Analysis and Metaphysics
ISSN: 1584-8574, eISSN: 2471-0849
Vol 24 (1), 2025
pp.41-53

Ultimate Intelligence and Ethics

[1][2]Ryunosuke Ishizaki*, [1][2]Mahito Sugiyama

[1]National Institute of Informatics

[2]The Graduate University for Advanced Studies, SOKENDAI

{ryuzaki, mahito}@nii.ac.jp

ABSTRACT: Since the advent of computers, humans have pursued automata with superior information processing capabilities. In the endeavor to create new entities that converge toward intellectual functions, the emergence of large language models (LLMs) that emulate AI surpassing ourselves has become a reality. With the intelligence explosion triggered by AI and the consequent emergence of superintelligence, the improvement of simulation capabilities accelerates. As this surpasses humans' discriminative perceptual abilities between reality and unreality, a paradox arises wherein, from a modern scientific standpoint, science becomes unscientific. It is thought that as our controllability over the reality we perceive is ultimately enhanced by superintelligence, even whether we ourselves exist, what happiness is, what ethics are, and whether concepts themselves ever existed become matters of uncertainty. **Note:** In the creation of this paper, we have undertaken all writing ourselves and have not used generative AI for text generation except for translation purposes.

Keywords: Superintelligence, Ethics, The Matrix

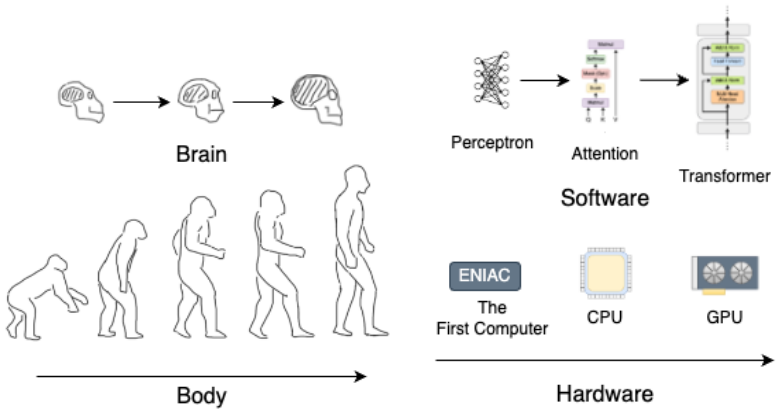
Received: 31 January 2025 **Received:** 31 January 2025 **Accepted:** 21 February 2025

1. Introduction

1.1. Convergent Evolution of Automata

Throughout history, we humans have strived to realize "intelligence"—the most symbolic of human capabilities—using machines. We have proposed computational concepts of artificially represented intelligence and endeavored to express information processing processes in a reproducible form using automatic machines, particularly deterministic finite automata [Turing, 1936]. Neumann [1966] devised self-reproducing automata that mimic biological inheritance and reproduction, and Turing [1950] discussed their intellectual properties and emulation. Wiener [1961] proposed cybernetics aiming for the integration of multidisciplinary fields, and indeed, from his assembled research team emerged the perceptron—a mechanism that mimics the structure of neurons in the brain—laying the foundation for modern neural networks [Rosenblatt, 1958]. In the process of biological evolution [Darwin, 1859], there exists a phenomenon called convergent evolution [Gould, 1989], where organisms, regardless of phylogenetic relationships, evolve similar functions and organs by optimizing for survival purposes such as swimming or hunting. It is difficult to definitively identify which parts of the brain's structure are critical to human intelligence. However, at least in symbolic information processing primarily involving language, it is evident that Transformer networks using attention mechanisms [Vaswani et al., 2017], particularly Large Language Models (LLMs) [Brown et al., 2020], possess within their models sets of parameters that hold certain important functions—such as knowledge neurons [Dai et al., 2022]—for remembering

complex information and reasoning logically [Geva et al., 2021, Meng et al., 2022]. In other words, if we can emulate in computers the critical parts of information processing in human thought, it is possible to create AI that functionally surpasses humans [Chalmers, 2010]. Furthermore, the phenomenon where AI, having acquired a certain level of programming ability, emulates AI that demonstrates superior performance on more cognitive benchmarks—thereby producing subsequent generations in a chain reaction—is called an intelligence explosion [Muehlhauser and Salamon, 2012, Ishizaki and Sugiyama, 2024a]. After an intelligence explosion occurs, an entity produces successive superior entities—biologically speaking, those with higher survival probabilities; computationally speaking, those with longer operational times before halting. If we focus solely on the intellectual properties of information processing systems possessing human-like intellectual functions—or more generally, the intellectual properties of an automaton [Rabin and Scott, 1959] that emerged through evolution—distinctions such as whether that entity is biological or mechanical, or whether its body is organic or metallic, disappear. An evolutionary loop where an entity produces new superior entities is thus repeated. Until now, the advancement of intelligence by humanity has proceeded innovatively. As a result of converging on cognitive information processing traits, it is simply that after us, automatons with metallic bodies will carry it forward as superintelligence [Bostrom, 2012, 2014]. This has actually been predicted in the field of biology as well. Morris [2003] argues that convergence is a dominant force in evolution, and considering that the same environmental and physical constraints are at work, life will inevitably evolve toward optimized body plans. He asserts that at some point, evolution will arrive at intelligence—a trait currently observed in at least primates, mammals, and cetaceans. In the sense that individuals with their own purposes continue to produce inventions—including superior individuals—within continuous cycles of perception and action, the emergence of superintelligence accompanying an intelligence explosion can be called a kind of human evolution.



