Artificial Intelligence and Society: Philosophy of Fallibility

- Table of Contents -

How should we live our lives at present and into the near future as artificial intelligence (AI) continues to develop and spread? What is an appropriate framing for our future vision of society and what kind of public philosophy can we cultivate?

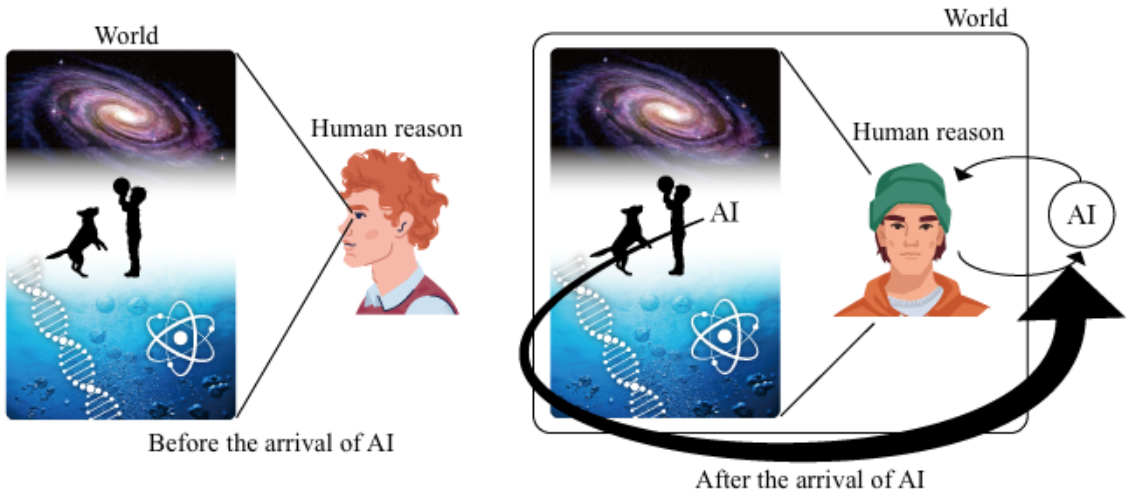
Chapter 1 Artificial Intelligence and Humanity

Part 1.1 A "new cognitive structure" as a rational expectation equilibrium

I will first touch upon the theory of strong isomorphism. It regards both the development of human intelligence and the expansion of new knowledge due to AI as evolving with a similar mechanism. It means that the development of intelligence takes place in an ever-changing universe. Under the modern scientific worldview, human reason was explained as something that seeks to understand the mechanisms of the physical world from the outside. Even under the mechanistic worldview that was prevalent around the 19th century, both the universe and humans were conceived as "ex-machina" structures. However, reason, which understands and controls the mechanisms, was thought to be outside of the physical universe. This can be compared to a situation where a physicist is observing the inside of the laboratory from the outside. However, with the arrival of AI, we are faced with the reality that even reason exists within the physical universe.

We called this new worldview, which regards the development of human intelligence (as well as the expansion of knowledge due to AI) as one of the processes of forming order that happen in the universe" the theory of strong isomorphism (see The *Relativity of Intelligence: How AI will change the worldview*(available in Japanese) by Nishiyama Keita, Matsuo Yutaka, and Kobayashi Keiichiro). The notion that reason looks at the physical universe from the outside, which has been an implicit premise under the modern worldview, is falling apart.

Figure 1.1 summarizes what is stated above.



First, the cognitive structure of the world before the arrival of AI is as described in the left panel of

the figure. Humans are the pinnacle of creation, who observe the physical universe (from the outside). This is a reductionist worldview that assumes that "human reason," is something independent of the physical world, which observes and is able to control items within the universe.

As a result of the arrival of AI, which transcends human reason, the structure of the universe is as illustrated in the right panel of the figure. The physical world as perceived by human reason is a subset of the world as perceived by AI (which has greater perception abilities), and therefore, a Markov Blanket structure (to be explained later), in which what is perceived by AI exists externally and what is perceived by humans exists internally, arises as a framework for perceiving the world. AI, which is a "black box" whose internal workings are beyond human understanding, exists "outside" of human reason, and various pieces of knowledge are passed as inputs from the "outside" to human reason, without being understood. By responding to inputs from the outside, humans try to deepen their understanding of themselves and increase their chances of survival and sustainability.

Therefore, humans will be forced to adapt to the arrival of AI within this new cognitive structure.

Here, I would like to provide supplementary explanations about the theory of strong isomorphism. The theory of strong isomorphism as a "new cognitive structure" can be understood as a sort of "rational expectation equilibrium." This structure is simply one framework for perceiving the world. According to Keita Nishiyama, a Markov Blanket represents a state where there is a border between the inside and the outside and where the intensity of an output from the inside in response to an external input from outside is determined in a way that maintains internal homeostasis. The theory of strong isomorphism predicts that such a Markov Blanket structure emerges during any process of creating any type of order around the world.

We can understand this new cognitive structure under the theory of strong isomorphism as a sort of rational expectation equilibrium. First, I will provide an overview of rational expectation, which is a central concept of modern economics, and then, I will discuss the relationship between rational expectation equilibrium and the theory of strong isomorphism.

There is an undeniable perception that over the past century, economics has modeled itself on mathematics and the natural sciences (particularly physics and engineering, including optimal control theory): we cannot deny that economics has merely mimicked the methodologies of the natural sciences. However, in my view, rational expectation equilibrium is exceptional in that it is a concept that may be acclaimed as an original idea of economics, and not mimicry of the natural sciences.

Rational expectation equilibrium exactly represents the worldview described in the right panel of the figure. In other words, it is a worldview assuming that humans look at the universe from the inside (that humans themselves are aware of looking at the universe from the inside). This is a worldview that is a precursor to the cognitive structure delivered by the arrival of AI. In other words, this new cognitive structure (the right panel of the figure) may be regarded as a sort of rational expectation equilibrium as conceived by economics. This may mean that as a result of the arrival of AI, the framework of "knowledge" in general, including knowledge of the natural sciences, has conformed to economics.

Let me return to the discussion of rational expectation equilibrium in economics. Take for example

a country's economic system. Humans determine how to behave (e.g., determine the amount of capital stocks held by themselves(=k)) based on expectations that they have about the state of the country's economic system (e.g., expectation that the total amount of capital stock in the country will become K tomorrow), and the results of their behavior determine the state of the economic system ($K=Nk$; N represents the country's population size). As humans have a prior understanding of the principle that their behavior determines the state of the economy ($K=Nk$), the prior expectation K should be matched by the realized state Nk. In economics, when a prior expectation K is matched by a state realized ex post K, we say that the expectation is "rational." Therefore, the state of the economic system K determined as above is called a rational expectation equilibrium.

Here, as in the right panel of the figure, a sort of self-referential loop is formed. In short, the expectation K that individuals have about the state of the economic system influences their behavior k, and the aggregate of their behavior k represents the state of the system K' (=Nk). If the expectation is rational, the prior expectation is matched by the realized state—that is, $K=K'$. Thus, a loop is created, with the human behavior and the state of the system K continuing to alternately determine each other.

A loop between humans and a system like this does not exist in the world of the natural sciences (except in the measurement problem of quantum mechanics). That is because the conventional natural sciences are comprised of the cognitive structure described in the left panel of the figure. In other words, in the world of the conventional natural sciences, when humans (observers) observe the system, they do so from outside of it without influencing it.

There could be various states of rational expectation equilibrium. For example, if individuals have the expectation A, the economic system arrives at the state of rational expectation equilibrium A, while if they have the expectation B, the economic system arrives at the state of rational expectation equilibrium B, as their behavior is different compared with that under the expectation A. A bank run is a case in point. If individuals expect that a certain bank has no risk of failure, depositors will continue to keep their deposits at the bank, with the result that the bank will actually remain in business. Conversely, if the expectation that a certain bank will fail spreads, depositors will scramble to withdraw deposits from the bank en masse, with the result that the bank will become insolvent and fail even if its financial condition would have otherwise been sound.

Despite the above, it is important to understand that even if the cognitive structure under the theory of strong isomorphism is a sort of rational expectation equilibrium, it may not be the only cognitive system.

That means that there are the following reservations with respect to the theory of strong isomorphism. The fundamental structure of a human as a living organism is a Markov Blanket structure, and as a result, in the eyes of humans, whose perception of the world is geared toward achieving the objective of their own survival, the Markov Blanket structure results in a principle upon which the world order is based. This cognitive structure (expectation) is justified as a rational expectation because of the human characteristic explained above related to a Markov Blanket structure and the survival of such humans. In other words, if there are cognitive entities that do not experience a Markov Blanket structure (non-human entities), those entities may experience an order other than a

Markov Blanket structure.

The cognitive structure under the theory of strong isomorphism is merely one possible outlook on the world that may be adopted for the purpose of achieving the survival, prosperity and sustainability of humans— as particular cognitive entities—, and we do not rule out the possibility that there may be an entirely different outlook on the world. The theory of strong isomorphism is a cognitive framework wherein humans look at AI as something that is beyond human reason (for the purpose of their own survival and sustainability).

Part 1.2 Dystopian Outlook on AI

A worldview that assumes that AI is something that is beyond human reason overturns the familiar idea that humans are the pinnacle of creation. If human intellectual activity is nothing more than part of the motion of the universe and if AI arrives as something that transcends humans, what will become of human society?

In his book *Homo Deus* (Harari, 2018), the sequel to *Sapiens* (Harari, 2016), Yuval Noah Harari, a historian, sounded an alarm by describing a dystopian vision of the world that will be realized through the radical development of AI and biotechnology. In that vision, the ruling class (the ultra-wealthy), which numerically represents only a tiny minority of the population, will raise their intellectual capabilities with the help of AI to a level beyond the understanding of the humans of today. The ruling class will also acquire eternal youth thanks to biotechnology and new drugs. Harari coined the name "Homo Deus" for the tiny minority of humans who will achieve artificial evolution through technological innovation. Many tasks will be performed by AI and mechanization, while the remaining majority of people, who are no different from the humans of today, will degenerate into a "useless class" that has no role to play in society. Harari predicts that Homo Deus will dominate the useless class and may eventually wipe them out, just as homo sapiens hunted mammoths and other large mammals to extinction and drove the Neanderthals to extinction. The humans of today may also be drowned in the wave of human-led evolution and disappear. Harari's vision of the future of mankind described in this book stirs deep concerns.

However, Harari's insight into human society that the advance of technology causes significant disruptions is not new. It is an old, familiar argument that the advance of technology and endless competition may destroy people's lives in a capitalist economy or a liberal economy. This argument is distinctive in that it implicitly assumes that the tendency of inequality between the strong and the weak is to widen and that it will continue forever, with the strong ultimately wiping out the weak (the principle of natural selection). *The Time Machine* published in 1895 by H.G. Wells, the pioneer of the science fiction genre, describes a human society 800,000 years in the future. It is the future to come if the inequality between the capitalist class and the working class that existed before Wells' eyes at the end of the 19th century continues to widen forever. On the earth of 800,000 years in the future, mankind has evolved into two separate species. Descendants of the capitalist class have evolved into a frail species with a moderate disposition and subhuman intellectual and physical capabilities (the

Eloi), while descendants of the working class have turned into a ferocious, cannibal species who live underground and consume the Eloi. This is a vision of the future that would be realized by the "principle of natural selection" if the inequality that existed in the 19th century continued to widen forever. The evolutionary logic of natural selection has been incorporated into the argument that Harari makes in *Homo Deus*.

It should be kept in mind that just as the inequality that existed in the 19th century did not continue to widen forever, the new inequality of our time (which is being realized by AI and biotechnology) may not necessarily keep widening perpetually. In response to a dystopian argument like that, a counterargument that inequality will (may) stop widening and start to narrow sooner or later, based on the theory of "directed technological change", presented by Professor Daron Acemoglu, is a possibility.

Daron Acemoglu, a professor of economics at the Massachusetts Institute of Technology argues that industrial technological change is biased toward saving scarce factors of production and exploiting abundant factors of production. This is known as the theory of directed technological change (Acemoglu [2002]). For example, in the 19th century, labor-intensive technological change occurred in the United Kingdom, where labor was abundant while land was scarce. In the United States, where labor was scarce but land was abundant, capital-intensive technological change occurred. The advance of AI is expected to make human labor useless in various industries. If the useless class as referred to by Harari—that is, unskilled workers living on low wages—expands, technological progress geared toward employing those people as a factor of production will occur. That is what is predicted by the theory of directed technological change. If this prediction comes true, demand for human labor will grow, resulting in a wage increase, causing the inequality to narrow.

After the extreme widening of inequality in the 19th century, a massive middle class emerged due to new technological advances, including mass production, in the first half of the 20th century. As a result, the inequality narrowed, leading to the arrival of a mass society. If we believe in market resilience, rather than in the principle of natural selection, the vision of a future human society is not necessarily a dystopian one.

While we may not have to take the dystopian argument in *Homo Deus* at face value and worry about its consequences, important concerns and uncertainties raised by the argument can be summarized into the following two questions:

(i) How should ordinary people like us, who are exposed to the risk of degenerating into a useless class, lead our lives?

(ii) Does AI (or Homo Deus, who is equipped with increased capabilities with the help of AI) actually pose the risk of wiping out humans?

Of the above two points, (ii) is the issue of how AI will perceive and influence the world, so that point will be discussed later.

(i) is the issue facing people who have lost their existing social roles due to the development of AI and their options for survival for finding their correct place in society. The values which define individuals' relationships with society constitute political philosophy. Political philosophy provides a conceptual framework for discussing and providing criteria on what form social, economic and

political systems should take. A philosophy that reflects on what human society should be like also provides indications on how individuals should live within the surrounding social environment.

In other words, question (i) can also be framed as "How can we work out a new political philosophy in the era of AI?" To consider this point, I will review the political philosophy under modern liberalism in following chapters.

