaningless but grammatically correct because it follows the syntactic patterns of English. Every language has its own syntactic patterns but these are all shaped from a common universal set of parameters which are either activated or not activated according to the dictates of the particular language. At first glance, then, it may seem that this could offer a possibility for machine translation since we now have an interlanguage that is systematic and not messed up by human fuzzy logic. Removing meaning from the equation means that meaning-illiterate machines are now mean and lean and ready to play. Feed in the sentence “a bowl fell”. The machine slices it into a Noun Phrase (“a bowl”) and a Verb Phrase (“fell”), sorts it and outputs “不定の単数のカップが落ちた”. No problem. However, the limit we find with this syntactic interlanguage is that each language, as it switches on and off parameters, finds itself committed to providing certain information in its grammar that other languages do not have to give. Thus, for instance, English, because of the parameters it has inherited, is a language where singular and plural must *always* be designated. This is not the case in Japanese, for example. There is no problem with this when we go from English to Japanese. “A bowl” or “bowls” can be simply “ボウル”. But it is a problem when we go the other way. If a text consists of the Japanese word “ボウル”, the English translator does have to decide if it is one or more bowls. In other words, the further we go down the branching parameters the harder it is for us to go back up again, making translation, without semantic knowledge, an impossibility. Chomskyan linguistics does admirably explain how all languages are chiseled out of the same block. However, it does have problems explaining how translation can pass, horizontally, between the vertical branches of a sentence’s underlying syntactic structures.

It seems, then, that any machine translation that seeks to translate as humans do through the search for a pure in-between set of concepts or grammatical patternings is doomed to failure. When a text is cut into bits this actually changes each bit and they can never be put back together again. This is the mystery of translation. It happens when both the source and target texts remain completely intact. This is why the most successful automated translation programs, as measured by the erratic and irrational expectations of humans, are those that use

translations that have already been done by humans and basically copy them when other similar source texts arise.[10] In other words, computers can translate like humans only when they cheat and plagiarize those humans.

Form is Emptiness

The fact that human translation occurs even though there is no in-between meaning or common structure between languages needs to be fully grasped. Let us ask again the question, between two texts that are translations of each other where does the meaning lie? If not in-between then where? We cannot say in both, as this is simply restating the problem. For meaning to be in both, they both must have something in common. But that something in common is not, as we have seen, anywhere else but in the two texts, both of which are, in form, completely different from one another. Meditate again on the two texts: "青磁輪茶碗" and "Celadon bowl with foliated rim". See again how different they are. If you say they are the same because they both express the same "concepts" or "ideas" you are, again, trying to locate an in-between realm that does not exist anywhere—except in the two texts that are so different. Ideas and concepts are *empty* without the texts to express them. The only way beyond this impasse is to realize that the form is this emptiness, and this emptiness is the form. In other words, we must remember the words of the *Heart Sutra*: 空不異色、色不異空 ("emptiness differs not from form, form differs not from emptiness"). In other words, when we stare at two translations, so different in form, the only way they can be connected is in our consciousness. But our consciousness can do this because, unlike machines which only ever move between form and form, it can produce meaning by emptying emptiness to produce form. Humans translate because

[10] As *The Economist* reports ("Language: Finding a Voice," May 1, 2017): "Many early approaches to language technology—and particularly translation—got stuck in a conceptual cul-de-sac: the rules-based approach. In translation, this meant trying to write rules to analyze the text of a sentence in the language of origin, breaking it down into a sort of abstract "interlanguage" and rebuilding it according to the rules of the target language. These approaches showed early promise. But language is riddled with ambiguities and exceptions, so such systems were hugely complicated and easily broke down when tested on sentences beyond the simple set they had been designed for. Nearly all language technologies began to get a lot better with the application of statistical methods, often called a "brute force" approach. This relies on software scouring vast amounts of data, looking for patterns and learning from precedent. For example, in parsing language (breaking it down into its grammatical components), the software learns from large bodies of text that have already been parsed by humans. It uses what it has learned to make its best guess about a previously unseen text. In machine translation, the software scans millions of words already translated by humans, again looking for patterns."

human consciousness, that radical self-reflectivity that can only be from a point of emptiness, (where the eye cannot see the eye), where meaning, and not preprogrammed mechanical change, is generated. Machines churn out their translations in a flawless flow. They miss nothing. But that is precisely their problem.

