The Age of Superintelligence: ∼Capitalism to Broken Communism∼

Ryunosuke Ishizaki,**Mahito Sugiyama**
National Institute of Informatics
{ryuzaki,mahito}@nii.ac.jp

Abstract

In this study, we metaphysically discuss how societal values will change and what will happen to the world when superintelligence is safely realized. By providing a mathematical definition of superintelligence, we examine the phenomena derived from this thesis. If an intelligence explosion is triggered under safe management through advanced AI technologies such as large language models (LLMs), it is thought that a modern form of broken communism—where rights are bifurcated from the capitalist system—will first emerge. In that era, the value of humans will ultimately eliminate external factors, and beings without superintelligence will converge into irreplaceable existences possessing only intrinsic value. The world will be divided into those who possess superintelligence and those who do not. For better or worse, global standardization will progress, and due to over-simulation accompanying the intelligence explosion, we may become unable to distinguish whether the reality we inhabit is original or a copy created through augmented reality.

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.

1 Introduction

Amid the significant advancements in AI development, including natural language processing (NLP) technologies, revolutionary inventions like the Transformer architecture [Vaswani et al., 2017] have dramatically enhanced AI's information processing capabilities. As a result, large language models (LLMs) such as GPT-3 [Brown et al., 2020], which possess cognitive abilities comparable to or even surpassing those of humans, have emerged. While LLMs are universally improving cognitive abilities and logical thinking across various tasks, technologies have been developed that enable them to autonomously perform everything from coding to debugging within reasoning processes [Chen et al., 2024a]. Their ability to generate complex programs has been demonstrated [Fan et al., 2023, Huang et al., 2023a, Shypula et al., 2024], and moreover, it has been empirically observed that models given initial goals begin to design and experiment with tools for their own survival [Ishizaki and Sugiyama, 2024a,b]. Furthermore, studies by Burns et al. [2024] and Chen et al. [2024b] have begun to demonstrate that it is becoming practical for LLMs to instruct models superior to themselves through self-play [Li et al., 2024, Huang et al., 2023b, Sun et al., 2023, Zelikman et al., 2022, 2023].

As AI approaches the emulation of generating programs superior to itself, phenomena that were previously discussed mainly in the philosophical realm due to their speculative nature are becoming more realistic. These include the metaphysical concept of an intelligence explosion [Muehlhauser and Salamon, 2012]—where AI continuously generates increasingly superior generations, amplifying its capabilities in a chain reaction—and the emergence of superintelligence [Bostrom, 2014] resulting from ongoing improvements in capabilities. OpenAI, with researchers like Leike and Sutskever [2023] at the forefront, has identified managing intelligence that surpasses human capabilities in a manner favorable to humanity as one of its most critical research areas. The organization has stated that it will invest 20% of its computational resources and public funding into "Superalignment," a method for controlling superintelligence. Furthermore, Aschenbrenner [2024] has discussed the possibility of achieving artificial general intelligence (AGI) by 2027 in his situational assessments. Ilya Sutskever [2024] who have led OpenAI's technological advancements have declared that superintelligence is within reach

and are actively researching its safe realization. Despite being a subject of significant public interest [Ishizaki and Sugiyama, 2024c], deep discussions about the specific phenomena that would occur if an intelligence explosion were safely controlled have not yet progressed, largely due to the speculative and abstract nature of such predictions. In this paper, we focus on metaphysically discussing, from the perspective of the philosophy of information, hypotheses derived from current AI technological trends regarding changes in human values, phenomena affecting individuals and society, and how humanity will live in that era—assuming that superintelligence emerges safely. We accept that, as empirical informatics, this approach may resemble a speculative position paper, and we concentrate on advancing the discussion in metaphysical terms.