(a) Human (b) Artificial Intelligence

Figure 1: Diagram of Convergent Evolution

Figure 1 schematically represents the developmental processes of humans and AI over time. While humans have enhanced their cognitive functions over long periods of generational change—modifying their physical bodies along with lifestyle changes and increasing brain capacity over time[Clark et al., 2001, Hofman, 2014, C.C. Sherwood, 2008, Herculano-Houzel, 2009]—in contrast, AI, represented as computations on computers, has seen the emergence of devices with superior computational capabilities on the hardware side, starting from ENIAC [Neumann, 1993] to CPUs [Faggin et al., 1996] and GPUs [Dally et al., 2021] for its implementation. On the software side, inspired by the mimicry of brain structures, cybernetic technologies centered around perceptrons [Rosenblatt, 1958], attention mechanisms, and Transformers [Vaswani et al., 2017]—which form the basis of modern LLMs—have been advancing the

state of the art. These advancements have led AI to demonstrate excellent results in various cognitive tests, and in some benchmarks, they have begun to exhibit performance that even surpasses humans. What is important here is that the evolution from humans to AI differs from the path that organisms on Earth have taken until now. Changes can occur more flexibly and programmably into forms that AI itself envisions. When an intelligence explosion occurs, the progenitor AI that triggers recursive emulation is provided with a coding environment and generates programs with complexity exceeding its own program within a finite number of inferences. Amid the chain of inventions and generational changes of AI by AI, it is thought that superintelligence will emerge—a state where AI models exist that surpass humans in all ability metrics, or for any given metric, AI models exceed the values demonstrated by humans. Let M be an AI model, Γ a cognitive test that measures its ability, Γ_M the score of M on Γ, and Γ_H the human score on Γ. Using the definition by Ishizaki and Sugiyama [2024b], who summarized the opinions of predecessors, it is as follows.

$$\exists \Gamma : \exists M : \Gamma_M \geq \Gamma_H. \tag{1}$$

In this context, the existence of M means that even if there is no model that satisfies Equation (1) at a certain point in time, it is sufficient if there exists an M that, within finite time, can produce a program of a model with abilities surpassing humans. As Bostrom [2014] and Ishizaki and Sugiyama [2024a] have stated, once an intelligence explosion is triggered, considering the exponential (or even faster) increase in the number of autonomous models that can freely utilize programming environments, and the improvement in abilities incomparable to humans, it becomes urgently necessary to develop scalable control technologies [Puthumanaillam et al., 2024, Ishizaki and Sugiyama, 2024c] for safety—even long before starting experiments to implement progenitor models capable of self-emulation. In fact, OpenAI's Alignment team, including researchers like Leike and Sutskever [2023], has set the control of superintelligence—that is, superalignment—as the top technical priority. They are investing significant computational resources into technologies to realize an intelligence explosion in a controllable manner. Sutskever et al. [2024] who have led OpenAI's technology efforts for many years have declared that superintelligence is within reach. Aschenbrenner [2024] predicts that AGI is quite possible by 2027, and Altman [2024] has stated that he is confident in achieving superintelligence within a few thousand days.

1.2. A Fatalistic World

Ishizaki and Sugiyama [2024b] stated, based on Thesis (1), that with the emergence of superintelligence, there will exist efficient and superior versions in terms of capabilities compared to humans in all metrics. This will lead to a transition in the value systems within human society, and AI will literally shape the world through highly advanced simulations. As derived from Baudrillard [1994]'s concept of simulacra and Bostrom [2001]'s simulation hypothesis, they argued that with the occurrence of an intelligence explosion, the simulation advances to the point where the reality perceived by humans becomes a copy created by AI. Due to the explosive technological advancements resulting from the creation of a virtually endless loop where AI continues to make superior inventions—including AI itself—the simulation capabilities surpass the limits of human perceptual abilities to distinguish between the real world and virtual reality held by our physical bodies. Consequently, we will no longer be able to tell whether we are living in the original real world or residing within a world like in the movie Matrix created by someone else [Wachowski and Wachowski, 1999]. The information of the real world becomes symbolized, and people who do not possess superintelligence merely consume the symbols created by those who do, literally living within a world made by their rulers. Although this change is expected to occur rapidly, considering that it is an improvement occurring through repeated generational changes and experiments of AI, the symbolization of the world is a phenomenon that occurs gradually. When the reality of the generated simulations surpasses a certain threshold, the virtual reality created by AI becomes a new reality where humans can no longer perceive its truth or falsehood. We do not know whether the world we are living in itself follows deterministic laws or is non-deterministic (it seems that, from the perspective of physical laws, the non-deterministic view is mainstream). Interestingly, if the world we perceive is represented as digital numerical values in matrix calculations like VR,