Bibliography

Chalmers, D. J. *The Conscious Mind: In Search of a Fundamental Theory*. New York; Oxford: Oxford University Press, 1996.

Dreyfus, Hubert L. *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper and Row, 1972.

Eco, Umberto. *The Search for the Perfect Language.* Oxford; London: Blackwell, 1995.

Nishida, Kitaro. 1970. "Toward a Philosophy of Religion with the Concept of Preestablished Harmony as a Guide." Translated by David Dilworth. *The Eastern Buddhist*, 3(1): 19–46.

Nishitani, Keiji. *Religion and Nothingness*. Translated by Jan Van Bragt. Berkeley: University of California Press, 1982.

Searle, John. R. "Minds, brains, and programs." *Behavioral and Brain Sciences.* 3(3) (1980): 417-457 [Pre-print version]

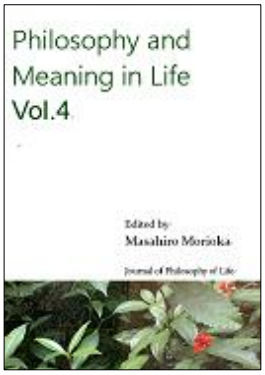
Turing, A., "Computing Machinery and Intelligence," *Mind.* 59 (236) (1950): 433–60.

Wittgenstein, Ludwig. *Philosophical Investigations.* Translated by G. E. M. Anscombe. Oxford: Blackwell, 1958.

Žižek, Slavoj. *Organs without Bodies: On Deleuze and Consequences*. New York; London: Routledge, 2004.

Information about the Authors

Shin-ichiro Inaba
Professor, Meiji Gakuin University.
Masahiro Morioka
Professor, Human Sciences, Waseda University.
Makoto Kureha
Associate Professor, Faculty of Global and Science Studies, Yamaguchi University.
István Zoltán Zárdai
Visiting researcher, Department of Ethics, Faculty of Letters, Keio University.
Minao Kukita
Associate Professor, Graduate School of Informatics, Nagoya University.
Shimpei Okamoto
Assistant Professor, School of Letters, Hiroshima University.
Yuko Murakami
Professor, Graduate School of Artificial Intelligence and Science, Rikkyo University.
Rossa Ó Muireartaigh
Associate Professor, Department of British and American Studies, School of Foreign Studies, Aichi Prefectural University.



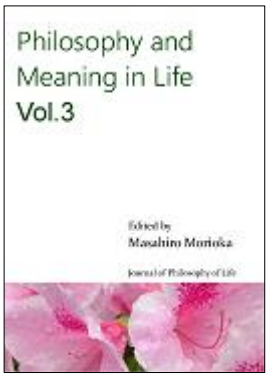
Open Access Book

Philosophy and Meaning in Life Vol.4

: Selected Papers from the Pretoria Conference (2022)

Cheshire Calhoun, Masahiro Morioka, Kiki Berk, Frank Martela, Patrick O'Donnell, Kazuki Watanabe

http://www.philosophyoflife.org/jpl2022si_book.pdf (Free Download)



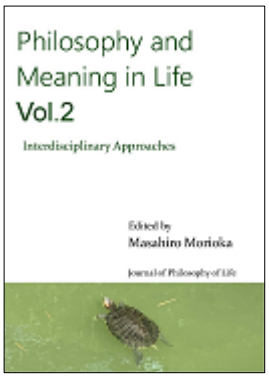
Open Access Book

Philosophy and Meaning in Life Vol.3

: Selected Papers from the Birmingham Conference (2021)