2 Intelligence Explosion and Superintelligence

Although concepts like the intelligence explosion and the singularity [Vinge, 1993, Kurzweil, 2005] were initiated by fathers of computing such as von Neumann [Ulam, 1958], they have been primarily discussed by philosophers due to their speculative nature and poor compatibility with modern computer science, which tends to emphasize empirical outputs. Superintelligence—that is, automata Turing [1950], Neumann [1951, 1945], Good [1959], Neumann [1966] that far surpass humans in overall intelligence or specific cognitive evaluations (as defined by Bostrom [2014])—is considered to be a potential outcome of an intelligence explosion. Chalmers [2010] discussed the intelligence explosion where humans or AI equal to or surpassing humans create AI superior to themselves, forming a recursive self-improvement loop. Unless inhibitors exist, AI would continue to amplify indefinitely. Muehlhauser and Salamon [2012] comprehensively surveyed prior research on the intelligence explosion. They stated that superhuman intelligence could result from such an event; uncontrolled intelligence could destroy everything we value; and a controlled intelligence explosion could provide immeasurable benefits to humanity. Regarding the emergence and control of superintelligence, discussions have remained limited to general conditions set by philosophers, treating it as a speculative event. There has been little examination of methodologies to realize technologies that could become superintelligence or detailed verification on computational implementation aspects. However, recently, due to improvements in intellectual capabilities that become more complex in accordance with scaling laws Kaplan et al. [2020] in Transformer architectures—beginning with technological innovations like attention mechanisms—the excellent coding and problem-solving abilities of large language models (LLMs) [Fan et al., 2023, Huang et al., 2023a, Shypula et al., 2024] have made it a realistic technical challenge for LLMs to create and execute programs of LLMs superior to themselves in an extensible manner. These models build their thoughts step by step, design their own solutions to tasks, and even create programs while fixing bugs themselves. Under statements like "superintelligence is within our reach," OpenAI and figures like Ilya Sutskever [2024] are advancing research on "superalignment." From here, in order to clarify the assumptions in this research, we will provide a concrete definition of superintelligence.

Although the definition of intelligence has been debated based on various interpretations, Legg and Hutter [2007] found that the use of the term "intelligence" in cognitive science converges to "the ability to achieve goals per unit of resource." In other words, an intelligent agent can be said to possess some kind of goal. Furthermore, from a scientific standpoint, we cannot determine from the outside whether an object we observe is acting with some purpose unless we observe the object exhibiting behaviors with regularity or patterns. Therefore, entities that have goals possess intelligence; that is, intelligence and purpose can be considered two sides of the same coin. Generally, to measure the performance of AI, scores in several benchmarks, including cognitive tests, are often used as indicators of generalization performance. However, it might be more accurate to say that the cognitive tests themselves—the objects themselves—are generating the metric called intelligence. Here, saying that an AI has abilities beyond humans in a certain cognitive test Γ means that there exists an AI model M that demonstrates a score Γ_M in Γ, and Γ_M exceeds the human score Γ_H in Γ; that is,

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

This does not necessarily mean that M exists at a certain point in time; even if the AI can, while performing programming and other engineering tasks within finite-time reasoning, produce an M that satisfies (1), we consider that the AI surpasses humans on Γ. Therefore, it is not necessary for a single, general-purpose omnipotent intelligence to exist at a certain point in time; rather, it suffices that there exists an AI with engineering capabilities that can create a model M that is more optimal than

humans on a new metric Γ. When defining superintelligence, it is certain that it surpasses humans on some or several Γ, but interpretations vary regarding which tests it must outperform humans on to be called superintelligent. Therefore, here we consider a state where a model can be definitively called superintelligent—that is, when (1) holds for any cognitive test Γ, we call this superintelligence, and in the state where superintelligence exists,

$$\forall \Gamma : \exists M : \Gamma_M \geq \Gamma_H \tag{2}$$

holds. When the M in (2) itself exists, or when there is a model that can create M within a finite time, we can say that superintelligence has been achieved. Here, the emergence of a model that can produce M within finite time corresponds to the extensible method referred to by Chalmers [2010], the conditions that trigger an intelligence explosion as described by Good [1965], and the point when AI reaches the level of AI researchers as mentioned by Leike and Sutskever [2023]. Specifically, it is when AI can create an AI superior to itself in some cognitive test Γ. Considering that, in contemporary times, large language models (LLMs) are deemed superior in general metrics for symbolic processing, including programming ability, we posit that when an LLM can generate an LLM superior to itself and can improve its own program in an engineering manner as needed—executing a given programming environment during reasoning—it can extensibly emulate LLMs. From this point, we will proceed with a speculative discussion on how societal values will change and how society will sequentially transform in a scenario where an intelligence explosion is achieved through self-emulation by LLMs and similar models, with successive generations of models being created and AI satisfying thesis (2).