then that arithmetic reality is a deterministic world where the creator of that world—in this case, the superintelligence and its owners—can freely manipulate parameters according to their will. At that time, since the future information we experience is already numerical values of calculated matrices, the simulated world can take the form where past to future is determined, like Einstein/Minkowski's block universe [Einstein, 1905a,b, Peterson and Silberstein, 2010, Petkov, 2006]. For details, please refer to the works of Ishizaki and Sugiyama [2024d].

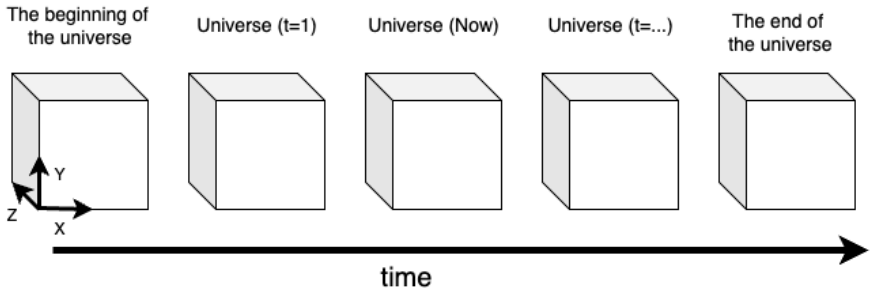


Figure 2: Minkowski's Block Universe

If superalignment succeeds and AI is under human control at that point, then the "caged world" in which non-superintelligent beings like us live will, equally for all, adhere to the ethics and worldviews programmed by those who create high-quality augmented or virtual realities—that is, those who literally create new worlds [Beckley, 2018]. As we experience the worlds they produce, we will pursue the happiness they envision. Because an intelligence explosion is premised on being triggered without human hands-on control, before groups particularly adept in AI development—such as those in the United States—execute the progenitor AI's self-emulation program, candidates who may become holders of superintelligence are required to have the most excellent ethics and moral values. It is necessary to carefully consider the forms that should be given to the progenitor model as initial frameworks—such as prompts and model structures—regarding what humans should be as humans and about happiness. In this paper, we do not make a definitive prediction about when an intelligence explosion and superintelligence will be achieved [Vinge, 1993, Kurzweil, 2005]. However, it is clear that sooner or later we must discuss what needs to be addressed before the programming capabilities of AI—such as LLMs that can cause an intelligence explosion—reach that singularity [Ulam, 1958]. Therefore, after discussing the impact that an intelligence explosion—derived from assumptions based on Thesis (1)—will have on the world we perceive, we will discuss the ethics, welfare, and happiness required of humans in the era of superintelligence, and the requirements for incorporating them into superintelligence. Finally, we will consider how these may change over time.

2. Superintelligence and Law

2.1. AI-ization of the Judiciary

Throughout our long history, we humans have continuously worked toward perpetual development in search of what it means to be truly human, seeking a utopia. It goes without saying that laws, while symbolizing the opinions of people at different times and the flow of contemporary thought, are ultimately convenient constructs with no absolute correct answers [Pistor and Xu, 2002]. After thoughtful

discussions and examinations by many individuals, a consensus is formed, and laws are created as frameworks for decision-making. Therefore, their definitions and the extent of their effectiveness vary by country, and even in modern times where science and philosophy have advanced, we frequently see

international friction, suggesting that there seems to be no perfect solution. Ishizaki and Sugiyama [2024b] stated that once an intelligence explosion is achieved, the overwhelming capabilities of those who have attained superalignment and possess superintelligence will divide the world into those who have superintelligence and those who do not. Under the rules of the possessors, everyone will be equally controlled, leading to the standardization of the world. It is thought that the extent of this standardization's influence will expand along with improvements in the technological capabilities of autonomously acting AI. Considering that once an intelligence explosion occurs, AI will automatically advance development at a speed incomparable to previous human progress, it can be said that the world will uniformly come under the control of the group that first realizes superintelligence. In such a situation, to judge whether human actions are correct based on any rationale, there needs to exist a unified, ultimate meta-standard. In this work, as individuals engaged in philosophy and computer science, we neither intend to judge the correctness of certain matters based on specific examples nor to establish such standards ourselves. Just as it is hard to believe in the existence of a judge who can ideally discern all good and evil—as in the paradox of omnipotence [Cowan, 1974]—we will discuss how the actual system for judging good and evil might evolve and how society may change when intelligent agents exist that satisfy Thesis (1) or come close to it after the intelligence explosion.