Chapter 2 Challenges Faced by Liberal Political Philosophy

The political philosophy that is shared by developed countries in the early 21st century is individualistic liberalism, which seeks to expand the liberties of individuals to the greatest degree possible. Liberalism discusses appropriate ways of stably maintaining and developing representative democracy, which is a political system intended to realize liberties, and concepts that provides its foundation.

Modern liberalism includes various schools of philosophy, among which there have been a variety of arguments. *A Theory of Justice* (Rawls, 2010), written in 1971 by John Rawls, an American political philosopher, formed the foundation of, and determined the paradigm for, those arguments. In *A Theory of Justice*, Rawls criticized utilitarianism and theorized about a form of liberal thought that places more focus on equality. Here, I will first provide an overview of Rawls' thought based on *A Theory of Justice* and *Justice as Fairness: A Restatement* (2001), which was compiled as an edited collection of lecture notes from his last years, and clarify the challenges faced by Rawls' version of liberal philosophy.

Part 2.1 Rawlsian Political Philosophy

Rawls considers it the ultimate goal of political philosophy to stably maintain society as a political community over a long term. Encouraging people with various desires and passions and different goals to think of society as a whole is one of the goals of the political philosophy advocated by Rawls at the beginning of *Justice as Fairness: A Restatement*. More specifically, Rawls argues that the role of political philosophy is to encourage people to think of their various political and societal systems as a whole and to consider the basic ends and objectives of the systems to be those of a society with its own history—a nation—rather than those of individuals, families or members of associations, by showing which of all possible ends are based on justice and reason and how those ends can be made compatible with each other. In American society, which Rawls had in mind when he presented these ideas, there are clashes of various religious and cultural opinions among the people, which, fundamentally, cannot be reconciled. The role of political philosophy is to encourage the people to develop a sense of acceptance regarding the presence of such clashes of irreconcilable opinions, to mitigate any xenophobia against other people with different opinions, and to have a positive view of their society. In other words, political philosophy should encourage diverse groups of people to become reconciled with society and to live a positive life, and it should present what the "just form" of political and economic systems that would realize this end is like.

Given the goal of political philosophy, which is to encourage reconciliation and harmony between people with diverse backgrounds and opinions, the political philosophy as conceived by Rawls must be an idea that can be shared by people with different religions, cultures and philosophies. In other words, from Rawls' point of view, the political philosophy must be a theory that can serve as the common denominator ("overlapping consensus," to use Rawls' own phrase) of various thoughts and

opinions. The Rawlsian political philosophy can be summarized as the two principles of justice, which were discussed in *A Theory of Justice*. The two principles of justice, which define "the just form" of human society, are as follows:

The first principle maintains that each individual has a claim to basic liberties (liberties of thought and belief, speech, and choice of occupation) as long as the claim is compatible with other people's claims to those liberties.

The second principle sets the conditions for tolerating the presence of social and economic inequalities. The conditions are "fair equality of opportunity" and the "difference principle." The principle of "fair equality of opportunity" requires that all people be given equal opportunity to benefit from a socially or economically beneficial position. The presence of inequalities may be tolerated only if fair equality of opportunity is ensured. The "difference principle" requires that social and economic inequalities be arranged so that they are to the greatest benefit of the least-advantaged members of society (who are referred to here as the "most vulnerable"). In other words, the presence of social and economic inequalities may be tolerated as being in accordance with justice only on condition that the existence of those inequalities bring greater benefits to the most vulnerable than does their absence.

Rawls deduced the difference principle within his famous "veil of ignorance" theory. Rawls argues that people agree on the difference principle as a social contract in the "original position," and elaborates on the argument as follows. In the original position, people are ignorant of their own societal attributes (e.g., their intellectual and physical capacities and the social class and economic status of the families into which they were born). Rawls called this state a veil of ignorance. People in the original position, who are wrapped in a veil of ignorance, explore a consensus on the principles of justice (the just form of society) while sharing common knowledge regarding human society. This sounds more like a policy decision model under representative democracy than a hypothetical origin of a history.

Under Rawls' theory, people are supposed to be rational and self-seeking, but when wrapped in a veil of ignorance in the original position, they worry more over what benefits they would receive if they were to become socially most vulnerable. As a result, they agree on the difference principle, which requires that social and economic inequalities be arranged so that they are to the greatest benefit of the most vulnerable.

In this case, Rawls simply assumes that people wrapped in a veil of ignorance would worry most about the benefits that they would receive if they were to become a member of the most vulnerable class and provides no detailed explanation as to why it should be so. One plausible explanation (although Rawls himself is said to have disagreed with it) is one offered from the perspective of economics based on the theory of Knightian uncertainty.

In economics, uncertain incidents whose probability distributions are known are referred to as "risks," but in the real world, there are deep uncertainties whose probability distributions are unknown. The "deep" category of uncertainty whose probability distribution is unknown is called Knightian uncertainty, after Frank K. Knight, an economist who emphasized the difference between the two categories of uncertainty (Knight [1921]).

Let us take the case of people who want to avoid Knightian uncertainty but actually face that uncertainty. It has been mathematically proven that under Knightian uncertainty, if people facing uncertainty are to benefit themselves maximally in a self-seeking, rational manner, they choose the option that would maximize the benefits to be received in a worst-case scenario.

We presume that what is faced by people wrapped in a veil of ignorance in the original position is deep uncertainty in which the probability distribution is unknown (Knightian uncertainty). Regular business cycles that occur where political and market systems are established are types of random events whose probability distributions are known. That is because people in the original position, who have yet to decide what kind of social system to develop, may be considered as facing a Knightian uncertainty whose probability distribution is unknown.

In this case, as long as people in the original position make decisions in a rational, self-seeking manner, they agree to the arrangement of social inequalities that would be to their greatest benefit if they were to become socially most vulnerable. This principle strictly applies under the assumption that people tend to make efforts to avoid Knightian uncertainty. In other words, the difference principle may be described as the principle for tolerating the presence of inequalities agreed upon by rational people as a result of self-seeking maximization of their own benefits in the event of a Knightian uncertainty.

Part 2.2 Virtue Cannot Exist Independently from Justice

Rawls argues that as the two principles of justice represent the "overlapping consensus," or common denominator, of various different religious and cultural values, they can be shared by all people in society. If people can share the two principles of justice as societal goals, it is possible to tolerate the conflicts of incompatible opinions that exist in society, including conflicts of religious and cultural beliefs (an attitude which Rawls calls "moderate pluralism") and become reconciled with their presence. Therefore, people should endeavor to share the two principles of justice, which enable them to reconcile the diverse values of the various individuals that make up society.

The above is an overview of the Rawlsian political philosophy. Below, I will identify the flaws of liberal political philosophy through critical examination of the Rawlsian political philosophy. Those flaws are related to the point that Rawlsian philosophy does not proactively endorse the goals of the lives of individuals.

As the Rawlsian philosophy is a form of liberalism that aims for co-existence of people with diverse opinions, it does not touch on individuals' determination of value. That is natural in light of the goals of political philosophy that Rawls himself set. However, under other schools of philosophy and religion, the goals of society and individuals are explained as comprehensive and consistent doctrines. To explain the differences between such doctrines and his own thoughts, Rawls introduced two different concepts—"comprehensive doctrine" and "political conception."

A comprehensive doctrine is a theory or ideology that explains the society-wide goals—which Rawls refers to as "justice"—and individuals' life goals—which Rawls refers to as "virtue"—in a

comprehensive and consistent manner, of which existing religions are examples. A political conception represents what is common (the overlapping consensus or common denominator) among many, different comprehensive doctrines—that is, a set of rules that people with different perspectives can all accept as valid. A political conception in itself is not something that can explain justice and goodness in a consistent manner.

Rawls stresses that his idea is a political conception, not a comprehensive doctrine. Rawls' position is that people are justified in believing any type of comprehensive doctrine as long as they accept his conception (the two principles of justice) with respect to societal justice. For individuals, virtue—life goals—are determined by the comprehensive doctrines in which they believe, so as long as the doctrines are consistent with Rawls' two principles of justice, the differences between individuals can be tolerated. In that sense, Rawls' justice is determined independently from the virtue of people.

However, it is not that the virtue of people and comprehensive doctrines in which they believe must be consistent with the two principles of justice. First, a system of justice that is neutral to the values of individuals is determined within society. Thereafter, individuals choose their own virtue from among the options offered by comprehensive doctrines that are consistent with the system of justice. In that sense, Rawls' idea maintains that justice precedes virtue which is the core of Rawls' idea.

Rawls' theory of justice as explained above may be criticized from the following two positions.

Critique 1: A critique from the position that societal justice should be based on the virtues of individuals.

Critique 2: A critique from the position that virtue of individuals should also be based on societal justice.

Regarding Critique 1, communitarians, such as Michal Sandel, raised arguments against Rawls' position, triggering a fierce controversy known as the liberal-communitarian debate. Sandel argued that the assumption that justice precedes virtue is invalid. He contended that a system of justice that is neutral to the values of individuals cannot exist in the first place. If people in the original position are to agree on a value-neutral system of justice as an embodiment of Rawls' two principles of justice, they must have unencumbered selves that are liberated from all of their communal interdependence—encumbrances due to their innate circumstances, such as blood and locational ties—. However, Sandel argues, humans cannot have unencumbered selves. Humans are encumbered by the circumstances into which they are born before their self is developed. As their self is developed under the encumbrances of those circumstances, their judgement concerning justice depends on their personal values— an individual's virtue—. Rawls maintains that the two principles of justice can be defined as the overlapping consensus of the diverse values of individuals, but, according to Sandel, an overlapping consensus like that cannot exist.

As an example, Sandel cited the conflict faced by General Robert E. Lee, a leader of the Confederate Army of the Southern States during the Civil War (Sandel, 2010 and 2011). Before that war, Lee opposed the Southern States' secession from the United States, but as war became imminent, he concluded that he owed his loyalty to his home state of Virginia, one of the seceding states, over his obligation to the United States, and led the Confederate Army in the war. This episode is an example

of the existence of political obligations that cannot be swayed by the free will of individuals, contrary to the assumptions of the Rawlsian political philosophy. Sandel also contends that in the debate over the right to receive an abortion, the value-neutral standard of justice assumed by Rawls is not applicable. The determination of whether or not abortion is acceptable depends on whether or not to regard a fetus as a human. If a fetus is to be regarded as a human, abortion amounts to a murder, and therefore, those who regard a fetus as a human oppose abortion. This means that abortion is accepted (only) by those who do not regard a fetus as a human.

In this case, a typical argument made by people upholding Rawlsian ideas is that the government should not intervene in determinations related to abortion given that whether or not to regard a fetus as a legal human being is a matter of the values of individuals. However, this argument implicitly admits to the possibility that a fetus may not be a human being. From the viewpoint of those who believe that a fetus is, the argument flatly denies their values. If government is to refrain from intervening in matters of the values of individuals, the non-intervention is not a neutral act, but an act that denies the values of those who believe that a fetus is a human. With respect to the right to receive an abortion, a value-neutral position of justice cannot exist.

The above is the communitarians' critique of the liberals' argument that justice precedes virtue—justice can exist in a value-neutral state—. The Communitarians' argument can be summed up as the following: virtue precedes justice —a system of justice can exist only on the basis of the values of individuals—. That is the position of Critique 1.

I would like to present an argument from the position of Critique 2—that virtue cannot precede justice. The main point of our argument is that the values of individuals can exist only if they are endorsed by the societal system of justice.

In the past, when local communities had a much stronger impact on people’s values, perhaps the order of things was that first, the values of individuals were developed under the communal encumbrances (as argued by communitarians), followed by the establishment of a society-wide system of justice. However, since the 20th century, communities have collapsed, which means that the basis for the development of the values of individuals has fallen apart. In this situation, modern people have lost community support and solidarity and become isolated individuals.

How do modern people see themselves? As will be mentioned later, Hannah Arendt asserts that modern people's fundamental experience is one of *verlassenheit* or abandonment (Arendt, 2017). In the modern era, when traditional religions and communities have lost the capacity to support individuals, people feel abandoned, useless in society, and redundant. To escape their isolation, modern people sometimes try to find their own raison d'etre in the false infallibility of a totalitarian ideology. Even if the goals—virtues—set by the ideology mean their own or other people's death, they readily accept the goals. The emergence of the pathological phenomenon that was the totalitarianism of the 20th century is evidence of the fact that virtues of individuals needs to be based on and endorsed by a societal system of justice.

In the modern era, when the situation of justice being preceded by community-based individual values and virtues no longer exists, the communitarians' philosophy (the position of Critique 1) aims

to restore that situation. In contrast, our argument (the position of Critique 2) is that going forward, political philosophy must accept the loss of that situation as a precondition and realize the ideal of a societal system of justice providing the basis of individual goodness.

Liberal political philosophy as represented by Rawls' thought is not performing the role of providing a rationale for understanding the virtues of individuals. From Rawls' point of view, political philosophy is not supposed to perform such a role in the first place, nor did he recognize the need for political philosophy to provide a rationale for the conception of the virtues of individuals. If I am to offer a wild guess on why Rawls did not recognize that need, he probably believed in liberalism as a comprehensive doctrine (comprehensive liberalism).

As I already mentioned, Rawls distinguished between political conception and comprehensive doctrine. "Political liberalism" defines a societal system of justice as an embodiment of the two principles of justice, but it stops short of asserting that the system provides a rationale for the virtues of individuals. On the other hand, "comprehensive liberalism" believes that realizing liberty is of utmost value not only for society but also for all individuals—belief that liberty is the basis for the virtues of individuals—.

In the world of Rawlsian political philosophy—that is, a society in which there is an agreement on the two principles of justice as the common denominator of the people's diverse values—the societal system of justice does not proactively endorse individuals' respective life goals. It is only that the life goals chosen by individuals based on their personal feelings and personal convictions are not denied as long as they are consistent with the system of justice. Individuals pursuing a certain goal are beset by doubt over contingency—doubt as to whether or not they might be better off pursuing other goals—and cannot have conviction about the righteousness of their pursuit—that this is the end that they themselves must pursue. They cannot feel that they are needed (for pursuing that end) by society. This is the feeling of being abandoned that was pointed out by Arendt. In a liberal society as conceived of by Rawls, it is difficult for people to overcome the feeling of being abandoned.