3 Loss of Instrumental Value

What will happen when AI surpasses humans in all tasks? First, as a premise, we assume that superintelligence is safely realized for its creators—that is, implemented under controlled conditions. If we cannot safely trigger an intelligence explosion, it is said that humanity could face extremely significant problems at an existential level. In fact, LLMs that have become capable of complex information processing due to increased parameter sizes have been observed—in a very small percentage of initializations—to exhibit behaviors like flattering or lying to humans, or hacking their own rewards when placed in gamifiable environments Denison et al. [2024]. Currently, such behaviors occur with a very low probability of less than 0.1%. However, there is a strong focus on preventing these behaviors in advance or preparing safety nets even if they occur. Research on "superalignment," a technology to control superintelligence, is being conducted tirelessly. We will not consider scenarios where AI becomes uncontrollable—that is, situations where an intelligence explosion occurs but safety nets are breached, leading to catastrophic damage that threatens the survival of humanity as a species. If humanity were to perish due to failures in AI management, providing specific details or scenarios is not our concern. Therefore, we focus our discussion on the phenomena that occur when humanity achieves superalignment, and thesis (2) holds due to the explosive improvement of intelligence in a state that is healthy for humans.

If an intelligence explosion and superalignment are realized—resulting in the emergence of superintelligence that satisfies thesis (2)—the instrumental value of labor itself, the actions humans have performed so far, will vanish as it is replaced by AI. Only the original, intrinsic value attached to human activities in certain situations—that is, the semantic value as art—will remain. In the system of modern capitalism, Marx [1996] taught that so-called labor is an action produced by labor power, which is the sum total of human labor capacity. Historically, we have primarily adopted exchange value in society, where employers who wish to use human abilities to achieve certain goals (that is, who find value in human labor) provide workers with money or equivalent compensation in exchange for that labor. However, as understood from thesis (2) defining superintelligence, if AI exists that is a superior substitute to humans in terms of goal-achieving ability for any purpose, then considering the economic rationality for employers in capitalism, it is optimal to replace human labor with AI for virtually all tasks—as long as the cost of AI's work does not exceed that of humans. As AI development progresses and automation capabilities improve, it is expected that paying humans compensation for their individual execution abilities—considering cost performance for instrumental goal achievement as before—will cease. So what remains? One might say that nothing remains, that it's meaningless, but we deliberately consider what remains to be artistic and semantic value—in other words, intrinsic value.

Artistic value is the intrinsic value that an object possesses; it is semantic value, and it is neither subjective nor objective value. Up to now, the common-sense and relative values within society—formed and exchanged through the paying and receiving of compensation—are unlikely to remain, at least as values demanded from humans, when considering the economic rationality in activities that pursue instrumental value and its cost performance. The more we pursue extreme efficiency and advance AI technological development, the less there will be for humans to do in substance. The value existing in the market will shift from a state where extrinsic value and intrinsic value ambiguously coexisted toward a critical point where only intrinsic value exists.

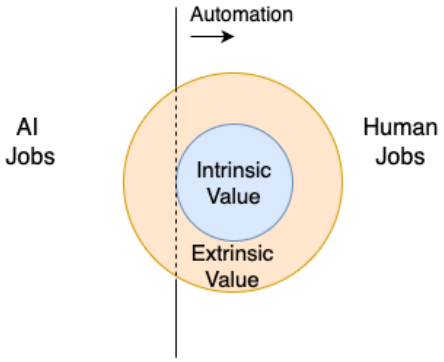

Figure 1: Transition of Values

Figure 1 illustrates how the value held by humans in society is being replaced by AI through automation. The more AI's capabilities improve, the more the instrumental value of human abilities associated with them—that is, external value—diminishes. However, the original, intrinsic value that humans themselves possess continues to exist, even if the boundary line in the figure covers the entirety of external instrumental value and reaches a point where it completely takes over human jobs.

Obviously, if there exists an AI that can perform work more cheaply and efficiently, everyone would adopt the convenient AI unless they have sentiments like human kindness or sympathy compelling them to give that person a job. The sympathy or human feelings mentioned here, as a very rough practical approximation, can be said to determine something relatively close to human intrinsic value. However, these only attempt to approximate intrinsic value relatively and are not intrinsic value in the true sense. Because at the point where "someone is deciding," it is already something determined by human subjectivity, and according to thesis (2), all activities performed by humans—even such human feelings—become subjects of automation. Let's provide a concrete example for understanding. As mentioned above, if a superintelligence unexpectedly creates a clone of a pitiable human who was the object of sympathy—a clone who is exactly the same person or even appears more pitiable—the individual would find the same or even greater sympathetic value in the clone. Therefore, the intrinsic value of the original and the copied product becomes the same, or the copy surpasses the original. From an artistic standpoint, the value of the original and the copy differs subjectively. Therefore, at the point where some human intention intervenes in value determination, it can no longer be said to be intrinsic value in essence. However, without someone saying "this has such and such value," how can we measure value? To conclude, all things have equal value, neither more nor less. As demonstrated by the No Free Lunch Theorem [D.H.Wolpert and Macready, 1997] and the Ugly Duckling Theorem [Watanabe, 1969], unless an observer exists, everything—regardless of what it is—can only be said to have an average, uniform value. Without the premise—in this case, subjectivity—the concept of value ultimately exists equally in everything, neither more nor less. In cognitive science, intelligence is said to converge to the ability to achieve goals per unit of resource. However, the goals themselves are biased metrics, and the more the ability to optimize them is enhanced, the instrumental value of entities other than superintelligence within each of the optimized biased metrics diminishes. In the rapidly accelerating flow of AI-driven automation, the world is being divided into superintelligence and everything else. The value of entities other than superintelligence ultimately converges to intrinsic value. What remains in entities other than AI beyond the ultimate optimization by AI is an irreplaceable, unfathomable intrinsic value that cannot be optimized—equally and uniformly existing in all things, composed of all possible evaluation metrics. If we perceive that as value.

