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)