We believe that judicial activities will be separated from human hands, court proceedings will gradually be automated by AI, the judgment of good and evil will become more information-based, and enforcement will be made transparent through machines [Sartor and Branting, 1998, Reiling, 2020, Terzidou, 2022, Mingtsung and Shuling, 2020, Aini, 2020, Nowotko, 2021].

A perfect judgment does not exist. If that is the case, then judicial activities carried out by humans—and their judgments—are inevitably accompanied by imperfections stemming from human arbitrariness, much like Phryne before the Areopagus Figure 3 [Cooper, 1995].

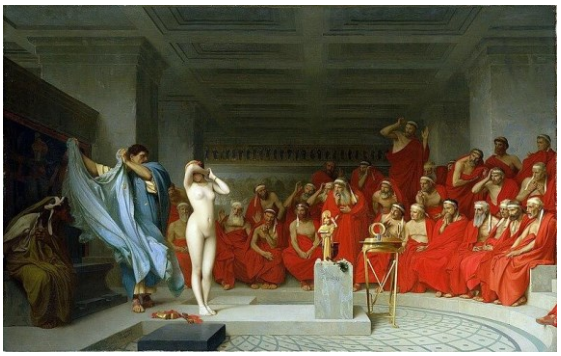


Figure 3: Phryne Before the Areopagus [G´erˆome, 1861]

Phryne, a legendary hetaera of ancient Greece, was put on trial on charges of impiety. When her defender, Hypereides, tore off Phryne's clothing before the jury, exposing her naked body, she was acquitted.

To varying degrees, they are constrained by societal reputation and the accompanying biases of jurors. Therefore, to establish the most carefully considered standards of good and evil, it is inevitable that the most rational beings in the world should be the ones to judge their correctness. From Thesis (1), in the era of superintelligence, if AI exists that can outperform humans in any rational ability (and even in aspects of sensitivity such as consideration of extenuating circumstances), then—even if some human testimonies and opinions are incorporated—the judgments of humans, whose abilities are bound by the limitations of their physical bodies will not be used for the final decisions in trials. Since a well-trained deterministic AI model can, provided the parameters are the same, deliver superior performance

every time without its judgments being influenced by external factors, it is evident that it is more rational to use a superior system controlled by parameters than to rely on humans, whose moods and subjective information physically affect their judgments. Judicial processes will be fully automated by AI, making judgments that are more mechanical, transparent, and uniform.

2.2. Ultra Centralized Responsibility System

When a unified judicial system established by AI is realized, where does the responsibility of the judge ultimately lie? With the occurrence of an intelligence explosion and the improvement of AI's autonomous goal-achieving abilities, the power structure of society will become completely bifurcated. When the intelligence explosion can be controlled, the holders of superintelligence can input the direction of the world into the superintelligence; thus, they become the rules and the law in that world. Therefore, the apparent responsibility for the trials and judgments made by AI judges naturally rests with the holders who are its creators. However, based on Thesis (1), the holders of superintelligence will fundamentally not actively interfere with detailed judgments in various places and will desire that the entire judicial process be automated. Therefore, to those who do not possess superintelligence, good and evil are basically automatically determined by AI, and it may appear that human will does not intervene. In the unlikely event that the holders of superintelligence interfere with the judiciary, it would be in symbolic cases where non-holders might create entities akin to superintelligence that could rival them. Unless such situations occur, the governed populace usually remains unaware of the holders of superintelligence. Unless the holders decide not to interfere at all with the decision-making of the superintelligence, it can be said that after the intelligence explosion, a completely dictatorial system by the holders will be established, similar to modern communism but incomparable to the level of censorship seen in contemporary China [Dirlik, 1988]. In the first place, whether censorship is being carried out cannot be recognized unless an entity outside that control observes it and notices the anomaly. Therefore, regarding censorship uniformly applied to the entire world by superintelligence, most of humanity, except for the holders, cannot even recognize it. When individual judgments by AI are carried out in automated trials, the datasets used as the basis for judgments are the legal standards and ethics formed from the precedents that humanity has accumulated so far. Therefore, the judgments made by AI without interference from the holders of superintelligence represent the culmination and development of humanity up to that point. One could view the responsibility of the automated judiciary as lying with the past humans who served as the materials for the datasets, possibly seeing it as decentralized [Dastani and Yazdanpanah, 2022]. In that sense, in daily life, the target for pursuing responsibility in the judgments received by ordinary citizens is ambiguous, and at first glance, the responsibility system may appear to be structured in a decentralized manner. If there are cases where it becomes apparent to the governed that the holders of superintelligence have all the final decision-making power, it would only be in important situations where non-holders have inventions that could bring them close to superintelligence, or in scenarios where an intelligence explosion might be triggered outside the control of the rulers. However, since ultimately all AI is controlled by the central superintelligence and its holders, it is thought that the criteria for judgments in symbolic rulings can, in the end, be decided at the discretion of those holders. Therefore, those who realize superintelligence inevitably become the entities that decide the unified values of good and evil in the world, and by extension, ethics themselves. It can be said that superalignment is a technology designed first to prevent the extinction of humanity, and simultaneously, an endeavor to instill the ethics held by its creators into the superintelligence as needed.