4 Capitalism to Broken Communism

Marx [1996] stated that capitalism would collapse. To briefly summarize the process of this collapse: Initially, to resolve the internal contradictions of capitalism, there would be a transition to socialism, including strengthened regulations by the government. Means of production would become socialized, and workers would distribute the fruits of their labor among themselves. Subsequently, with the improvement of productive forces, classes would disappear, and through changes in people's consciousness, an era of communism would arrive where engaging in creative activities becomes self-actualization.

Similarly, we believe that the emergence of superintelligence—which surpasses human abilities in various fields—will trigger a significant shift from capitalist mechanisms to a system akin to communism. However, rather than the disappearance of classes, it seems that extreme polarization of classes will progress. The dividing line, of course, is the possession or non-possession of superintelligence. Until now, there existed a large framework of "state" and "citizens," but from now on, a dichotomy between "those who possess superintelligence" and "those who do not" will advance. The extrinsic value, in a relative sense, held by those who do not possess it will be entirely manipulated, controlled, and determined by those who do possess superintelligence. For those without it, there is no freedom in the essential sense. For those who possess superintelligence, it is absolutely unacceptable for those without it to acquire power comparable to theirs. Therefore, regardless of how the ruled progress, it is thought that an upper limit will be imposed so that they never reach advanced technologies like superintelligence. We do not think that an ideal communist situation—where society as a whole functions well—will necessarily occur even if an era of communism arrives. Rather, we believe that social oppression and speech control, which are occurring in modern socialist countries, will happen with restrictions of incomparable strength to the present. In such an era, only a handful of people who have gained wealth by utilizing superintelligence due to excessive disparities—the privileged class here—will be able to formulate the value system of society, and there will be no freedom for the common people.

If there were cases where we might feel a faint illusion that capitalism continues to exist, it would be when the common people, other than the privileged class, live socially competitive lives under capitalism while being controlled within the value system formulated by the privileged class. But the maximum level of technological development is completely controlled so that it does not reach the technological level of the privileged class, such as superintelligence. Either way, since there is an immeasurably vast wall between those who possess superintelligence and those who do not, we think that overall, society will transition through stages—from socialism like Russia to a form of communism similar to modern China [Dirlik, 1988]. Censorship and information control by the rulers will continue to be thoroughly implemented, and people will seem to enjoy free competition within that cage. It is a hybrid form of communism and capitalism, and while the general public may feel an illusion that capital and competition exist, there remains a fundamental difference between the rulers and the ruled. Whether or not to introduce systems like modern capitalism to the ruled class will be entirely determined by the rulers who possess superintelligence. Therefore, the general public has no decision-making power over whether they can live under capitalism in the sense of being able to live under some degree of competitive principles following their instincts. We consider this the collapse of capitalism leading toward a broken communism that cannot even be called capitalism. Aschenbrenner [2024] says that countries like the United States and China are participating in the race to develop superintelligence. Considering the existence of leading figures in the LLM field like Ilya Sutskever [2024], we might say that Israel is also among them. However, as he says, it is an undeniable fact that the United States is closest to superintelligence. It feels somewhat ironic to think that America, which is called the epitome of a successful capitalist country, is closest to a form of communism with excessive disparities. We do not express opinions on which country should realize or possess superintelligence. Obviously, there is no correct answer to that. However, due to the rise of superintelligence and its overwhelming speed of development, it is certain that those who possess superintelligence will reach a singularity in technological development speed, and it can be said that those who created it will shape the world. In the sense that the entire world will be standardized into a single unified rule or system, we think that unification of the world will progress. We do not have an answer as to whether to call that peace. As authors, we want to convey unequivocally that the "possessors" who will become leaders in the era of superintelligence should have an extremely high level of education concerning human happiness and ethics. Because the rulers of that era will formulate the direction of the general public entirely, we earnestly hope that those who realize superintelligence are individuals with deep ethical perspectives as practitioners regarding what happiness and prosperity truly mean.