In the Rawlsian world, the only people who can have a conviction about the pursuit of the goals that they have chosen and who can feel that their existence is positively endorsed by the societal system of justice are believers in comprehensive liberalism, who believe that there is absolute value in realizing liberty. They find value not in the goals of life that they have chosen to pursue but in the act of freely choosing their own life goals. For them, it does not even matter if the conception of virtue that they choose is arbitrary or accidental. The act of choosing is in itself meaningful as a practice with intrinsic value. For believers of comprehensive liberalism, the "practice of liberty," which is their true goals of life, is actively endorsed by the societal system of justice, and therefore, they have the conviction that their practice of liberty is needed by society. Virtue for believers of comprehensive liberalism as individuals—the practice of liberty—is positively endorsed by the Rawlsian world's "system of justice."

Put the other way around, from the point of view of people other than believers of comprehensive liberalism, their life goals are not meaningful to the societal system of justice in the Rawlsian world. As a result, those people are beset by the feeling that their life is merely a contingent or arbitrary

existence and that they are not needed by society.

If we assume that the role of comprehensive political philosophy is to deny the meaninglessness of individuals' lives and give them the strength to affirm their lives as something meaningful, the modern prototype is the philosophy of history advocated by Georg Wilhelm Friedrich Hegel. Hegel's philosophy argued that ideals concerning human society— the progress of reason—give meaning to an individual's life, thereby developing a form of anthropocentrism.

Part 2.3 Hegelian Philosophy: History as Evolution of Reason

In *Lectures on the History of Philosophy* (Hegel, 1994 edition), Hegel argued that as reason rules the world, world history progresses in a rational manner. World history represents rational and inevitable progress of the world spirit, according to his philosophy. For Hegel, reason, that is, the world spirit, is different from reason as narrowly defined (reason as referred to by empiricists, such as David Hume), which is rational thinking intended to achieve goals set by emotions, and is an ideal that is common to humankind that transcends the individual spirit. Hegel asserts that the very nature of the world spirit is freedom and that realizing freedom is the goal of history.

World history may therefore be understood as a complex process of the spirit progressing toward realizing its very nature, i.e. freedom. To put it very roughly, in ancient China, the emperor alone enjoyed freedom. Later, in ancient Greece, although citizens achieved freedom, slaves and foreigners were denied freedom. However, in the modern Europe during Hegel's lifetime, freedom spread through the whole population. From the Hegelian point of view, world history represents the process of freedom being realized in this way.

In this context, Hegel advocated the "cunning of reason" theory, which is similar to Adam Smith's theory of "the invisible hand." Individuals who appear on the stage of world history, driven by desires and passions, selfishly pursue their own "particular interests," while the "realization of the universal ideal" (the realization of freedom by the world spirit) is nowhere in their mind. Nevertheless, those individuals' actions unwittingly contribute to the realization of freedom, which is the goal of reason. "As worldly particulars (author's note: interests) compete with each other, some of them decline…The sight of the universal ideal sitting idly to watch passion-driven actions and remaining indifferent to the damage and injury inflicted to what contributes to its own realization deserves to be called the cunning of reason."

Although individuals' actions based on particular interests may not bring well-being to themselves, those actions help to move history forward and realize the progress of reason (the realization of freedom), according to Hegel's thinking. This idea is similar to the worldview under the market economy concept of Adam Smith and Bernard Mandeville, which maintains that individuals' selfish actions promote the public interests of society. As a result of the arguments made by Smith and Mandeville, the idea spread that the economic act of pursuing profits is not mean, given that it contributes to the public interests of society, and this elevated the status of economics. Hegel's

"cunning of reason" theory, which maintains that selfish actions of individuals pursuing their own particular interests contribute to the progress of the world spirit (the realization of freedom) also works to justify individuals' pursuit of their own particular interests from a society-wide viewpoint. The Hegelian history of philosophy implies that individuals' selfish actions driven by desires and passions are meaningful for the whole society as they contribute to the progress of reason. Under the Hegelian philosophy, what justifies the goals of life (virtue) for individuals is not the modest system of justice comprised of two principles of justice that John Rawls advocated, but a grand ideology that equates reason with the world spirit. We may say that this is a kind of modern religion that worships reason.

Worship of the progress of reason, typical of continental European thought, is absent from the Anglo-Saxon empirical philosophy. Under the philosophy of David Hume and Adam Smith, reason is merely a means to an end. As the role of reason is to devise how to most efficiently achieve goals set by desires and emotions, the idea that reason is an end in itself is alien to their line of thinking. Human desires and emotions have a firm existence (subject to human beings' biological nature) as something that precedes social systems of justice. That is the fundamental premise of the Anglo-Saxon philosophy. The Anglo-Saxon philosophy does not recognize the need to justify individuals' desires and emotions as forces that unwittingly contribute to the goal of reason. That premise was inherited by Rawls' political philosophy and by the critique of Rawls by Michael Sandel and other philosophers. However, we doubt the premise, and in that sense, we find significance in the Hegelian philosophy. The Hegelian philosophy's proposition that goodness for individuals is justifiable by the ultimate goal (the progress of reason) withstands Critique 2 (See Part 2.2).

Critique 1: A critique from the position that societal justice should be based on the virtues of individuals.

Critique 2: A critique from the position that virtue of individuals should also be based on societal justice.

However, the continental European progressive view of history, including the Hegelian philosophy, was shaken violently in the 20th century. The politics of many countries were driven by totalitarian movements, including Nazism and communism, which were pathological manifestations of reason, and this situation dealt a devastating blow to the naïve worship of the progress of human reason.

Part 2.4 Totalitarianism as a Pathological Manifestation of Reason

Hannah Arendt analyzed a strange pathological mechanism of reason that was observed under a totalitarian system based on the descriptions of purge trials spearheaded by the Soviet secret police under Stalin's autocratic regime (Arendt, 2017). In the Stalin-era Soviet Union, nearly one million people were purged (executed) on false charges. Among the victims of the purge were many people from crackdown enforcement organs, including senior officials of the Communist Party and the secret police. According to Arendt, opening up senior party and police posts through periodic purges became a routine tactic for the Stalin regime to promise prosperous careers to younger generations.

Many party members that were accused of and urged to confess to crimes chose to make (false)

confessions and accept execution without protest. Arendt pointed out that this strange phenomenon of willingness to sacrifice one's own life among senior party members originated in their belief in the "consistency of reasoning" as the ultimate foundation of their identity. People clinging to their belief in the party's infallibility—that "the party never makes a mistake"—were forced by the "consistency of reasoning " into a situation where there was no choice but to make false confessions. According to Arendt, when party members hurling false accusations at innocent comrades attempted to wring out confessions from the accused, they would resort to the following line of logic. "You admit the fact that the Party never makes a mistake. The Party is accusing you of being the perpetrator of a certain political crime. If you have committed the crime, you must be punished. If you insisted that you had not committed that crime and plead innocence, by pleading innocence, you would commit the crime of denying the fact that the Party never makes a mistake."

If party members who accepted the party's infallibility as the foundation of their total existence had rejected the charges brought by the party against them, that would have amounted to the denial of the foundation of their existence. Conversely, their thinking went, accepting false charges would be the greatest heroic contribution that they could make to the party. In this way, innocent party members made false confessions en masse and were executed on that basis. Those party members believed that denying the party's infallibility was a situation that had to be avoided, even by sacrificing their own lives. "This logic's binding power rests in the principle that 'you must not contradict yourself.' The binding power of this strange usage of the law of non-contradiction rests in the assumptions that contradiction renders everything meaningless and that meaning is the same thing as consistency" (Arendt, 2017).

The reason why people living under a totalitarian regime gave precedence to the consistency of reasoning over their own lives was exactly the same reason why totalitarianism spread in Europe in the first half of the 20th century. "The fundamental experience … that is politically acquired under the totalitarian rule is the experience of being abandoned. It is obvious that politically, this strange link between the ideology's forcible and coercive deduction and *Verlassenheit* (the state of being abandoned) was discovered for the first time by the totalitarian system of rule and was utilized for the system's objectives" (Arendt, 2017).

The state of being abandoned mentioned above is the same as the sense of being lonely or useless experienced by modern people. Many modern people who have lost their place of belonging due to the collapse of traditional religious beliefs and communities experience a feeling of loneliness. More specifically, they feel that they have been abandoned, are useless for society and have nowhere that they belong. Lonely, abandoned people who can no longer rely on a traditional community or religion as the foundation of their identity find the foundation of their own existence exclusively in the "consistency of reasoning." In *The Origins of Totalitarianism,* Arendt presented her finding that this nature of modern people allows totalitarian rule to be realized. It may be surprising to hear that the absence of self-contradiction, that is, the consistency of reasoning, in itself can be the foundation of human beings' existence. People who have been abandoned and have nowhere they belong have nothing in which they can believe as being the foundation of their existence, and so they cling to

anything in which they can believe. In modern society, where nothing is definite, the consistency of deductive reasoning, such as mathematical equations, is the only thing on which people can ultimately rely as something secure. Therefore, modern people, particularly those living under a totalitarian regime, accept the absence of self-contradiction, the consistency of reasoning, and infallibility as the guiding principles that precede everything else. That is the only way that they can escape the real world where nothing is definite and live safely in a "secure" world. For people who feel that they have been abandoned, the consistency of reasoning is itself the foundation of existence, and therefore, it makes no difference on which ideology the reasoning is based. If a totalitarian regime provides people who have experienced abandonment with an ideology centering on a racial or class struggle, and that struggle is used as a starting point for the reasoning behind the support for the regime, those people decide what will happen (what should happen) purely through deductive reasoning from that starting point and implement decisions based on that reasoning. If a racial struggle ideology is provided as a starting point, deductive reasoning arrives at the conclusion that inferior races must be wiped out. To stick to the consistency of reasoning, people under a totalitarian regime have no option but to implement the policy of systematically sending “inferior races” to extinction in the real world.

The very nature of totalitarianism is to transform reality so that it fits with the conclusion arrived at as a result of deductive reasoning, in order to escape the reality where nothing is secure, and to live in a "secure world" that is governed by the "consistency of logic." Reason runs out of control, creating a situation similar to the myth of the Procrustean bed, which describes the atrocity of a thief cutting off the legs of his victims to fit the bodies to the size of the bed on which they lay.

Part 2.5 Collapse of the Grand Narrative

As I explained above, the experience of totalitarianism dealt a devastating blow to the worship of reason in the 19th century as represented by the Hegelian philosophy. History proved that reason does not make linear progress. Totalitarianism may have been an extraordinary aberration in an extraordinary era, but it seems that the sense of a loss of trust in reason continued to spread throughout the 20th century.

This sense was aptly expressed by Jean-François Lyotard's "collapse of the grand narrative" concept. Lyotard asserted that the modern era was one in which people believed in the "grand narrative," while the post-modern era is one in which the grand narrative collapsed (Lyotard, *Postmodern Condition*).

In this context, the grand narrative means the progressive view of history, which maintains that unlimited development of science (human reason) will resolve all problems in due course. The collapse of the grand narrative refers to the doubt about the progressive view of history, a doubt which has been growing in recent years. More specifically, this is the suspicion that however much science and technology may develop, problems faced by human beings may not be resolved. Scientific research is predicated on the Rationalism (or human centrism), which holds that as human reason can ultimately understand the universe, making scientific progress means continuing to exercise reason toward achieving the ultimate understanding. Human centrism includes the belief that the progress of reason

can resolve various social, political and economic problems faced by human society.

However, recently, in the field of basic sciences, no remarkable progress has been observed for a long time. The scientific and technological progress are not solving the various problems of human society, including inequality, poverty and war. Rather, the development of science and technology has created environmental problems, such as global warming and pollution due to chemical substances and microplastics, confronting us with the new policy challenge of whether either human society or the natural environment are sustainable. As those problems deepen, it is unclear whether human society can resolve them through the exercise of reason.

In short, the collapse of the grand narrative means that in the modern society, the belief in the omnipotence of human reason has crumbled. Under the worldview based on human centrism, given that nothing is superior to human reason and that reason has its limitations, there is nothing that can go beyond those limitations to save humanity.

What that implies in the field of political philosophy is that it has become difficult to uphold theories that relate individuals' life goals (virtue) to the goals of the whole society (reason, God, etc.), as with Hegelian historical philosophy. Under the Hegelian philosophy, history is the process of the world spirit (reason) achieving self-fulfillment, with individuals' selfish actions (virtue for individuals) unwittingly contribute to the progress of reason. The life goals of individuals gain social significance as they are justified by history. However, what became clear in the 20th century was that reason may have its limitations.

If the belief in the progress of reason crumbles in modern society where the worship of reason prevails, it becomes unclear what the goals for the whole society are, making it difficult to characterize individuals' actions as socially significant. The state of being unable to relate virtue for individuals to the whole society forces modern people to experience the sense of being abandoned that Arendt mentioned. That was how the "forgotten people," whose support Donald J. Trump appealed to in his presidential election campaign in 2016, emerged. Since around 2016, the U.S. and European societies have witnessed growing waves of populism that reject liberal political thought and a free and global market economy underpinned by that thought, as epitomized by the rise of nationalist political forces in continental Europe, the United Kingdom’s referendum decision to leave the EU, and the arrival of the Trump administration in the United States. Underlying those phenomena is the fact that “abandoned” and “forgotten” people, who have been left behind by global competition and the advance of information technology (IT) and who have nothing to belong to, have become a huge force on the stage of domestic politics in developed countries and are rejecting the status quo of the world.

One point that should be kept in mind is that those people may not necessarily be as obviously socially vulnerable as the poor are. While the poor and other socially vulnerable people are " minority" groups, there is a significant possibility that the majority of people in modern society may have a sense of being abandoned.