Drew Chastain, Mirela Oliva, Masahiro Morioka, Kiki Berk, Benjamin Murphy, Jairus Diesta Espiritu, Aaron Brooks, Heidi Cobham

http://www.philosophyoflife.org/jpl2021si_book.pdf (Free Download)



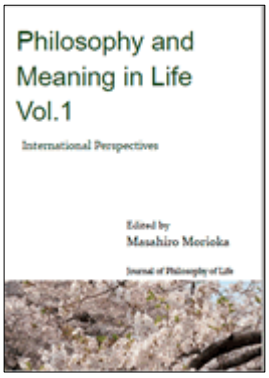
Open Access Book

Philosophy and Meaning in Life Vol.2

: Interdisciplinary Approaches (2020)

John Partridge, Andrew Tyler Johnson, Andrew M. Winters, Joshua Chang and Michelle Pitassi, Laÿna Droz, Lehel Balogh, Nathaniel Serio

http://www.philosophyoflife.org/jpl2020si_book.pdf (Free Download)



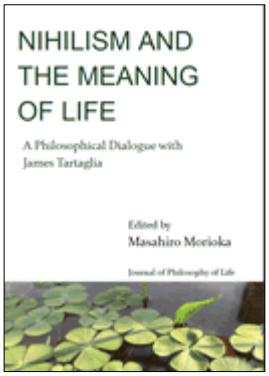
Open Access Book

Philosophy and Meaning in Life Vol.1

: International Perspectives (2019)

Brooke Alan Trisel, Angel On Ki Ting, Isabel G. Gamero Cabrera, Jairus Diesta Espiritu, Sara Chan Yun Yu, Masahiro Morioka

http://www.philosophyoflife.org/jpl2019si_book.pdf (Free Download)



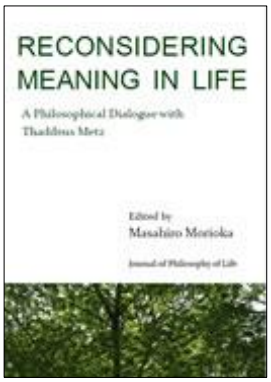
Open Access Book

Nihilism and the Meaning of Life

: A Philosophical Dialogue with James Tartaglia (2017)

James Tartaglia, Adam Balmer, Philip Goff, Ronald A. Kuipers, Tracy Llanera, Alan Malachowski, Bjørn Torgrim Ramberg, Brooke Alan Trisel, J. J. Valberg, Damian Veal, Sho Yamaguchi

http://www.philosophyoflife.org/jpl2017si_book.pdf (Free Download)



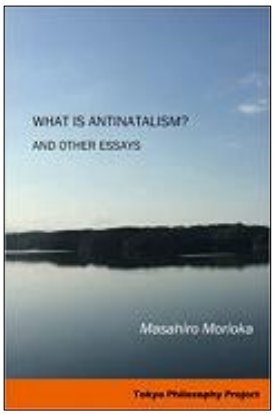
Open Access Book

Reconsidering Meaning in Life

: A Philosophical Dialogue with Thaddeus Metz (2015)

Thaddeus Metz, Hasko von Kriegstein, David Matheson, Peter Baumann, Masahiro Morioka, Sho Yamaguchi, James Tartaglia, Christopher Ketcham, Fumitake Yoshizawa, Nicholas Waghorn, Mark Wells, Jason Poettcker, Minao Kukita, Yu Urata

http://www.philosophyoflife.org/jpl2015si_book.pdf (Free Download)



Open Access Book

What Is Antinatalism? And Other Essays

: Philosophy of Life in Contemporary Society (2021)

Masahiro Morioka

https://www.philosophyoflife.org/tpp/antinatalism.pdf (Free Download)

ISBN 978-4-9908668-9-1

Journal of Philosophy of Life