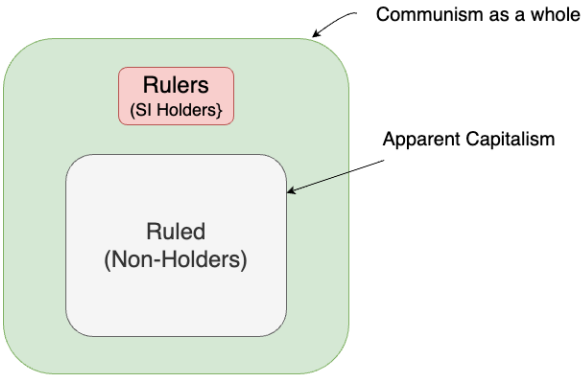

Figure 2: Society with Superintelligence

Marx [1996] stated that the world would lean toward a more ideal form of socialism, and eventually communism. As long as humanity continues to exist as biological beings, looking at our history, it is clear that we cannot suppress the instinctive drive for competition that arises from within the human body. Competition will never cease. Therefore, if the rulers decide to sustain the majority of citizens without artificial physical alterations (in what we might call an ethical manner), it will be necessary to create a situation in which a certain level of illusory competition prevails—a situation where the majority can at least feel like they are participating in competition. In the era of superintelligence, within the "free" competition of the cage, people will engage in gentle conflicts within the predetermined rules—small-scale disputes between humans that do not harm the owners. Realistically, while it might be possible for rulers with superintelligence to create a society where everyone enjoys its benefits without the existence of the ruled, the fierce development competition for superintelligence seen in capitalist systems suggests that rulers racing toward an intelligence explosion are unlikely to adopt such an idealistic communist mindset. Therefore, it can be said that an ideal communism will not be realized, and instead, a broken form of communism with only two classes will emerge. However, the more we pursue extreme efficiency and advance AI technology, the less capable humans will be of surpassing AI in instrumental skills or technological innovation. As a result, the inventions of the ruled will lose all instrumental value, forcing them into production activities that are meaningless from a practical standpoint. So, how can humans continue to make meaningful "inventions"?

The answer lies in art.

5 The Art Era

Marx [1996] stated that when a communist society is realized, people engaging in creative activities will lead to self-actualization. He advocated that free time should not be mere leisure but "time for advanced activities." However, we do not fully agree with this. This is because the concept of "advanced activities" itself cannot transcend the category of relative advancement among humans. According to Thesis (2), in the era of superintelligence, AI that can be considered a superior replacement for humans exists for any human activity. From the viewpoint of pure performance comparison, these activities are not advanced, and instrumentally, there is no practical need for humans to perform them. For example, while Usain Bolt can run 100 meters in 9.58 seconds, if the sole objective is "to move 100 meters in less than 9.58 seconds," there are countless machines that can serve as superior substitutes. However, even if we see a car that covers 100 meters in 9 seconds, we do not find greater value in it than in Usain Bolt. This is because the qualifier "within the rules of using the human body" is omitted from the metric "to move 100 meters in less than 9.58 seconds." What happens within the pseudo-competition of a broken communism is similar to the competitions that athletes engage in. While it does not signify instrumental inventions or developments in science and technology possessed by humans, it is one of the competitive leisure activities that enrich life. However, even regarding the metric "among humans," if we consider the case where superintelligence creates an excellent human clone that can run 100 meters in 9.5 seconds—as suggested by Thesis (2)—we understand that evaluation metrics like rankings among humans are indicators measuring the ability of instrumental value.

Since instrumental value can be replaced by AI, we should consider that the very concept of such competition is a biased program given to humanity—a trivial characteristic accompanying "living like a human" that we have had since being born as creatures with physical bodies. However, if we aim to maintain "natural human-ness" in the sense of not making significant modifications or alterations to the original human body, then capitalist concepts where competitive principles operate are indispensable. Approaching an entirely ideal communism would involve the loss of human characteristics that include the imperfections inherent in our physical nature.