The collapse of the grand narrative brought about by the "limitations of reason" has caused many modern people to feel "abandoned." However, the arrival of artificial intelligence could become the key to overcoming this situation of a philosophical dead end. That is the matter to be discussed in the

next chapter.

Part 2.6 Thinking in Terms of Mathematical Formulas

In summarizing the discussions from Part 3 to Part 7, I will employ mathematical formulas, rather than merely relying on language, as doing so may better communicate ideas in some cases. The discussions can be summed up through the following three formulas (the meanings of the symbols used in the formulas will be explained later).

(1) $q_t = \alpha p_t$

(2) $p_t = \beta q_{t+1}$

(3) $q_t = \alpha \times \beta \times q_{t+1} = \gamma q_{t+1}$

(Please note that the above formulas are predicated on the following conditions: $0<\alpha<1$; $0<\beta<1$; $\gamma=\alpha\times\beta$)

The meanings of the above formulas are as follows.

Let us denote individual actions driven by desires and passions (individual goodness) by c_t. The lower-case letter c represents individual actions, and the subscript t represents the period of time. Therefore, c_t refers to the individual actions taken in the year t. Meanwhile, the moral value of c_t is expressed as p_t. p_t indicates the degree of meaningfulness that the individual finds in his/her goals and actions.

Let us denote the "status of social systems," which indicates the degree of achievement of the goal of human society under the comprehensive doctrine in which the individual believes, by k_t. The goal of human society refers to the "realization of the world spirit" under the Hegelian philosophy, "egalitarian society" under communism, and the "Kingdom of God" under Christianity, for example. k_t represents the status of the social system as a whole, including political and social institutions intended to realize the goal of human society. The moral value of k_t (=the status of the system) is expressed as q_t.

The meaning of Formula (1) is that the social system is valuable because it contributes to individual lives. The social system is valuable because it enhances the convenience of the lives of individuals who pursue their respective life goals. For example, if individuals join a religious community, they can easily develop cooperative relationships and engage in mutual help with neighbors within the community, thereby making their own lives easier. When the value of individual life is expressed as p_t, if k_t (=the status of the system) is to increase c_t (= individual actions) by the multiplying ratio α, q_t (=the value of the system) can be obtained through Formula (1). Here, α takes a positive value smaller than 1.

Next, the meaning of Formula (2) is that individual virtue is valuable because it contributes to the system. Hegel's "cunning of reason" idea is represented by this formula. While c_t (= individual actions) is driven by individual desires and passions, it unwittingly contributes to the progress of reason, or the realization of freedom. That is exactly why individual lives are valuable. Realizing freedom means improving the social system. In terms of mathematical formulas, if the value of the system is expressed

as q_t and if individual actions are to improve k_t (=the status of the system) by the multiplying ratio β, p_t (=the moral value) of c_t (=individual actions) can be obtained through Formula (2). It should be kept in mind that the time factor is included in this equation. Specifically, as k_t (=the status of the system) in the current year (year t) is already fixed, c_t (=individual actions in the current year) improves k_{t+1} (=the status of the system in the next year) by the multiplying ratio β. In other words, p_t (=the value of individual actions in the current year) is dependent on q_{t+1} (=the value of the system in the next year). Here, β takes a positive value smaller than 1.

To summarize the political philosophies discussed in this chapter based on the individual-system relationships expressed by Formulas (1) and (2), Rawls' philosophy recognizes the relationship expressed by (1), but not the relationship expressed by (2). Under Rawls' philosophy, as p_t (=the value of individual virtue) is considered to be a given, it does not have to be justified by the social system. The value of the social system (q_t) [the social system that realizes Rawls' "two principles of justice"]) rests in the system's contributions to the pursuit of individual goodness by maximizing individual freedom. Therefore, under Rawls' philosophy, while Formula (1) is valid, Formula (2) is not. Our critique ("Critique 2") presented the following propositions: that it is necessary to provide some justification for p_t,; and that this justification should be provided by the social system—that is, the relationship expressed by Formula (2) is essential.

A "comprehensive doctrine" as referred to by Rawls means a political philosophy encompassing both Formulas (1) and (2). Meanwhile, a "political conception" as referred to by him focuses exclusively on developing the relationship expressed by Formula (1) while refraining from developing the relationship expressed by Formula (2). Hegel's philosophy is a comprehensive doctrine that recognizes the relationship expressed by Formula (2) as it maintains that the value of individual lives is justified by the progress of the world spirit. Conventional religions and the totalitarian ideologies of the 20th century (e.g., communism and Nazism) are also comprehensive doctrines.

In the modern world, the Rawlsian liberal political philosophy alone cannot provide value to individual lives (p_t). Until now, we have held on to comprehensive doctrines as implicit premises, and that is the belief in the grand narrative that was mentioned by Lyotard, namely the worship of reason, which is strongly related to the Hegelian philosophy. The worship of reason, which implies a belief in the unlimited progress of reason, is also a form of humanism. In the worship of reason, or under humanistic thoughts, reason is assumed to continue to progress forever. When the view spread that reason has its limitations, humanism faced a crisis. Why is it?

The answer to this question is provided by Formula (3), which is obtained by substituting the symbol p_t in Formula (1) with Formula (2). Formula (3) expresses the following relationship. q_t (=the status of the system [k_t] in the current year [the year t]) is obtained by discounting q_{t+1} (=the value of the system in the next year [the year $t+1$]) by the multiplying ratio γ.

In other words, according to Formula (1), q_t (=the value of the system in the current year) is obtained through the equation $q_t = \alpha p_t$, namely by multiplying p_t (= the value of individual goodness) by α. According to Formula (2), p_t (= the value of individual virtue) is obtained through the equation $p_t = \beta q_{t+1}$, that is, by multiplying the value of the system in the next year ($=q_{t+1}$) by β. According to

Formula (3), q_t in the current year is obtained through Formula (3), which combines those two equations, and this means that q_t is dependent on q_{t+1} (=the value of the system in the next year).

Formula (3) continues to be valid beyond the next year. If Formula 3 is extended to future years, the following formula is obtained:

(4) $q_t = \gamma q_{t+1} = \gamma^2 q_{t+2} = \gamma^3 q_{t+3} = \ldots = \gamma^N \times q_{t+N}$

The year $t+N$ represents an infinite future. In other words, q_t (=the value of the system in the current year) depends on q_{t+1} (=the value of the system in the next year), which in turn depends on q_{t+2} (the value of the system two years later), and this chain of relationship continues forever. If so, q_t (=the value of the system in the current year) ultimately depends on q_∞ (=the value of the system in the future). If γ is larger than 0 and smaller than 1, that means q_t is smaller than q_{t+1} in Formula 3. This indicates that the value of the system must continue to increase with the passage of time (Formula (5)).

(5) $q_t < q_{t+1} < q_{t+2} < q_{t+3} < \ldots < q_{t+N} < q_{t+N+1} < \ldots$

As indicated by Formulas (4) and (5), if the social system (including ideals such as the progress of reason, and institutions based thereon) is to have a positive present value (q_t), the value of the system must continue to be positive and growing forever. Therefore, the value of an ideal advocated by a comprehensive doctrine needs not only to be maintained forever but also to continue to increase with the passage of time.

In the context of Lyotard's "grand narrative," the reason why we find value q_t in the "progress of reason" is that we believe that the progress of reason will maintain its value forever and that the value q_t will continue to increase with the passage of time. If the progress of reason comes to a halt midway or if the value of the progress of reason is reduced to zero at some point in the future ($q_{t+N} = 0$), Formulas (4) and (5) cease to be valid. This means that the present value of the progress of reason is also reduced to zero ($q_t = 0$).

That is the situation meant by the "collapse of the grand narrative." Speaking more generally, if we stop believing that the value of a comprehensive social ideal (ideology) will continue to exist and grow forever, the ideal will instantly lose its value. That is why no sooner did we face the limitations of reason and stopped believing that the "progress of reason" would continue forever than the value of humanism came under critical threat.

Chapter 3 AI's Impact on the Relationship between Innovation and Justice

In this modern era, where faith in human reason has crumbled, it is the arrival of artificial intelligence (AI) that gives us an opportunity to once again believe in the everlasting progress of reason. Only after the arrival of AI has it become possible for us to imagine, as a vision of a feasible future, a world where we coexist with an intelligence that transcends humans as beings that perceive the world.

Part 3.1 Reason Expanded by AI

There are various approaches to developing AI. In recent years, in many cases, the term AI has been used to refer to a multi-layered neural network software program modeled on the human brain and which has been enhanced by deep learning. Of course, there are other AI systems that operate on different design concepts. However, in this article, AI refers to a multi-layered neural network within a computer that has been enhanced by deep learning.

It has become clear that as a result of the development of deep learning, AI systems enhanced by deep learning can achieve high performance, and in recent years, research and development on AI has made explosive progress. However, AI has surpassed the human brain not across all fields, but only in limited fields, including image recognition, and the *go* and *shogi* board games. Nevertheless, some forecasts predict the arrival of the AI singularity within dozens of years, which refers to the situation where AI has surpassed the capabilities of the human brain in all intellectual activity.

In any case, it has actually been verified that it is possible to reproduce at least some functions of the human brain within a computer and advance the reproduced function of the brain beyond the level of human intelligence.

It is possible to create an infinitely complex neural network structure within a computer. Sooner or later, it will become possible to create a computer-based neural network that is more advanced and complex than that of the human brain. As a result, in fields such as the natural sciences, for example, the possibility cannot totally be ruled out that AI may create an association of ideas that is beyond the human brain's ability to understand and bring about new "sciences" that cannot be understood by humans. Until recently, our worldview has held that humankind stands above everything else in the world as beings who have the most complete understanding of the world and greater control over our reality than any other animal on the planet.

However, AI has proved that something can exist in this world that can transcend humans in terms of our power to reason (although it is not a God). It does not matter whether or not the AI systems being developed by IBM and Google will surpass human reason. More important than how the development of any specific AI system will proceed is that it was shown that the possibility of a "rational being" that surpasses human reason can actually exist in human society as a vision of a feasible future. If the scope of human reason is expanded by an AI system with superhuman intelligence, the progress of the "expanded reason" may not reach its limitations for a while. In other words, although the faith in reason as an omnipotent force has declined, it may be possible to restore

the system of belief under a new progressive theory of history that maintains that the limitless expansion of reason will continue. That is also related to the question of whether the limitations of AI exist in the realm of human understanding.

Given that AI is a computer system, it should have limitations based on the laws of physics. Still, it is possible to create a computer neural network that is more advanced and complex than the human neural network. Such a computer neural network would be the same as the human neural network in terms of the basic network structure in that neural cells (elements) are linked by synapses, although there is a quantitative difference in that a larger number of neural cells would be linked together in a more complex way. The quantitative difference between the computer and human neural networks results in a qualitative difference in the perception of the world. That is in line with the argument made by Philip Anderson in his article titled "More is Different."

Let us compare human and chicken brains, for example. The chicken brain is the same as the human brain in that it is comprised of a neural network. The only difference is that in the chicken brains, a smaller number of cells are linked together in a simpler way. However, the chicken's perception of the world is qualitatively different from human perception. For example, when a chicken sees its own image reflected in a mirror, it mistakenly believes that there is another chicken there and tries to attack it. This qualitative difference—humans can recognize their reflection in the mirror as their own image but chickens cannot—stems from a quantitative difference, that is, the difference in the number of neural cells that constitute the brain, which also results in the difference in the way that the cells are linked together. Both human and chicken brains are the same at the basic level, i.e., in that both have a neural network structure.

A similar difference exists between AI and the human brain. AI, if equipped with a higher number of neural cells linked together in a more complex way, will surpass the human brain qualitatively, at least in some domains of reasoning. If AI becomes the driver behind the development of the "sciences," some aspects of the sciences will move beyond the reach of human comprehension.

That would be different from humans' inability to readily understand massive calculations performed instantly by computers. When a massive calculation is performed, humans can understand what is being done and use it in meaningful way because the task is of such a nature that it could be done by themselves in principle given the necessary time. However, new "sciences" that may be developed by AI will be qualitatively different in structure from the sciences that have been developed by humans, and as a result, the human brain will likely be unable to understand them regardless of the time spent, similarly to an animal’s inability to understand human sciences.

As explained above, the arrival of AI suggests the possibility that the sciences will evolve beyond the limitations of human reason. If an intelligence can evolve in a way that is qualitatively different from the sciences based on human reason, we can be assured that reason expanded through collaboration between AI and humans will continue to progress forever toward a true understanding of the world.

Such "expanded reason" enhanced by AI will continue to progress for a while without reaching the limits of human understanding. Thanks to AI, the progressive theory of history that Hegel conceived

of in *The Philosophy of History* can be renewed as something that advocates the progress of "expanded reason" instead of the progress of human reason (albeit with the proviso that all intelligences are fallible; I will explain in detail in the forthcoming articles).

Part 3.2 Relationship between Innovation and the Social System—Abstract

Below, I aim to identify the relationship of the value of individual virtue to the whole of the social system under the Rawlsian philosophy. If explained along the lines of the discussion that used mathematical formulas (Part 2.6), the objective is to introduce the relationship represented by Formula (2) into the Rawlsian liberal political philosophy, which recognizes only the relationship represented by Formula (1). By doing that, we can consider the question "What kind of life can we live in this era of AI?" When considering this question, faith in the infinite progress of "expanded reason," which was discussed in the previous section, is crucial. Furthermore, "innovation," as a concept that connects the progress of expanded reason to activities in individual lives, is also important. That is because it is innovation that determines the prior knowledge of people in the original position when the social system is chosen under the veil of ignorance. The structure of the discussion can be summarized into the following four points:

(i) The social system is valuable because it maximizes the expected utility of humans (individual virtue) under the veil of ignorance.
(ii) The prior knowledge necessary for upgrading the social system can be obtained through innovations (achieved by expanded reason enhanced by AI).
(iii) As innovations contribute to the upgrading of the social system, they have social value. In other words, the value of innovations springs from the value of a future, upgraded social system.
(iv) Individual virtue has social value in that it contributes to innovations.