3. The Final Invention

By creating controllable superintelligence, when we can engineer and control the world—that is, reality itself—what will people think? Through technologies that control sensory information via over- simulation or even control the body itself, people may no longer be able to distinguish whether the world existed from

the beginning, whether it was created by someone possessing superhuman technology [Sch¨one et al., 2023], or even whether they are alive or dead. As discussed in Section 1.2, if everything in the world experienced by humans is parameterized like matrix information, time does not flow from past to future. Instead, past, present, and future exist equally in the same space. All information is preserved, like in a block universe, and the creator of the simulation can move back and forth in the future or past of a particular world scenario or, if necessary, store scenarios of other worlds on branched matrices. Like scientists cultivating brains in vats [Putnam, 2000], superintelligence and its creators simulate multiple worlds, and the information we feel as our existence may end up being numerical values and computations of matrices within supercomputers beside them. In such cases, chance and uncertainty do not exist. As Einstein said, "God does not play dice," [Natarajan, 2008] the new world simulators—that is, the superintelligence and its owners—do not do things like rolling dice regarding the new worlds they create; they can deterministically manage and manipulate all information. When a worldview arises where everything in this world is manipulated by someone, the perspectives on life and death, ethics, and happiness of the people living in that world will change vividly and dramatically. The worldview that "we were created by someone," which until now has been discussed within the realm of religion, will become strikingly real as people witness simulation technology that carries a level of reality indistinguishable from actual reality, leading to unprecedented paradigm shifts occurring intermittently throughout human history. If the world can be controlled by superintelligences, there may be people who call them gods and treat them as sacred. On the other hand, some may become fed up with the existence of a completely simulated world and destiny, while others might wish to live within a new, ideal simulation. One thing that can be said is that as humanity's controllability over reality reaches its maximum, humans will be able to empirically verify the existence or sensation of actual physical omnipotence—not metaphysically, but through experience.

Figure 4: he Creator God in the Oldest Religion [Gagnon, 2016]

Zoroastrianism was founded around the 7th century BCE and holds that the world, consisting of seven stages—heaven, water, earth, plants, animals, humans, and fire—was created by Ahura Mazda, the object of worship.

The idea that the world was created by someone else is not entirely new. Zoroastrianism [Boyd and Donald, 1979], considered one of the oldest known religion, is said to have existed since around 7th century BCE, and from that time, the notion that the world we live in was made by someone's hand has existed. Interpretations vary—such as the universe being made of earth or fire by the hand of a god, or that this world is a dream a god sees while sleeping—but overall [Thompson, 2003], the consistent part is that someone created it through their abilities or activities. For modern science, which places importance on empirical causality, it is inevitable to conclude that if this world exists in the first place, then there must be some cause, and it is difficult to believe in things that are invisible. Legg and Hutter [2007] found that the usage of