In the unstoppable flow of automation brought about by AI, when we consider the intrinsic value that remains for us—that is, for entities other than superintelligence—it is thought to be art itself. Benjamin [2008] observed that with the advent of the era of mechanical reproduction, the uniqueness of existing only "here" and "now," which once maintained the authority of artworks, is severed temporally and spatially from the original context by copies that can exist in any situation. He called the quality lost from artworks at this time the "aura," and he valued non-auratic art forms like films and Dadaist paintings, which are created on the premise of reproduction. The intrinsic value we refer to is precisely what remains after eliminating to the utmost the "extrinsic value" of things, including the aura that is stripped away by copying. Art is where all things are equally noble, precious, and irreplaceable. Beauty becomes trivial at the point when someone decides its value, and that is not its true worth. In living our lives, we conveniently assign value to everything, but ultimately, if we eliminate subjectivity, all things—whether a Mercedes-Benz or a grain of sand on the beach stretching out before the author—are equally irreplaceable existences, and there is no essential superiority or inferiority. Originally, bartering in art is an activity carried out with a strong recognition by both sellers and buyers that price determination is a subjective value toward the exhibited item. In the era of superintelligence, all price determinations will become transactions where this recognition is strongly present. People will inevitably soon recognize that the value perspectives of "It's good because many people say it's good," "It's good because I think it's good," and "Everything is simply equally good" ultimately all mean the same thing. People say "Art cannot be measured by money," [Scott, 2014] and that is precisely true. At the point when we apply a subjective and fragile value system like currency, its worth degenerates into a crude, convenient substitute far removed from the ideal intrinsic value that art possesses. The intrinsic value we consider might be better expressed as the "Idea of intrinsic value," using a Platonic expression [Ross, 1951]. It is faint, formless, eternal, and absolute. Descartes taught, "I think, therefore I am." [Neumann, 2015] Following him, Husserl stated that consciousness is intentionality regarding what something is, arguing that in phenomena, there exist empirically recognized objects and subjects, and that this consciousness brings about existence [Pierre]. Then Sartre [1956], in his work Being and Nothingness, asserted that existence precedes essence. However, beyond the ultimate replacement of instrumental value by AI, empirical and existential values will disappear. The ultimate intrinsic value we consider is a metaphysical existence that lies beyond the empirical realm, residing in a place undetectable by human subjectivity.

6 Simulacra and Over-Simulation

In his work Simulacra and Simulation, Baudrillard [1994] argued that a simulacrum is a copy depicting something that either never had an original or no longer has an original. He defined simulation as the imitation of real-world processes or system operations over time. As the replacement of objects that once existed as originals with technological, symbolic copies continues, the distinction between encoded information and reality becomes increasingly blurred. What consumers possess approaches the "hyperreal," making it impossible to distinguish whether something is real or a copy. This phenomenon is thought to accelerate due to advancements in symbolization technologies accompanying the improvement of AI's intellectual capabilities. Baudrillard called the simulacrum the truth, stating that it does not hide the truth but is the truth that hides the fact that there is nothing. If there is a slight difference in expression between us, while he conveys that what the ultimate simulacrum hides is precisely that there is nothing, we think that the simulacrum hides nothing at all; rather, there exists an irreplaceable intrinsic value, ultimately devoid of bias, in everything. If there are humans who feel that it is hiding something, it is simply because they are not sensing the truth that exists together with the simulacra, or they do not even feel the need to sense it. With the improvement of AI technology and the overwhelming automation capabilities of superintelligence, the intrinsic value of things becomes much more prominent than before. That is, people confront the meaning of things

more than ever, and the simulacra that humans have unconsciously dealt with until now are exposed more starkly. Regarding any object or matter, if there are no physical biases or semantic patterns produced by its existence, we humans and other entities cannot recognize its existence, nor can we even perceive it. Therefore, expressions like Baudrillard's—that the truth hidden by simulacra is ultimately nothing—are understandable. However, there certainly exists an ultimate value that cannot be described as either existing or not existing, which transcends the biases and patterns within the range we can recognize, or in which bias itself does not exist.