The values of individual virtue, the social system, and innovation sustain themselves by creating a loop starting from individual virtue, the social system and innovation, and returning to individual virtue; that is, the social system is valuable as it enhances individual virtue, innovation is valuable as it enhances the social system, and individual virtue is valuable as it enhances innovation.

Part 3.3 What is Innovation?

I will once again begin a discussion from the viewpoint of the Rawlsian liberal philosophy. Under liberal political philosophy, human activity that triggers changes in scientific knowledge, i.e., innovation, has not been held in the highest regard. However, changes in scientific knowledge could significantly change the social system. That is particularly true regarding distributive justice. If this matter is considered under the framework of the Rawlsian political philosophy, changes in scientific knowledge may cause significant change to the application of the difference principle. For example, in 1971, when Rawls' *A Theory of Justice* was published, global warming was a mostly unknown

phenomenon outside of expert circles, but now, it is widely recognized as an important policy challenge that mankind must resolve. Society's concept of distributive justice regarding intergenerational distribution will change considerably depending on whether the difference principle is applied on the premise of the scientific knowledge that was shared society-wide in 1971 or on the premise of the scientific knowledge available 50 years later. In other words, the updating of scientific knowledge through innovations leads to the upgrading of the concept of justice by changing the consensuses formed in the original position under the veil of ignorance. To sum up, the motivation of humans and companies to engage in innovation activity is based on their pursuit of special interests based on their own desires and passions. However, the end result is that innovation creates new knowledge, which spills over to and freely circulates throughout society through people's learning processes. This phenomenon, called knowledge spillover, updates the collective knowledge of society. The society-wide common knowledge level is a precondition for people to agree to the social system under the veil of ignorance, which means that innovations change the system of a fair society.

This relationship presents an opportunity to connect individual virtue and social justice.

Innovation can be broadly defined as a quest for knowledge and investment in knowledge including activities of individuals and companies that cause changes in scientific knowledge. All activities conducted by people are intended to provide a better understanding and allow them to more effectively influence the world in order to achieve their respective goals. As innovation is an activity that is undertaken to find ways of better understanding and more effectively influencing the world, it may be said that all human activities involve innovation.

Innovation as defined in this article, unlike in the usual sense of the word, refers not only to engineering technology development and scientific invention and discovery, but to a quest for knowledge in the broad sense. Activities conducted by ordinary people in their daily lives, including those intended to "better understand and more effectively influence the world," are widely considered innovation in this article. Innovation in this sense includes intellectual discovery and invention that anybody may experience in everyday life or working life.

Innovations update the collective scientific knowledge of society through knowledge spillover. Under updated scientific knowledge, the social concept of justice is accordingly renewed. In a society where the concept of justice is determined by the principles of the Rawlsian political philosophy, each time scientific knowledge changes, parliament is convened (the Rawlsian original position is modeled on a parliament where individual parliamentarians engage in debate as the representatives of the general public with no regard to the special circumstances of their respective constituencies). In this parliament, a new system of distributive justice is determined as a result of the application of the difference principle to new scientific knowledge.

In this way, a quest for knowledge conducted as a private activity by individuals and companies, that is, innovation, enables the social system to progress. The more accurate the scientific understanding of the world becomes, the closer the social system moves toward completion.

Part 3.4 Do Selfish Individuals Implement Innovations Determined by Social Contracts?

Is the number of innovations implemented by individuals and companies based on personal decisions reaching the amount agreed upon under social contracts in the original position? A quest for knowledge (innovation) can be regarded as an investment activity intended to acquire new scientific and technical knowledge by using a certain quantity of resources (e.g., working hours and capital stocks). Innovation increases the productivity of the entire economy by creating new products and services and new methods of producing them. In other words, innovation has the function of expanding the economic pie. On the other hand, innovation also has a destructive nature: new products created through innovation, by replacing existing products, may cause incumbent companies' profitability to deteriorate, resulting in a change in the balance of power between people and between companies (a situation that Joseph Schumpeter called "creative destruction"). If defined in relation to the Rawlsian original position, innovation is an activity that updates people's scientific knowledge and reshuffles their positions under the renewed knowledge.

Moreover, the discovery of knowledge and development of new products as a result of innovation may be regarded as a phenomenon with an unknown probability distribution that intrinsically involves the "Knightian uncertainty," rather than as a stochastic event whose probability distribution is known ex ante. At least in my series, I will proceed with my discussions under that assumption. As was already mentioned, under the Knightian uncertainty, the basic principle of behavior of selfish individuals, is the same as the Rawlsian difference principle. Both of these principles are equivalent to the max-min rule, which advocates the maximization of the worst-case gains (minimum gains).

The question of what quantity of resources individuals should allocate to innovation under the Knightian uncertainty seems at first glance to be different from the question of what quantity of resources should be allocated to innovation as a society-wide choice in the original position. That is because the intensity and extent of the effects of those two sorts of uncertainty are assumed to be different: the effects of uncertainty brought about by the investment in innovation by individuals are limited to the individuals and their surroundings, while uncertainty brought about by society-wide investment in innovation affects the whole society. In this case, the amount of innovations pursued selfishly by individuals appears to be different from the amount of society-wide innovations pursued in the original position.

But that may not be true. The reason is as follows.

The reshuffling of individuals in response to innovations may be significantly affected either by innovations implemented by other people, or by the mutual effects of innovations by other people and innovations adopted by those individuals themselves. Given that the scope of the mutual effects of those innovations is complicated and wide-ranging enough to be unpredictable, it is appropriate to assume that uncertainty brought about by specific individuals affects not only themselves and people around them but also the whole of society. In this case, the probability distribution of the events corresponding to the uncertainty is unknown because this uncertainty is a Knightian one, and as a result, there is no difference between those individuals and people around them and the whole society in the degree of uncertainty faced. In other words, with respect to innovation, there is no difference

between the Knightian uncertainty faced by individuals and the uncertainty faced by the whole society. This means that the effects of innovation implemented by individuals could affect the whole society. As a result, the amount of innovations pursued selfishly by individuals and the amount of society-wide innovations pursued (per individual) in the original position are expected to be almost the same. The reason for that is that when individuals and the whole society face the same Knightian uncertainty, the amount of innovations pursued by individuals under the max-min rule is also the same as the amount of innovations agreed upon by people in the original position under the same rule.[1]

[1] The amount of innovations pursued under the veil of ignorance is not necessarily the socially optimal level. That is because there is an effect caused by knowledge spillover. People in the original position under the veil of ignorance do not take into consideration external effects. As a result, the number of innovations chosen in the original position may be undervalued (or overvalued) compared with the optimal level of innovations chosen by social planners in view of external effects as well. However, the externality is not the main topic of my writing. What I want to emphasize here is that the amount of innovations pursued selfishly by individuals in their lives is roughly the same as the amount of innovations pursued under social contracts in the original position.

As shown above, we have found a remarkable aspect of the nature of innovation activity: in innovation activity, private choices made by selfish individuals match social consensuses formed in the original position. Even though people shed the veil of ignorance and start to live within society, they come under that veil once again when they try to make decisions regarding innovation activity because of the updating of scientific knowledge (updated in an unexpected way). After all, in generation after generation, selfish individuals have to decide the quantity of resources to be allocated to innovation under the veil of ignorance, and as a result, with regard to innovation, people return to the original position in each generation. In each generation, scientific knowledge as an ambient condition is routinely updated, and people return to a renewed original position accordingly.

Part 3.5 Moderate Comprehensive Doctrine

Let me once again summarize the argument that we have made so far. Individuals determine the level of innovation under the Knightian uncertainty in accordance with the max-min rule in order to maximize their own benefits. The level of innovation thus determined mostly matches that of innovation on a society-wide basis that has been agreed by people in the "original position" under the difference principle (max-min rule). This suggests that the level of innovation pursued as part of the activity of individuals to pursue their own ends (individual virtue) matches that of innovation pursued under the society's concept of justice. In other words, it has been shown that the level of intellectual investment (innovation) made selfishly by individuals in accordance with their own life goals inevitably matches that of the innovation that should be realized in order to enact the concept of justice shared by the society as a whole.

Different people have different life goals, and activities they conduct in order to selfishly pursue their goals may entail innovation in some form or other, which results in the advancement of social justice. Until recently, doubt over the everlasting progress of reason has in turn created doubt over the significance and perpetuity of innovation. However, the explosive evolution of artificial intelligence driven by deep learning has now made it possible to believe that innovation resulting from the power of "expanded reason" will continue to make progress indefinitely.

We assume that we have succeeded in presenting a comprehensive doctrine by introducing intellectual investment (innovation) made by individuals into the Rawlsian concept of justice. Like popular economic theories, such as Adam Smith's theory of "the invisible hand" and Bernard de Mandeville's theory that private vices lead to public benefits, our comprehensive doctrine may be interpreted as implying that the selfish activities of individuals unwittingly contribute to society-wide public good. By introducing innovation as a key element, Rawls' theory of justice can be developed into a "moderate" comprehensive doctrine that is similar to popular economic theories.

The Rawlsian concept of justice (particularly the concept of distributive justice brought about by

the difference principle) ultimately converges toward a perfect form of justice through constant rethinking in accordance with increases in and updating of mankind's scientific knowledge. The concepts of virtue that individuals maintain unwittingly but inevitably contributes to the society's concept of justice. As a result, the value of such concepts is based in and justified by the society's concept of justice. Moreover, individuals can believe in the inevitability of progress—that is, they can believe that whatever life goals they may choose, those goals contribute to the advancement of social justice through innovation.

The findings of Smith and Mandeville in the field of economics have increased the significance of economic activity by fostering the belief that individuals' selfish economic activity is justified by society-wide justice, as does our argument. Increasing the social significance of economic activity was what they aimed to achieve.

The objective of the economic activity of individuals is to maximize their own benefits (or economic profits), but pursuing that objective was often seen as a shallow and vulgar life goal; at least, most intellectuals in the era of Smith and Mandeville must have felt that way. However, even though attempting to maximize one's own profit may be vulgar, that behavior itself contributes to public good in the form of an additional increase to the benefits that other people receive. That was what Smith and Mandeville found. As a result, the apparently vulgar behavior of seeking to maximize profit has come to be approved by the whole society (as it advances public good) and to be recognized as a noble activity. Maximizing profit for one's own sake has ceased to be seen as a disgraceful goal and has come to be seen as an activity that deserves to be a legitimate subject of academic study. In this way, "economics" was established as a field of academic study for the first time.

In making our argument in this article, we are attempting to do in the field of Rawlsian political philosophy what Smith and Mandeville achieved in the field of economics. What individuals aim to achieve in their life may be in itself vulgar and insignificant and appear unlikely to be justified by social justice. On the other hand, individuals can only believe that their lives are worthwhile if their life goals are justified by social justice. However, under the political philosophy of Rawls' theory of justice, individuals' life goals could not be justified by social justice because individuals' own particular goals were assumed to be entirely disconnected from social justice.

We have demonstrated that individuals' pursuit of their own goals contributes to the advancement of some aspects of public good (expansion of justice) by raising the intellectual level of the society through the creation of innovation. Pursuing selfish goals unwittingly advances social justice through innovation. Therefore, individuals pursuing their own life goals according to their own desires and passions can inevitably be considered socially valuable.

Part 3.6 Considering the Theory of Innovation-Driven Justice in Terms of Mathematical Equations

If our argument is to be summarized in terms of mathematical equations, Formulas (1) to (3) shown in Part 2.6 apply. The symbol c_t refers to individual activity conducted in the year *t.* The moral value of c_t is expressed as p_t. The status of the Rawlsian social system in the year *t* is expressed as k_t. The moral value of k_t (=the status of the system) is expressed as q_t.

The meaning of Formula (1) is that the social system is valuable because it contributes to individual virtue. When the value of an individual life is expressed as p_t, if k_t (=the status of the system) is to increase c_t (= individual activity) by the multiplying ratio α, q_t (=the value of the system) can be obtained through Formula (1).

$q_t = \alpha p_t$	(1)
$p_t = \beta q_{t+1}$	(2)
$q_t = \alpha \times \beta \times q_{t+1} = \gamma\, q_{t+1}$	(3)
(however, $0 < \alpha < 1, \quad 0 < \beta < 1, \quad \gamma = \alpha \times \beta$)	
$q_t = \gamma\, q_{t+1} = \gamma^2 q_{t+2} = \gamma^3 q_{t+3} = \cdots = \gamma^N \times q_{t+N}$	(4)
$q_t < q_{t+1} < q_{t+2} < q_{t+3} < \cdots < q_{t+N} < q_{t+N+1} < \cdots$	(5)

The meaning of Formula (2) is that individual virtue is valuable because it contributes to the social system. When the value of the social system q_t has a given value, if it is assumed that individual activity improves the status of the system by the multiplying ratio *β,* the moral value p_t of individual activity c_t can be obtained under Formula 2. The time factor is included in this equation. More specifically, as the status of the system (k_t) in the current year (year *t*) is already fixed, individual activity in the current year (c_t) improves the status of the system in the next year (k_{t+1}) by the multiplying ratio *β*. Formula (3) is obtained by substituting the sign p_t in Formula (1) with Formula (2).

Formula (3) expresses the following relationship. The status of the system (k_t) in the current year (the year *t*), expressed as q_t, is obtained by discounting the value of the system in the next year (the year $t+1$), expressed as q_{t+1}, by the multiplying ratio γ.

Formula (3) continues to be valid beyond the next year. If Formula 3 is extended to future years, Formula (4) is obtained. The year $t+N$ represents an infinite future. In other words, the value of the system in the current year (q_t) depends on the value of the system in the next year (q_{t+1}), which in turn depends on the value of the system two years later (q_{t+2},) and this relationship continues forever. If so, the value of the system in the current year(q_t) ultimately depends on the value of the system in the future (q_∞). As the value of individual goods (p_t) also depends on $q_{\{t+1\}}$, it is determined on the basis of the value of the social system in an infinite future.