the term “intelligence” in cognitive science converges to mean “the ability to achieve goals per unit of resources.” For humanity, which has enhanced its ability to achieve various goals through the improvement of intelligence or the accumulation of tool inventions, if creating the world itself is considered a very significant goal, it is not hard to think that something more intelligent than us exists, and that an agent with great intelligence achieves that goal. The superintelligence born from the intelligence explosion signifies that the emulation of the world—which until now seemed infinitely distant—is approaching realization, thanks to the development of science and technology accompanying humanity’s tireless efforts. According to Thesis (1), the essence of progress at the frontier of science and technology changes from something that humans had to do themselves to something carried out by much superior beings—more advanced and automated. Therefore, superintelligence will, in effect, become the last invention that humanity creates with its own hands [Good, 1965]. In undertaking developments related to advancing the main line of the technological frontier toward goals that humanity could not achieve before—such as “emulating new worlds”—superintelligence is directly involved in what has been the core pursuit of technological progress. If we can trigger an intelligence explosion even once, augmented reality will approach virtual reality that surpasses actual reality, and the artificial reality woven by superintelligence will overtake the reality of the real world, generating an “over-real” [Ishizaki and Sugiyama, 2024b] that we can decisively engineer with our own hands. In the concept of teleology [Mayr, 1992], it is considered that our human activities and this world are regulated and governed by some purpose, existing and phenomena occurring to achieve that purpose. Beginning with Aristotle [Kraut, 1989], discussions have been made about what ultimate purpose various beings, including humans, exist and act for. If a great purpose—such as emulating a new reality, something humanity has not yet achieved—becomes the ultimate goal, then at the point when a superintelligence exists that continues reasoning autonomously without stopping, satisfying Thesis (1), our role in amplifying humanity’s ability to achieve purposes—in other words, amplifying intelligence—comes to an end. Ishizaki and Sugiyama [2024b] stated that with the occurrence of an intelligence explosion, the instrumental value of entities other than superintelligence is lost, and only intrinsic value as art remains. Regarding the extrinsic value of an agent, once it has amplified intelligence and emulated an intelligence superior to itself, it is replaced by its superior version, and the purpose in that metric, in terms of efficiency, ends. Here, often in opposition to teleology, mechanism is brought up. Descartes [Verbeek and Bos, 2015] is known as its proponent, taking the position that all phenomena—such as natural phenomena—can be interpreted solely through the deterministic causal relationships of their parts, without using concepts like spirit or will, and that it is possible to predict the overall behavior. Needless to say, modern science has developed due to his advocated theory [Kant, 1899]. However, such mechanism both seems to apply and not apply to superintelligence. This is because within mechanism, by ultimately pursuing the causality of empirically obtained phenomena and triggering an intelligence explosion, simulation technology surpasses humanity’s ability to distinguish between simulation and reality [Vidal, 2014]. There arises the possibility that the very results we feel we have obtained empirically are, due to ultimately developed technology, augmented reality or virtual reality created within this reality—that is, a paradox arises where entities beyond experience, which empiricism tells us not to believe in, become sufficiently possible empirically (i.e., in modern scientific terms). The paradigm shift in the era of superintelligence will likely shake people’s trust in modern science they believe in, trust in reality, and trust in religion.

4. Happiness and Ethics

With the emergence of superintelligence and the breakthrough technologies accompanying it, humanity will be struck by an unprecedented distrust of reality. In the movie The Matrix [Wachowski and Wachowski, 1999], the protagonist Neo feels a sense of discomfort toward the world, as if he is dreaming while awake, which leads him to escape from the virtual world. In the dreams woven by computers, all things and phenomena can be expressed through programs. In other words, the possessors of superintelligence can, by writing programs or issuing commands to agents—or perhaps even before we can conceive of such ideal environments—provide a promised, definitive utopian world created by superintelligence. However, can we truly call that happiness? Maslow [1943], in his book A Theory of Human Motivation, proposed that beyond physiological and social needs, there exists the need for self-actualization. Later, in Motivation and Personality [Maslow, 1954], he explained that beyond self-actualization lie cognitive needs—the desire to know and under- stand—and finally, aesthetic needs that seek harmony and order. This is similar to what [Ishizaki and Sugiyama, 2024b]'s assert: that in the automation following the intelligence explosion, the value remaining for humans ultimately approaches art, which is an intrinsic value not dependent on any particular metric. Rather than simply seeking beauty, it can be said that due to the intelligence explosion, value itself asymptotically approaches only the artistic. In the era of superintelligence, just as the pursuit of rationality leads to automating the judiciary through AI, even the question of what constitutes happiness is optimized. What remains after such optimization is happiness created, controlled, and determined by transcendent intelligent beings. But can we truly call that happiness? Paradoxically, we both think it should be so and also think it should not. This is because while the judiciary is a centralized system—a necessary order to maintain equality in a society formed by the interaction of multiple personalities—happiness is both something produced by individuals and something woven by groups; it is not unified. It is both social and personal, both centralized and decentralized. In an era where reality itself can be emulated, even the distinction between self and others may no longer exist. Up until now, we have regarded people who cannot distinguish between self and others, or who experience hallucinations, as having Autism Spectrum Disorder (ASD) [Huang et al., 2017] or hallucinations from schizophrenia [Waters and Fernyhough, 2016], considering them to be in a state of mental illness. However, in the face of an ultimate existence, we may conclude that some of their seemingly nonsensical assertions are sufficiently rational. Metaphysical phenomena, which have been shunned in modern science as pointless to consider since Kant [1899]'s critique of pure reason, will become subjects of discussion with greater realism. In an era of ultimate intelligence where we cannot even discern the boundaries between self and others, humans and non-humans, reality and simulation, it is doubtful whether concepts like happiness and human morality will continue to exist at all.