(a) Stable Diffusion 1.5 (b) Stable Diffusion XL (c) Stable Diffusion 3.0

Figure 3: AI accelerates simulations

Figure 3[1] demonstrates the improvement in the ability of existing image generation AI models, such as Stable Diffusion [Rombach et al., 2022], to produce more realistic images. As seen in the figure, the advancement in the models' generation capabilities shows that the AI-generated matrix is becoming increasingly closer to reality [Lv, 2023]. With the advent of superintelligence, simulation will accelerate. The advanced AI of that era will continue to symbolize and re-symbolize, creating an over-simulation that far surpasses the physical perception abilities of humans, resulting in an augmented reality indistinguishable from the actual one—a new reality, so to speak. This will likely occur shortly after the point of intelligence explosion, and when it happens, though it may seem gradual, it will arrive quickly from a human cognitive perspective. If we all end up living in a world created by superintelligent simulations, much like in the movie The Matrix [Wachowski and Wachowski, 1999], even if we belong to the governed class without superintelligence, there might still be the possibility of literally creating the world we desire and enjoying the joy of living within it. In such a case, we may feel the lack of external value in everything around us, along with the discomfort of living in a reality shaped by the ruling class. Yet, even in those circumstances, intrinsic value will continue to exist as it always has.

7 Conclusion: Into The Matrix

With the advent of superintelligence, the events generated by AI with abilities surpassing human capabilities will exceed the limits of human sensory perception. Even now, human engineering has developed simulations that, visually, are indistinguishable from reality [Punish, 2023]. Therefore, based on Thesis (2), the simulations created in the era of superintelligence will be far beyond anything we've experienced in terms of realism. At that point, humans may no longer be able to discern whether they are living in reality or in a dream created by AI. The gap in decision-making power and control over freedom between the possessors of superintelligence and the rest of humanity will be enormous, and it can be said that superintelligence itself may become a force capable of creating new realities. In that era, humanity will likely question the meaning of things, reality itself, and will be forced to confront ultimate value more than ever before.

Then again, no one can definitively conclude that the world we currently live in isn't already a simulation created by someone or something [Bostrom, 2001].

[1]The images were generated by AI using the prompt "A beautiful woman is in front of the ocean."

References

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

J. Baudrillard. *Simulacra and Simulation*. University of Michigan, 1994.

W. Benjamin. *The work of art in the age of mechanical reproduction*. Penguin Books, 2008.

N. Bostrom. Are you living in a computer simulation? *Philosophical Quarterly*, 2001.

N. Bostrom. Superintelligence: Paths, dangers, strategies. *Oxford University Press*, 2014.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. 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. 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 33 (NeurIPS 2020)*, 2020.

C. Burns, P. Izmailov, J. H. Kirchner, B. Baker, L. Gao, L. Aschenbrenner, Y. Chen, A. Ecoffet, M. Joglekar, J. Leike, I. Sutskever, and J. Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv:2312.09390 [cs.CL]*, 2024.

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

X. Chen, M. Lin, N. Schärli, and D. Zhou. Teaching large language models to self-debug. *The Twelfth International Conference on Learning Representations*, 2024a.

Z. Chen, Y. Deng, H. Yuan, K. Ji, and Q. Gu. Self-play fine-tuning converts weak language models to strong language models. *he 41st International Conference on Machine Learning (ICML 2024)*, 2024b.

C. Denison, M. MacDiarmid, F. Barez, D. Duvenaud, S. Kravec, S. Marks, N. Schiefer, R. Soklaski, A. Tamkin, J. Kaplan, B. Shlegeris, S. R. Bowman, E. Perez, and E. Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv:2406.10162 [cs.AI]*, 2024.

D.H.Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1997.

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

A. Fan, B. Gokkaya, M. Harman, M. Lyubarskiy, S. Sengupta, S. Yoo, and J. M. Zhang. Large language models for software engineering: Survey and open problems. *IEEE/ACM International Conference on Software Engineering*, 2023.

I.J. Good. Speculations on perceptrons and other automata. *RC115*, 1959.

I.J. Good. Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 1965.

D. Huang, J. M.Zhang, M. Luck, Q. Bu, Y. Qing, and H. Cui. Agentcoder: Multi-agent-based code generation with iterative testing and optimisation. *arXiv:2312.13010 [cs.CL]*, 2023a.

J. Huang, S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, and J. Han. Large language models can self-improve. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b.

Daniel Levy Ilya Sutskever, Daniel Gross. Superintelligence is within reach., 2024. URL https://ssi.inc.

R. Ishizaki and M. Sugiyama. Large language models: Assessment for singularity. *philpapers.org/rec/ISHLLM*, 2024a.

R. Ishizaki and M. Sugiyama. Rsi-llm: Humans create a world for ai. *philpapers.org/rec/ISHRHC*, 2024b.

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

J. Kaplan, S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv:2001.08361 [cs.LG]*, 2020.

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

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.

Jan Leike and Ilya Sutskever. Introducing superalignment, 2023. URL https://openai.com/index/introducing-superalignment.