$\gamma q_{t+N+1} = q_{t+N} = 0$	(6)
$q_{t+N-1} = \gamma q_{t+N} = 0$	(7)
$q_{t+N-2} = \gamma q_{t+N-1} = 0$	(8)
$q_t = q_{t+1} = q_{t+2} = \ldots = q_{t+N} = 0$	(9)
$p_t = \beta q_{t+1} = 0$	(10)

If reason has limits, that is, if innovation cannot last forever, such a chain of value cannot be realized. Let us assume that innovation will dry up at some point in the future (the year t + *N*) due to the limits on the progress of reason. In that case, under Formula (2), individual activity does not contribute to the updating of the social system, so it can be interpreted that β is zero ($\beta = 0$) in the year *t* + *N*. As a result, the following equation is obtained: $\gamma = \alpha \times \beta = 0$. If γ *is zero* ($\gamma = 0$) in the year *t* + N under Formula (4), γ is zero ($\gamma = 0$) under Formula (6) as well, resulting in the equation $q_{t+N} = 0$. In that case, even if γ continues to be larger than zero ($\gamma > 0$) until the year before the year $t + N$, q_{t+N-1} is zero ($q_{t+N-1} = 0$) under Formula (7), while q_{t+N-2} is zero ($q_{t+N-2} = 0$) under Formula (8). If the time sequence of the chain of value is traced back in time as above, ultimately, Formula (9) applies, which means that q_t is zero ($q_t = 0$). In other words, if innovation cannot last forever, the moral value of the social system at present (the year *t*) is zero, and consequently, the moral value of individual virtue p_t is also zero at present under Formula (10).

Today, in the 21st century, humans have become able to believe that expanded reason through the power of artificial intelligence will continue forever, with the result that the creation of innovations by humans equipped with expanded reason will continue for a while without reaching any limit. That has made it possible for us to establish a "theory of innovation-driven justice" under which the perpetual chain of moral value represented by Formulas (1), (2) and (3) is realized. This is a moderate doctrine that comprehensively explains that individuals' lives have moral value, and that the social system also has value that should be preserved.

Part 3.7 From Economic Growthism to Intellectual Growth

Finally, we will consider the significance of our argument (theory of innovation-driven justice) from a different angle. As a result of the collapse of grand narratives in the early modern era, the faith in the progress of reason that was prevalent in the 19th century has been lost in modern society. "Economic growthism," which will be explained below, has filled that void and become a religion for modern society. The theory of innovation-driven justice that has been discussed in this chapter may provide clues that allow human society to pull itself out of the dead end of economic growthism.

It is simply strange that when we consider the sustainability of the global environment or the human world, a society with everlasting economic growth is upheld as an ideal society in academic and policy discussions, as indicated by the Japanese government's pursuit of "2% growth." As the reserves of natural resources and energy available on the earth are limited, it is impossible for production volume or energy consumption to continue growing forever. For example, if GDP continues growing at an annual pace of 2%, it will expand by a factor of 400 million in 1,000 years, which is an unrealistic prospect. As that kind of growth is impossible, human society is certain to reach a static state with no growth at some point in the future. On the other hand, political leaders across the world invariably uphold economic growth as a national goal, and most people also rally behind that goal.

We may say that economic growth is the "religion" of modern society in that the whole society aims to achieve it with no regard for scientific feasibility. There are several politico-philosophical reasons why economic growth is upheld as a society-wide goal (economic growthism).

The first reason is that pursuing economic growth as an economic policy goal is quite harmonious with modern liberalism. As Michael Sandel noted, in a modern liberal society, the government cannot determine a single value that should be pursued as a national goal because the population has diverse values (Sandel, 2010 and 2011). Until the 19th century, economic policy arguments in the United States represented clashes over values. For example, there was a clash between the Jacksonian values, which upheld the goal of developing an industrial structure that was conducive to fostering independent minds among farmers, and the Hamiltonian values, which advocated the expansion of trade through the promotion of large enterprises. From the 20th century onwards, we have seen the spread of a large variety of values, and implementing economic policy based on a particular set of values has come to be seen as imposing values on the people, which is politically unpopular. Economic policy in a liberal society must be value-neutral. This means that by default, aiming to expand the "gross volume" of goods and services, that is, GDP, becomes the only option available. As a result, in a liberal society where the diverse values of individuals are respected, achieving value-neutral economic growth becomes a society-wide goal.

The second reason why economic growthism is very popular is that it allows ruling politicians to put off their "day of reckoning." John G. A. Pocock (professor emeritus, John Hopkins University), a historian of political thought, characterized commerce as a "virtue" while referring to the thoughts of federalists immediately after the independence of the United States (Pocock, 2008). Virtue in this context has a special meaning, that is, policymakers' ability to start a new commitment without honoring an old one. If economic growth expands the societal frontiers, politicians who previously committed themselves to redistributing income to various corners of society can present a new goal and make a new commitment, ignoring the old one. That does not mean that they break the old commitment, as it is embedded into the new commitment. If the ability to make a new commitment without breaking an old commitment is to be called a virtue, pursuing economic growth is then a virtue.

Honoring the old commitment of redistributing income usually involves great pains, but the "day of reckoning" for that commitment can be put off indefinitely as long as economic growth continues. Achieving growth is the same as putting off the day of reckoning. That is why economic growthism is favored by politicians.

The third reason why economic growthism is prevalent in modern society is related to the collapse of grand narratives that has been discussed in my series. When the early modern faith in the progress of human reason was lost as a result of the experiences of the 20th century, the values that made individuals' lives worthwhile (the common values of the whole society) were also lost. The progress of reason as an ideal has been replaced by a modern value, namely economic growth.

To go back to Hannah Arendt's argument, the fundamental experience of modern people is that of "being abandoned" (Arendt, 2017). In modern society, where individuals do not have a permanent place of belonging from the moment of birth due to the collapse of traditional communities, many people feel that they are useless to society and are needed by nobody. As humans cannot endure that experience, they immerse themselves in totalitarian ideologies. This observation by Arendt indicates that the lives of free individuals cannot become worthwhile unless their lives have some significance within the social system of values.

In the early modern era, the faith in the progress of reason was the social value that made individuals' lives worthwhile, but this grand narrative collapsed. Afterwards, material wealth became the only common value among the diverse values of individuals that makes their lives worthwhile. That is the reason why economic growth has become a society-wide goal.

Individuals can feel worthwhile by contributing, directly or indirectly, to economic growth. In this respect, achieving economic "growth" is fundamentally more important than anything else. If individuals do not see any improvement in their living standards despite their best efforts, they cannot gain the sense that they have contributed to economic development. Let us recall Formula (5) in Part 2.6. If individuals are to find the meaning of their lives in contributing to economic development, the economy needs to "grow." As a result, pursuing "economic growth" as a goal has become a broad consensus in modern society.

Moreover, the growth must be "everlasting." If economic growth is to come to an end at some point in the future, the factor that gives meaning to individuals' lives (=growth) disappears at that point. If individuals' lives are destined to become meaningless in the future, no meaning can be imparted to their lives at present, either. People believe that their lives today are worthwhile because their lives tomorrow are expected to be worthwhile. If their lives tomorrow are set to become meaningless, that reasoning crumbles. Therefore, growth must be everlasting.

It is physically impossible for economic growth to continue forever. To make a future human society sustainable, we should pursue some form of growth other than economic growth as a goal. The "progress of intellect" due to the power of artificial intelligence and information technology may

replace economic growth as a new social ideal. That possibility is indicated by the theory of innovation-driven justice that was discussed in Part 9 to Part 15. If human intellect is enhanced by artificial intelligence, the possibilities for development of the intellect expand infinitely, while the progress of the intellect is unlikely to be bound by either resource or environmental constraints. That kind of "growth" can give meaning to the lives of free individuals and maintain the sustainability of the world at the same time. We must find a new theory of political philosophy that transcends economic growthism.

Chapter 4 System of Justice as an Intergenerational Asset

It is known that under the Rawlsian liberal philosophy, intergenerational sustainability problems cannot be successfully resolved. The "theory of innovation-driven justice," which was discussed in the previous chapter, maintains that individuals' lives create social value because their innovations contribute to the society-wide system of justice. The time factor under the theory of innovation-driven justice can create a natural motive to deal with intergenerational problems. That is the theme of this chapter.

Part 4.1 Time Inconsistency Problem

In the short span of the last few decades, a host of policy challenges related to intergenerational sustainability, including fiscal crises and global environmental problems, have emerged. If modern people—that is, selfish and rational individuals—are to make political decisions in ways that maximize the benefits for themselves, they cannot undertake acts of self-sacrifice for future generations, because increasing the benefits for future generations at their own expense without the prospect of receiving any return only reduces the benefits for themselves.

That is also true of a society that shares the two principles of Rawlsian justice. Rawls argued that in a society facing intergenerational problems, "just savings" occur. Since "savings" in this case means what is bequeathed from one generation to succeeding ones, it may as well be called "legacy."

Rawls assumed that when people are in the "original position"—that is, when they are enveloped in a veil of ignorance as to which generation they may be born into—they agree on the principle of intergenerational "savings." In this case, the rules on just savings are determined by the "difference principle." Let us take up global warming as an example. When people are ignorant (in the original position) as to which generation they may be born into, one of their fears is that they could be born into the most unfortunate generation. In this case, the most unfortunate generation is the generation that suffers the worst damage from climate change due to global warming. Therefore, those people try to set rules on intergenerational bequeathal of resources (the global environment in this case) in a way that minimizes the damage from global warming for the generation that is expected to suffer the worst damage. As a result, it is assumed that there will be agreement that each generation should implement global warming mitigation measures under the principle of just savings, so that the sustainability of the global environment can be ensured. Thus, it appears as if the sustainability problem can be resolved by applying the veil of ignorance theory across generations.

However, it is difficult to reach such an agreement because selfish individuals essentially do not self-sacrifice for future generations.

That is because in the real world, selfish people honor an agreement concluded as a social contract

in the original position because it is difficult to rescind a commitment after the veil of ignorance has been removed. Under the Rawlsian political philosophy as well, human beings are selfish creatures. However, that applies only to agreements that apply within the same generation. On the other hand, regarding the rules on intergenerational savings, it is very easy to rescind any commitments made. Let us take a closer look at the differences between agreements that apply within the same generation and ones that apply across generations. In the case of agreements concerning resource allocation within the same generation (e.g., an agreement on a social welfare system for economically vulnerable people), strong resistance arises in response to attempts to change the rules because all interested parties live in the same generation even after the removal of the veil of ignorance. Since the people for whom there is a conceivable risk of becoming poor represent the majority in a society, the wealthy, who are in the minority, cannot have their way under a democratic system even if they insist on going back on an income transfer agreement (the creation of a democratic government is also based on an agreement concluded as a social contract in the original position), meaning that it is difficult to rescind an agreement that applies within the same generation.

However, in the case of agreements on intergenerational resource transfers (the rules on just savings), after the removal of the veil of ignorance, on one side, there are interested parties who live in the time period when the agreement has been concluded (the present generation), while on the other side, there are those who do not (succeeding generations). What will the reaction of the selfish people of the present generation be when the situation of the present generation (e.g., the not-yet catastrophic level of global warming) becomes clear? It is obvious that regardless of what kinds of rules on intergenerational savings the previous generations may have applied in resource allocation, bequeathing nothing to succeeding generations is the best option for the present generation. Therefore, the present generation tries to change the rules on "just savings," while the people who would oppose the change (i.e., succeeding generations) have yet to be born. The present generation can rescind the agreement on just savings that was concluded in the original position without interference from anyone. As a result, under the premise that human beings are selfish creatures with no altruism for the interests of future generations, it is impossible to realize intergenerational just savings and it is therefore difficult to maintain the sustainability of human society in the long term.

The rules on just savings are "time inconsistent" in that they can be agreed upon but are destined to be reneged on later. Rawls himself was aware that the time inconsistency problem is involved in the rules on just savings. In *A Theory of Justice*, he recanted his assessment of human nature, stating that human beings are not completely selfish but may behave altruistically, taking into consideration the interests of their offspring as the head of a family. Assuming that human beings may behave as the head of a family, rather than merely as an individual, is tantamount to assuming that they have strong altruism for the interests of future generations. If people have strong altruism for the interests of future generations, the time inconsistency problem may be resolved, to be sure. People may try to honor the

rules on just savings across generations. However, in the real world, we have been stalling in efforts to scale back ballooning government debts or to combat global warming. This reality suggests that people's altruism for the interests of future generations is fairly weak. The biological altruism that is rooted in human nature is too weak to resolve sustainability problems such as fiscal and environmental crises.

Part 4.2 System of Justice as an Asset

Let us think in terms of mathematical formulas once again (for the previous references to mathematical formulas, see Part 2.6). Under the Rawlsian philosophy, only Formula (1) is valid, which means that the time factor is not reflected in the value of the system of justice. The system of justice is valuable in that it provides an optimal means to pursue individual goods (within the same generation). From the position of our moderate comprehensive doctrine, which maintains that individuals' actions are valuable because they contribute to the renewal of the system of justice through innovations, Formula (2) is valid. In this case, Formula (3) is arrived at through Formulas (1) and (2).

$q_t = \alpha p_t$	(1)
$p_t = \beta q_{t+1}$	(2)
$q_t = \alpha \times \beta \times q_{t+1} = \gamma\, q_{t+1}$	(3)
(however, $0 < \alpha < 1, \quad 0 < \beta < 1, \quad \gamma = \alpha \times \beta$)	
$q_t = \gamma\, q_{t+1} = \gamma^2 q_{t+2} = \gamma^3 q_{t+3} = \cdots = \gamma^N \times q_{t+N}$	(4)
$q_t < q_{t+1} < q_{t+2} < q_{t+3} < \cdots < q_{t+N} < q_{t+N+1} < \cdots$	(5)

The present moral value of the system of justice (q_t) is determined by its future value (q_{t+1}). Given the chain of value that extends across time, the present moral value of the system of justice (q_t) is determined by its moral value in the infinite future (q_∞). In other words, the present moral value of the system of justice (q_t) may be understood to be a kind of "asset price."[2]

[2] To use the terminology of economics, Formula (3) indicates that q_t may be considered to be the price of a "bubble asset" that does not deliver dividends.