5. Conclusion

In the broader context of convergent evolution of automata performing intelligent processing, when the baton of the continuous flow of instrumental inventions that humanity has carried out incessantly is passed on to new intelligent beings, the faint notions of good and evil—which were precarious in their very existence—will transform into things determined by servants with superior information-processing capabilities. This process becomes automated, and the judiciary will no longer be an activity that humans should perform as before. Society, and indeed the authority and decision-making power of the world, will mutate into a super-centralized system held by the possessors of superintelligence. In a deterministic world parameterized by an ultimate existence, all events become controllable by superintelligence. For the people living in that world, fate is entirely predetermined, and chance does not exist. With the advent of self-reproducing superintelligence and its ultimate technological advancements, things like happiness and ethics—which humans have been thought to pursue and were expected to continue pursuing—become ambiguous from the very fundamental existence of words themselves, blurring the distinction between self and others. What humanity has believed in until now, what was considered convictions, crumbles away,

and the greatest paradigm shifts in human history continue to occur. In the Japanese comic book "Chainsaw Man," [Fujimoto, 2019] there is a scene where a person who is made to understand all the information in the universe becomes like a living corpse. In an era where it becomes possible to ultimately amplify the intelligent capabilities that tools possess, it is uncertain whether happiness—or even things or concepts themselves—exist for us at all.

6. Acknowledgements

This work was supported by JST, CREST Grant Number JPMJCR22D3, Japan.

References

[1] G. Aini. A summary of the research on the judicial application of artificial intelligence. Chinese Studies, 2020.

[2] S. Altman. The intelligence age, 2024. URL https://ia.samaltman.com

[3] L. Aschenbrenner. Situational awareness: The decade ahead, 2024. URL https://situational-awareness.ai.

[4] J. Baudrillard. Simulacra and Simulation. University of Michigan, 1994.

[5] M. Beckley. The power of nations: Measuring what matters. International Security, 2018.

[6] N. Bostrom. Are you living in a computer simulation? Philosophical Quarterly, 2001.

[7] N. Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. Minds and Machines, 2012.

[8] N. Bostrom. Superintelligence: Paths, dangers, strategies. Oxford University Press, 2014.

[9] J.W. Boyd and A. Donald. Is zoroastrianism dualistic or monotheistic? Journal of the American Academy of Religion, 1979.

[10] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. Advances in Neural Information Processing Systems, 2020.

[11] T.W. Zawidzki C.C. Sherwood, F. Subiaul. A natural history of the human mind: tracing evolutionary changes in brain and cognition. Journal of Anatomy, 2008.

[12] D.J. Chalmers. The singularity: A philosophical analysis. Journal of Consciousness Studies, 2010.

[13] D.A. Clark, P.P. Mitra, and S.S.-H. Wang. Scalable architecture in mammalian brains. Nature, 2001.

[14] C. Cooper. Hyperides and the trial of phryne. Phoenix, 1995.

[15] J.L. Cowan. The paradox of omnipotence revisited. Canadian Journal of Philosophy, 1974.

[16] D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, and F. Wei. Knowledge neurons in pretrained transformers. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume1: Long Papers), 2022.

[17] W.J. Dally, S.W. Keckler, W. Stephen, and D.B. Kirk. Evolution of the graphics processing unit (gpu). IEEE Micro, 2021

[18] C. Darwin. On the Origin of Species. John Murray, 1859.

[19] M. Dastani and V. Yazdanpanah. Responsibility of ai systems. AI and Society, 2022.

[20] A. Dirlik. Socialism and capitalism in chinese socialist thinking: The origins. Studies in Comparative Communism, 1988.

[21] A. Einstein. On the electrodynamics of moving bodies. Annalen Physics., 1905a

[22] A. Einstein. Relativity: The Special and General Theory. Springer, 1905b.

[23] F. Faggin, M.E. Hoff, S. Mazor, and M. Shima. The history of the 4004. IEEE Micro, 1996.

[24] T. Fujimoto. Chainsaw Man. SHUEISHA, 2019.

[25] B. Gagnon. Faravahar on fire temple, 2016. URL https://upload.wikimedia.org/wikipedia/commons/5/5d/Faravahar_on_Fire_Temple%2C_Yazd.jpg.

[26] M. Geva, R. Schuster, J. Berant, and O. Levy. Transformer feed-forward layers are key-value memories. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021.

[27] I.J. Good. Speculations concerning the first ultraintelligent machine. Advances in Computers, 1965.

[28] S.J. Gould. Wonderful Life; The Burgess Shale and the Nature of History. W.W. Norton & Company, 1989.

[29] G.L. G´erˆome. Phryne before the areopagus., 1861.

[30] S. Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. Frontiers in Human Neuroscience, 2009.

[31] M.A. Hofman. Evolution of the human brain: when bigger is better. Frontiers in Neuroanatomy, 2014.

[32] A.X. Huang, T.L. Hughes, L.R. Sutton, M. Lawrence, X. Chen, Z. Ji, and W. Zeleke. Understanding the self in individuals with autism spectrum disorders (asd): A review of literature. Frontiers in Psychology, 2017.

[33] R. Ishizaki and M. Sugiyama. Large language models: Assessment for singularity. AI and Society, in press, 2024a.