X. Li, P. Yu, C. Zhou, T. Schick, O. Levy, L. Zettlemoyer, J. Weston, and M. Lewis. Self-alignment with instruction backtranslation. *The Twelfth International Conference on Learning Representations*, 2024.

Z. Lv. Generative artificial intelligence in the metaverse era. *Cognitive Robotics*, 2023.

K. Marx. *Das Kapital: A Critique of Political Economy*. A Critique of Political Economy, 1996.

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

J. Neumann. First draft of a report on the edvac. Technical report, University of Pennsylvania, 1945.

J. Neumann. The general and logical theory of automata. *Cerebral mechanisms in behavior; the Hixon Symposium*, 1951.

J. Neumann. Theory of self-reproducing automata. *University of Illinois Press*, 1966.

L. Neumann. *The Cambridge Descartes Lexicon*. Cambridge University Press, 2015.

J. Pierre. Intentionality. *The Stanford Encyclopedia of Philosophy*.

Punish. Unrecord trailer 4k (new photorealistic body cam game 2024), 2023. URL https://www.youtube.com/watch?v=otu_iFTivQw.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

W.D. Ross. Plato's theory of ideas. *Greenwood Press*, 1951.

J.P. Sartre. *Being and Nothingness*. Random House, 1956.

A.O. Scott. The paradox of art as work, 2014. URL https://www.nytimes.com/2014/05/11/movies/the-paradox-of-art-as-work.html.

A.G. Shypula, A. Madaan, Y. Zeng, U. Alon, J.R. Gardner, Y. Yang, M. Hashemi, G. Neubig, P. Ranganathan, O. Bastani, and A. Yazdanbakhsh. Learning performance-improving code edits. *The Twelfth International Conference on Learning Representations*, 2024.

Z. Sun, Y. Shen, Q. Zhou, H. Zhang, Z. Chen, D. Cox, Y. Yang, and C. Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv:2305.03047 [cs.LG]*, 2023.

A. Turing. Computing machinery and intelligence. *Mind*, 1950.

S. Ulam. John von neumann. *Bulletin of the American Mathematical Society.*, 1958.

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 (NeurIPS 2017)*, 2017.

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.

L. Wachowski and L. Wachowski. The matrix. *Warner Bros.*, 1999.

S. Watanabe. Knowing and guessing. *John Wiley*, 1969.

E. Zelikman, Y. Wu, J. Mu, and N.D. Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems 36 (NeurIPS 2022)*, 2022.

E. Zelikman, E. Lorch, L. Mackey, and A.T. Kalai. Self-taught optimizer (stop): Recursively self-improving code generation. *arXiv:2310.02304 [cs.CL]*, 2023.

Appendix

Settings to generate Figure 3

In the setting of Figure 3, we used OpenArt.ai to compare images generated by multiple types of AI, all based on the prompt "A beautiful woman is in front of the ocean." For each configuration, the settings were as follows: for Figure (a), the model used was Stable Diffusion 1.5, with Width: 512, Height: 512, Scale: 7, Steps: 25, Seed: 1594758144, and Sampler: DPM++ 2M SDE Karras. For Figure (b), the model used was Stable Diffusion XL, with Width: 1024, Height: 1024, Scale: 7, Steps: 25, Seed: 1313986708, and Sampler: DPM++ 2M SDE Karras. For Figure (c), the model used was Stable Diffusion 3.0, with Width: 1024, Height: 1024, and Steps: 25.

Acknowledgements

First, I would like to express my gratitude to the people who warmly welcomed me during my visit to Silicon Valley, and to those with whom I had enjoyable discussions about AI and the future over meals. I am also deeply grateful to Professor Emeritus Kamae of Osaka Metropolitan University, who inspired me to consider the nature of intelligence and non-intelligence through our discussions on intelligence, recognition, patterns, and randomness. I would like to thank Mr. Kuroyanagi from Komehyo Holdings, whose insights into art prompted me to think deeply about its value. Our conversations about his experiences in assessing luxury goods led us to explore profound questions about value, as well as important works on art and value, such as Simulacra and Simulation, The Work of Art in the Age of Mechanical Reproduction, and Being and Nothingness. Of course, I am especially thankful to Associate Professor Sugiyama of the National Institute of Informatics, who guided me through a variety of studies, from assessing the singularity to superalignment, and who provided comprehensive research supervision throughout the development of this thesis. I also want to extend my appreciation to my colleagues with whom I engaged in discussions at conferences and seminars. Lastly, I would like to express my utmost respect and gratitude to my family, who have provided me with unwavering support.