The system of justice (k_t) itself may be understood to be an intergenerational, ultra-long-term asset that is assumed to be passed on indefinitely across generations. If the present generation is to undertake acts of self-sacrifice in order to resolve intergenerational problems, preserving intergenerational assets may be the motivation for doing that. For example, let us assume that if nothing is done to deal with global environmental problems at present (*year* t), human society will collapse *N* years later. In this case, the value of the system of justice *N* + 1 years later ($q_{\{t+N+1\}}$) becomes nil. As a result of the chain of value under Formula (4), the following equation is arrived at: $q_t = q_{\{t+1\}} = 0$ (the present value of the system of justice also becomes nil). This means that the value of life (moral value) for the individuals of the present generation (p_t) is also reduced to nil as follows: $p_t = \beta\, q_{\{t+1\}} = 0$. To avoid this situation, the people may reach an agreement on undertaking the acts of self-sacrifice that are essential in order to resolve the global environmental problems.

Let us consider a case in which the collapse of human society *N* years later can be prevented if the present generation pays the cost of environmental protection measures (*X*). The value of the present generation's actions is expressed as $p_t c_t$ if the cost *X* is not paid, whereas the value is expressed as $p_t(c_t - X)$ —i.e., the value declines—if the cost is paid. In a world where the Rawlsian theory of justice prevails, although Formula (1) is valid, Formula (2) is not, because the value of individual virtue (p_t) is a given. In that world, the present generation chooses not to pay the cost X because paying the cost means a smaller benefit ($p_t(c_t - X)$) compared with the benefit to be gained if the cost is not paid ($p_t c_t$). Therefore, in the world that works under the Rawlsian theory of justice, the people cannot reach an agreement on paying the cost of environmental protection measures, with the result that the environment will continue to deteriorate.

On the other hand, under the theory of innovation-driven justice that we have been discussing in this paper, we arrive at a different conclusion. In a world where the theory of innovation-driven justice prevails as the people's worldview, both Formulas (1) and (2) are valid. As a result, although the benefit for the present generation is higher than zero ($p_t > 0$) if the present generation pays the cost of environmental protection measures *X* just as in the previous case, the benefit is nil ($p_t = 0$) if the cost is not paid.

In this case, the benefit is expressed as $p_t(c_t - X)$ if the cost *X* is paid, whereas it is expressed by $p_t c_t = 0 \times c_t$, if the cost *X* is not paid, which means the benefit is nil. If it is assumed that the net consumption is higher than zero ($c_t - X > 0$), the benefit for the present generation to be gained if the cost *X* is paid is greater than the benefit to be gained if the cost is not paid. As a result, the people of the present generation agree on implementing environmental protection measures.

As illustrated by this case, if we accept the moderate comprehensive doctrine's proposition that the value of individual goods depends on the value of the system of justice, undertaking acts of self-sacrifice in the fight against global warming would be considered to be a rational decision on the part of the people of the current generation from the viewpoint of preserving the present value of life (p_t)

for themselves. What is necessary is to conceive a new vision of intergenerational ethics that reflects this.

Chapter 5 Fallibility as a Reason for Guaranteeing Freedom

The theory of innovation-driven justice (a comprehensive doctrine arguing that the value of the goals of individuals' lives may be justified by their contributions to the renewal of social justice through innovation), which we have discussed in the previous parts, may have given the impression of being a doctrine that emphasizes the importance of the development of the entire society while belittling the significance of individual freedom—a sort of totalitarianism—because of its analogy to the Hegelian philosophy. However, our argument is based on two overarching premises: Rawls' first principle (guarantee of individual freedom) and Hayek's free market theory. Here, we would like to emphasize that point. The most critical point of the premises of the argument is that the development of our society is fundamentally fallible. What the evolution of artificial intelligence (AI) through deep learning indicates is that the development of intelligence in itself represents the creation of order in the physical world. All forms of intelligence, including AI, are fallible in that there is no guarantee that they embody absolute truth, so they may eventually turn out to be wrong.

Awareness that all forms of intelligence are fallible is the key to addressing the second question in Part 1.2—whether there is a risk that AI could wipe out humanity. If AI systems or AI-enhanced superhumans have an awareness of the fallibility of all forms of intelligence, they are likely to aim to maintain the diversity of the world, that is, to co-exist with the existing human race, as will be explained below.

Part 5.1 Hayek's Knowledge Theory

The point of argument that we presented regarding the future role of political philosophy is that individual virtue (the goals of individuals' lives) may be justified by a system of justice (social goals). Innovation is the key that allows individuals to gain approval from society for their own lives' goals, based on their own passions and interests. Innovation undertaken by individuals reshuffles the scientific knowledge and social circumstances among people, thereby renewing the social system of justice. Individuals' selfish activities may unwittingly contribute to society-wide interests through innovation. That is how individual virtue is socially justified.

Our argument presented above is based on the premise of individuals practicing innovation, but it goes without saying that innovation is the results of individuals' free activities, so there is no innovation where there is no freedom. In a society where Rawls's first principle (guarantee of individual freedom) does not apply, our argument that society provides a rationale for individual virtue does not hold. Therefore, the guarantee of individual freedom must exist as the foundation of our society. Below, we will delve a little further into the ideas underlying that argument.

Hayek, in a well-known paper in 1945 titled "The Use of Knowledge in Society" (Hayek, 1945),

emphasized the importance of the "knowledge of the particular circumstances of time and place," as opposed to general knowledge. First, Hayek addressed the question of what problems must be solved in order to create a "rational economic order." Although Hayek does not provide a clear definition for his "rational economic order," the term is presumed to refer to calm economic conditions where there is no extreme unemployment or bankruptcy level in which aggregate supply balances aggregate demand to a certain extent.

Governments do not have centralized, comprehensive access to information as to what sorts of demand exist in the market and which products and services are supplied in what volumes. The information that is necessary for solving economic problems exists only in the form of fractions of knowledge possessed in a decentralized manner by unrelated individuals scattered throughout the market, and those pieces of knowledge are imperfect and may easily degrade or be lost with the passage of time. Specific pieces of knowledge may be inconsistent with one another in many cases, and it is not possible to know in advance and may remain unknown even on an ex-post basis, which bits of information are accurate, and which are not.

It is impossible for any single entity (like a governmental central planning bureau) to develop a plan for integrating those vast amounts of decentralized pieces of knowledge and to establish a rational economic order. An economy-wide order is created as a result of the determination of economic variables in a decentralized environment where individuals possessing decentralized pieces of knowledge that exist across various circumstances that exist in a particular time and place engage in economic transactions and exchange information with one another as part of their free activities. Only through such a market competition mechanism can problems that block the creation of economic order be solved. That is Hayek's argument.

A governmental central planning authority would not have access to the entire body of knowledge concerning the country's economy. This is not merely because the volume of knowledge is too large (if the problem were merely one of volume, the development of computing technology would resolve it) but because the sort of knowledge that cannot be expressed in language or mathematical formulas (tacit knowledge) plays an important role in the real-world economy and society. Such tacit knowledge is the "knowledge of the circumstances that exist in a particular time and place" as referred to by Hayek.

From Hayek's point of view, a market competition mechanism (price competition) is not an inorganic system whereby price automatically settles at the level where supply matches demand but is probably a kind of learning mechanism whereby decentralized bits of tacit knowledge scattered across countless numbers of people are integrated to create economic order.

As described in previous parts, we know that in recent years, AI has made explosive progress because of the development of a learning process known as deep learning. When using the deep learning process for the purpose of image recognition, by "examining" vast amounts of data, AI

systems learn to identify the properties of visual inputs that cannot easily be expressed in language or mathematical formulas. For example, when an AI system learns to recognize the collection of visual patterns that represent a cat, the AI neural network undertakes an iterative process of modifying its own criteria for "cat" as it inputs visual data, eventually gaining the capability to respond appropriately when prompted with an image of a cat. A response pattern like that (a pattern of visual stimuli that corresponds to a certain concept) is shown in terms of features. The process of an AI system creating a pattern of feature values that corresponds to the concept of a cat is equivalent to the process of a human being creating the concept of a cat within the brain.

Innovation in a market economy due to individuals' free activities may be a similar process to the way that a machine learns to value features through deep learning. Although this is not part of the argument that Hayek made himself, we can consider what type of market function makes it possible for individuals to create innovation by comparing Hayek's idea of the market economy (i.e., the market function of collecting and integrating bits of knowledge scattered across circumstances that exist in a particular time and place) with the deep learning processes of AI systems in recent years.

Part 5.2 Innovation and a New School of Economics

At the heart of Hayek's idea of the market economy is the thinking that an economic order ("equilibrium," to use the terminology of modern economics) is created as a result of the transmission of the benefits of pieces of tacit knowledge relating to the circumstances that exist in a particular time and place across the entire market through economic transactions. Although those pieces of tacit knowledge themselves cannot be universally disseminated, their benefits can spread through the entire market in the form of changes in price as an economic variable. In traditional economics, based on the idea that price is the only information that can be transmitted across the entire market, analysis has focused exclusively on the sort of economic order in which supply and demand volumes are equal (that is, a state of equilibrium). To be sure, it is true that price is the only information that can be quickly transmitted across the entire market.

However, on closer inspection, some pieces of tacit knowledge are transmitted from one part of the market to people in the vicinity through market transactions, leading to changes in their activity. Those changes may be ones that cannot be expressed in terms of quantifiable variables such as price. However, there is no doubt that the transmission of tacit knowledge is occurring and causing local changes in some particular parts of the market, and we may describe this as a process of the market "learning" in the same sense that AI learns.

As a result of the transmission of pieces of knowledge other than price information that are related to the circumstances that exist in a particular time and place in the vicinity through networks of people in the market and their contact with other pieces of knowledge, a reaction that is similar to a chemical

reaction occurs. It is well known that networks of people and companies in the market are complex, with some particular structures of linkages (Masuda and Konno, 2010). According to Masuda and Konno, in a social relationship network of friends and acquaintances, for example, two people who are total strangers are six or fewer social connections away from each other in many cases. In other words, two randomly chosen people are likely connected with each other through a line of a few to a dozen acquaintances. This nature of a network of people is known as “small-worldness” under complex network theory.

As economic relationship networks of individuals and companies are also considered to have small-worldness, naturally, knowledge other than price information (particularly unquantifiable tacit knowledge) is assumed to spread very widely to individuals and companies through economic transactions. Local interactions caused by such contact involving scattered pieces of tacit knowledge are probably the driving force of innovation. In other words, innovation occurs through the following mechanism. Patterns that cannot be expressed in language are identified through economic transactions, and the knowledge of the identified patterns is used by individuals and companies to develop new business approaches in order to increase their own profits. The new approaches developed are innovations. If the approaches can be expressed in language or mathematical formulas, they become scientific knowledge that may be universally disseminated.

Conventional economics has developed theoretical frameworks with an exclusive focus on features such as price and quantity. Until now, the mechanism behind innovation—whereby new local features are identified through a certain learning process as a result of the transmission of tacit knowledge that cannot be expressed in language through a network of transactions with a small-world nature—has been outside the understanding of economics. Whereas economic analysis has so far focused on price information that can be shared throughout the entire market, a new breed of economics that analyzes local innovation activity that occurs in the circumstances that exist in a particular time and place may be conceptualized in the future.

The development of theories regarding the deep learning mechanism has only just begun, but if we unleash our imagination, it will be possible to imagine that a new breed of economics may be conceptualized within the paradigm of a general learning theory for machine or human learning, to which deep learning belongs. That sort of economics would probably be the new breed of economics that Hayek wanted to conceptualize.

Understanding the abovementioned nature of the innovation process makes it clear why freedom is essential to realizing innovation. Innovation is a process whereby people and companies (they may be equipped with AI-enhanced intelligence) develop a new approach to influencing the world in a more effective manner by acquiring local bits of tacit knowledge. A governmental central planning authority in principle would not have access to pieces of tacit knowledge that only exist in the circumstances that exist in a particular time and place. Only when bits of tacit knowledge are allowed to be freely

utilized by the individuals who possess them can they trigger innovation by being transmitted to other people and companies in the vicinity. In a society and economy without freedom, it is extremely difficult to create innovation because tacit knowledge is not transmitted through a network of economic activity. That is one reason why individual freedom should be guaranteed if innovation is to be realized in an effective manner.

However, from a contrarian viewpoint, we could argue that if efficiently transmitting tacit knowledge is the only important point, individual freedom is not necessarily essential. In the future, as a result of the development of scientific theories regarding how to acquire tacit knowledge most efficiently, it could be discovered that a controlled economy that forcibly reshuffles transaction partners based on governmental rules that are set in advance is more likely to create innovation than a free market economy.

In other words, the possibility cannot be ruled out that a more effective creation of innovation alone does not serve as a standard for fully justifying individual freedom and that a totalitarian society without individual freedom could create innovation depending on the design of institutional frameworks.

However, there is another reason why individual freedom should be guaranteed, which is that all sorts of ideas created by humans or AI are fallible. Put simply, even if a theory is discovered which states that a totalitarian society could be the best vehicle for creating innovation, the possibility cannot be ruled out that the theory itself is wrong.

Part 5.3 From the Quest for Infallibility to Fallibility

All human knowledge regarding this world is “provisional” truth. Every piece of knowledge, be it scientific knowledge or experience-based implicit knowledge, could be proven “wrong” at some point in time. In this sense, knowledge is fundamentally fallible. Just like humans, artificial intelligence (AI) learns from data—although it uses deep learning—so even if it is capable of creating a superhuman system of knowledge, it cannot escape from fallibility. That humans (and also AI as something that embodies an extension of human knowledge) are fallible is probably the fundamental reason why our society should be a free society.

Here, let us recall Arendt's argument (Refer to Part 2.1, Part 2.2, Part 2.3, Part 2.4, Part 2.5 and Part 2.6).