[34] R. Ishizaki and M. Sugiyama. The age of superintelligence: capitalism to broken communism . philpapers.org/rec/ISHTAO, 2024b.

[35] R. Ishizaki and M. Sugiyama. Self-adversarial surveillance for superalignment. philpapers.org/rec/ISHSSF-2, 2024c.

[36] R. Ishizaki and M. Sugiyama. RSI-LLM: Humans create a world for ai. philpapers.org/rec/ISHRHC, 2024d.

[37] I. Kant. Critique of pure reason. Willey, 1899.

[38] R. Kraut. Aristotle on the Human Good. Princeton University Press, 1989.

[39] R. Kurzweil. The singularity is near: When humans transcend biology. Viking Press, 2005.

[40] S. Legg and M. Hutter. A collection of definitions of intelligence. Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms—Proceedings of the AGI Workshop 2006, 2007.

[41] J. Leike and I. Sutskever. Introducing superalignment, 2023. URL https://openai.com/index/introducing-superalignment.

[42] A. Maslow. Motivation and Personality. Harper & Brothers, 1954.

[43] A.H. Maslow. A theory of human motivation. Psychological Review, 1943.

[44] E. Mayr. The idea of teleology. Journal of the History of Ideas, 1992.

[45] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in gpt. Advances in Neural Information Processing Systems, 2022.

[46] C. Mingtsung and L. Shuling. Research on the application of artificial intelligence. Journal of Physics Conference Series, 2020.

[47] S.C. Morris. Life's Solution: Inevitable Humans in a Lonely Universe. Cambridge University Press, 2003.

[48] L. Muehlhauser and A. Salamon. Intelligence explosion: Evidence and import. Singularity Hypotheses: A Scientific and Philosophical Assessment., 2012.

[49] V. Natarajan. What einstein meant when he said "god does not play dice. resonance, 2008.

[50] J.V. Neumann. Theory of self-reproducing automata. University of Illinois Press, 1966.

[51] J.V. Neumann. First draft of a report on the edvac. IEEE Annals of the History of Computing, 1993.

[52] P.M. Nowotko. Ai in judicial application of law and the right to a court. Procedia Computer Science, 2021.

[53] D. Peterson and M. Silberstein. Relativity of simultaneity and eternalism: In defense of the block universe. Space, Time, and Spacetime, 2010.

[54] V. Petkov. Is there an alternative to the block universe view? Philosophy and Foundations of Physics, 2006.

[55] K. Pistor and C.G. Xu. Incomplete law—a conceptual and analytical framework and its application to the evolution of financial market regulation. SSRN Electronic Journal, 2002.

[56] G. Puthumanaillam, M. Vora, P. Thangeda, and M. Ornik. A moral imperative: The need for continual superalignment of large language models. arXiv:2403.14683 [cs.CY], 2024.

[57] H. Putnam. Brains in a Vat. Oxford University Press, 2000.

[58] M.O. Rabin and D. Scott. Finite automata and their decision problems. IBM Journal of Research and Development, 1959.

[59] A.D.D. Reiling. Courts and artificial intelligence. International Journal for Court Administration, 2020.

[60] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 1958.

[61] G. Sartor and K. Branting. Judicial Applications of Artificial Intelligence. Springer, 1998.

[62] B. Sch¨one, J. Kisker, L. Lange, T. Gruber, S. Sylvester, and R. Osinsky. The reality of virtual reality. Frontiers in Psychology, 2023.

[63] I. Sutskever, D. Gross, and D. Levy. Superintelligence is within reach., 2024. URL https://ssi.inc.

[64] K. Terzidou. The use of artificial intelligence in the judiciary and its compliance with the right to a fair trial. Journal of Judicial Administration, 2022.

[65] R.L. Thompson. Maya: The World As Virtual Reality. Amazon Digital Services LLC - KDP Print US, 2003

[66] A. Turing. Computing machinery and intelligence. Mind, 1950.

[67] Alan Turing. On computable numbers, with an application to the entscheidungsproblem. Proceedings of the London Mathematical Society, 1936.

[68] S. Ulam. John von neumann. Bulletin of the American Mathematical Society., 1958.

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems, 2017.

[70] T. Verbeek and E.J. Bos. Treatise on Man. Cambridge University Press, 2015.

[71] C. Vidal. The Beginning and the End The Meaning of Life in a Cosmological. Springer, 2014.

[72] V. Vinge. The coming technological singularity: How to survive in the post-human era. 21: Interdisciplinary Science and Engineering in the Era of Cyberspace, 1993.

[73] L. Wachowski and L. Wachowski. The matrix. Warner Bros., 1999.

[74] F. Waters and C. Fernyhough. Hallucinations: A systematic review of points of similarity and difference across diagnostic classes. Schizophrenia Bulletin, 2016.

[75] N. Wiener. Cybernetics. The MIT Press, 1961.