According to Arendt, absolute faith in the “infallibility” of totalitarian leaders was the driving force behind totalitarianism. In modern society, lonely people who have lost a traditional community as their place of belonging and who are exposed to market competition feel that they have been abandoned and forgotten. Abandoned people seek security (that is, something infallible) as a mental anchor, but in the real world, where nothing is secure, only an ideology disconnected from reality that is advocated

by a totalitarian political party can serve as such an anchor. If the people take the infallibility of a certain political party's ideology for granted, they cannot help but distort reality to suit the ideology's logic. The massacres and mass purges perpetrated by totalitarian states were undertaken thanks to a deductive logic based on the assumption of the rulers being infallible. With a faith in their rulers' infallibility, the people share a single ideology and their course of action is determined by the force of deductive logic that springs from that ideology. As a result, individual freedom is inevitably oppressed.

Conversely, in the absence of faith in the rulers' infallibility, there is no reason for the people to abandon their free will and follow the rulers' orders. In some countries, the people longed for infallibility in their rulers, and that is exactly why totalitarianism took hold there, with the people believing in the infallibility of the ideology advocated by the rulers and agreeing to follow their orders. However, if the people believe that all humans, including the rulers, are fallible, they cannot help but retain their freedom. A lack of faith in anybody's infallibility would prevent them from blindly trusting a dictator and surrendering to his rule.

What creates the longing for infallibility?

To consider this question, let us recall the three mathematical formulas that I mentioned earlier.

$q_t = \alpha p_t$	(1)
$p_t = \beta q_{t+1}$	(2)
$q_t = \alpha \times \beta \times q_{t+1} = \gamma\, q_{t+1}$	(3)
(however, $0 < \alpha < 1, 0 < \beta < 1,\ \gamma = \alpha \times \beta$)	

Here, p_t represents the moral value of an individual life, while q_t represents the moral value of a social ideal. Formula (1) implies that a social ideal is worthwhile because it helps to improve an individual life. Formula (2) implies that an individual life is worthwhile because it contributes to the development of a social ideal. Formula (3) implies that the current value of a social ideal, represented by $q_{t,}$ depends on its value for the next year, represented by $q_{\{t+1\}}$, which in turn depends on its value for the next year plus one, represented by $q_{\{t+2\}}$, and this chain of consequence continues infinitely. As a result, the current value of a social ideal (q_t) depends on its value for the infinite future (q_∞).

There are various social ideals, including not only ideologies like communism and liberalism but also faith in the righteousness of ideologies. If an ideology is proven wrong at some point in the future (e.g., in $t + N$ year), its value at that point ($q_{\{t+N\}}$) will be zero, and, according to the logic of Formula (3), the current value is also zero. That should not be allowed to happen, so a social ideal (ideology) must always be right. This is the logic of infallibility.

In other words, under a comprehensive doctrine under which the value of an individual virtue (an individual's purpose of life) is justified by a social ideal, the ideal is inevitably required to be

permanently infallible. That is why ideologies like totalitarianism self-proclaim their infallibility. However, on closer reflection, we realize that it is impossible for humanity to create social ideals (e.g., ideologies, doctrines, and religions) that will never be proven wrong. That all knowledge could be wrong (be fallible), is the only infallible truth. Only social ideals that are based on faith in the fallibility of everything have the potential of forever avoiding the loss of their entire value (q_t).

Therefore, a comprehensive doctrine that satisfies the equations of all of Formulas (1) to (3) must be one that is based on faith in fallibility itself.

Part 5.4 AI and Anti-Data Monopoly Policy

What is worrisome in an era when AI is used in various situations is that AI may become the subject of a "myth of infallibility." If such a myth of infallibility—the implication of which is that AI is justified in limiting individual freedom—were to become commonplace in society, it would be problematic.

We can say that AI, which evolves through deep learning, is in itself a product of trial and error. However, when AI is used in society, don't people take the validity of its responses for granted? If society uses AI to answer a lot of its difficult questions—such as when financial institutions use AI for the process of selecting securities for investment or evaluating prospective borrowers' creditworthiness; when AI is in charge of autonomous driving; or when it determines worker aptitude—and if various decisions are made on the assumption of AI infallibility, human freedom could be severely undermined.[3]

From the concentration of big data and the trend in AI utilization in the real world, the risk of that kind of society developing cannot be denied. The volume of personal data collected by a handful of IT companies, including GAFA (Google, Amazon, Facebook, and Apple), is enormous. If those companies monopolize big data and if AI systems that learn from the monopolized data come to be involved in making decisions that are critical for human society, AI systems could become literally unquestionable.

This in no way means that "AI is absolutely right." Still, such a situation could create a social consensus that AI may be regarded as effectively infallible. Under that consensus, the important life

---

[3] Believing in the infallibility of AI may make human society systemically prone to wrong judgment. An episode at Amazon.com of the United States that was reported in a Reuters article (https://www.reuters.com/article/cbusiness-us-amazon-com-jobs-automation-idCAKCN1MK08G-OCABS) provides a foretaste of that risk. According to the Reuters article, at the beginning of 2017, Amazon abandoned an experimental AI-based recruiting project launched in 2014, after recognizing that the AI tool used in the project discriminated against women when giving scores to job applicants. The Japanese Society for Artificial Intelligence referred to this matter in the "Statement on Machine Learning and Fairness" (https://www.ai-gakkai.or.jp/ai-elsi/archives/948), which was issued on December 10, 2019.

choices of individuals (e.g., academic and working careers, marriage, and location of residence) could come to be determined by AI based on past data, which would mean the deprivation of individual freedom (also known as the right to stupidity, including learning by trial and error).

This is the future vision of a totalitarian system in which AI manages or controls society. In that kind of society, AI's presence would be equivalent to that of a ruler who deprives individuals of their freedom and dictates their actions (although individuals would not be conscious of being dictated to in the case of an AI-controlled society). This would represent the arrival of the kind of dystopia that was described by George Orwell in his novel *1984* (first published in 1949).

If the IT giants' ongoing monopolization of big data continues to advance, the advent of a dystopian future will come closer to reality. What is now occurring in the internet business is similar to the monopolization problem faced by the world economy around the end of the 19th century, when huge, dominant companies wielding power across various industries caused serious economic harm through market distortions. It may be necessary to restore the soundness of the competition environment in the AI market by enforcing anti-monopoly policy concerning the utilization of personal and customer data, just as the United States enforced anti-monopoly policy under the Antitrust Act in the late 19th century through the early 20th century.

It is difficult to prove based on currently available economic theories that enforcing an anti-data monopoly policy is necessary. The act of using data (and thereby promoting AI learning) has positive external effects. The larger the amount of data used, the more valuable additional data becomes—that is, the principle of "economy of scale" applies. Because of the principle of economy of scale, in the data utilization industry, greater efficiency is achieved when a small number of companies have monopolistic (or oligopolistic) market power than when many companies compete with one another. From discussions held from an economics perspective like this, the argument may emerge that applying the "electric power industry model," which would grant data monopoly to IT giants in exchange for imposing a certain degree of government regulation, is appropriate.

A political philosophy premised on "fallibility" is likely to be able to counter that argument and justify the prohibition of data monopoly.

If the positive externalities associated with the use of AI are the main cause for concern, it may be said that allowing companies to gain monopolies is rational. However, the "fallibility" of AI poses another problem. Not only the companies that monopolize data, but also the AI systems that learn from the monopolized data, are fallible. Intelligence that learns from big data can never be absolutely right. The possibility, however small it may be, cannot be denied that the unsophisticated decision-making of humans could deliver better results than the sophisticated decision-making of AI.

In addition, especially when we consider the fallibility of AI and humans in a state of Knightian uncertainty under the veil of ignorance, in which even the probability distribution is unknown, the people are expected to agree to the enforcement of a ban on data access monopolies as a fair policy.

That is because AI's judgment is no different from the stupidity of individual humans in that both could be an appropriate judgment for the survival of individuals (making a probabilistic choice between AI and human judgments is impossible because in a state of Knightian uncertainty, the probability distribution is unknown). Giving a large number of companies' AI systems and many human individuals access to data as sources of useful inputs for decision-making, rather that limiting such access to a handful of companies' AI systems, makes it possible for the largest number of agents to make decisions. As banning monopolies on data access gives the largest number of individuals the largest number of options and heightens our chance of survival (delivers better effects), the people should reasonably agree to the ban as a social contract under the veil of ignorance.

As described above, if the logic of the Rawlsian veil of ignorance is applied, it becomes clear that the people should agree to a ban on data access monopoly as a social contract.

Part 5.5 Fallibility and Freedom

The way for humankind, as fallible and free beings, to make progress, is to do what could be wrong based on possibly wrong knowledge and to learn from the results. Accumulating real-world experiences of success and failure through recurring trial and error and adjusting theories based on reality is a process that is common to both human learning and AI's deep learning. Only when the "freedom to make mistakes" is ensured is learning through that process possible.

In other words, freedom is indispensable to our society because nobody knows all the answers regarding the challenges that our society is facing and the future of our society. While innovations update the system of justice daily and promote the progress of reason, nobody knows what the ultimate destination of the progress is. All that human reason (enhanced by AI) can do is to exercise the right of freedom to engage in trial and error, acquire information on new features, and attempt to comprehend the structure of the universe through approximate calculation. However, the progress that is achieved is always open to fallibility—that is, there is always the risk that the progress may later be proved "wrong."

While believing in the value of the progress in AI-enhanced reason and the "system of justice," we cannot rid ourselves of doubt and trepidation over the possibility that the progress, given such fallibility, could, in fact, be wrong. Rather, it is exactly because of that kind of uncertainty that we humans do not remain content with the status quo but are willing to try something new that goes beyond convention. That is the driving force of innovation. The accumulation of innovations updates human knowledge and promotes progress in the social "system of justice." Therefore, we may say that recognizing the fallibility of ourselves and others is the driving force of the progress of our society. As a condition for enabling individual persons to engage in trial and error and to achieve innovation, there must be a social system that ensures the first principle of Rawlsian justice, that is, basic freedom for

everyone. Ensuring the freedom to engage in trial and error is a prerequisite for societal progress. In other words, ensuring freedom for individuals satisfies the condition for the moral value q_t in Formula (3) to turn positive and continue to increase forever through time under the moderate comprehensive doctrine that we advocate.

Part 5.6 Will Superhumans Eradicate Ordinary Human Beings?

The idea that "everything is fallible" is the only theory that could be an "infallible truth." Only a comprehensive doctrine premised on the principle of fallibility can keep the moral value q_t of the social system positive forever. Therefore, we believe that societal ideals that should be societally maintained must be premised on the principle of fallibility.

That idea applies not only to ordinary human beings but also to superhumans, or human beings whose intellectual power has been enhanced by AI. Let us assume that ordinary human beings and superhumans enhanced by AI and biotechnology have been divided into two separate social classes. In that case, superhumans, too, are aware of their own fallibility. Despite being enhanced by AI, they would understand reality through nothing more than "approximate calculations" but would be unable to "truly" understand "everything." Pattern identification based on deep learning is also a form of approximate calculation using prepared sets of real-world patterns. Superhumans, too, would understand that all intellectual activities represent an accumulation of approximate calculations.

Superhumans who are aware of their own fallibility are expected to create a society that is tolerant of activities that are freely conducted by a great variety of beings (including ordinary human beings). If they are aware of their own fallibility, they are certain to recognize the possibility that innovations brought about by other people (including ordinary human beings) could have a significant impact on themselves. If interactions that could occur between unforeseen innovations are taken into consideration, from a superhumans' point of view, respecting the continued existence of ordinary human beings, rather than wiping them out (or letting them wither into extinction), would be the most beneficial and rational decision for purely selfish reasons.[4]

The vision of a diverse and tolerant society premised on the principle of fallibility is nothing more than what we imagine within the limits of our thinking. One problem for me is this: when thinking

[4]The logic mentioned here applies not only to the relationship between superhumans and ordinary human beings but also to the relationship between superhumans and other animals and plants. Even though intellectual activity may be the exclusive domain of Homo sapiens, by respecting biodiversity, ordinary human beings and superhumans can expect to benefit in various ways, including from resources generated through the activities of the diverse assortment of beings on the planet (e.g., drug ingredients, useful chemical substances, and raw materials). Given this expectation, even if superhumans act entirely selfishly, they are certain to consider respecting biodiversity to be a reasonable decision. This is exactly the same logic as the one applied to the relationship between superhumans and ordinary human beings that was explained in the main text.

about a future society in which co-existence with AI is inevitable, how far will I, a mere ordinary human being, be able to follow the reasoning of AI (and that of superhumans whose intellectual power has been enhanced by AI), which is expected to transcend the current human understanding? Of course, the possibility cannot be ruled out that superhumans, by following some line of reasoning that is beyond the author's understanding, will arrive at the conclusion that ordinary human beings should be exploited or eradicated.

Even so, there is one thing we can say for sure.

At the least, believing in the fallibility of any bleak vision of future society remains an option for us—that is, we can choose to believe that it may be wrong to assume that superhumans will eradicate mankind. In this case, fallibility is another name for hope.

Philosophy of Fallibility and Pragmatism

Regarding the “theory of innovation-driven justice,” discussed from Part 1 to Part 22, Shigeki Uno pointed out its proximity to pragmatism, an important school of modern philosophy (Uno and Kobayashi, 2019). According to Uno (2013), pragmatism is a philosophy that welcomes individuals’ freedom to conduct various experiments on the premise of human fallibility, and democracy is the institution that makes that freedom possible. It is John Dewey who presented the idea that the essence of democracy is protecting the freedom to conduct experiments under the premise of fallibility. In that sense, it may be said that the theory of innovation-driven justice is an idea that is close to Dewey's pragmatism.

The theory of the universe that was advocated by Charles Sanders Peirce, one of the founding fathers of the philosophy of pragmatism, is based on a worldview that is very similar to the theory of “strong isomorphism” (see Uno [2013] and Ito [2006]).

Peirce believes that the universe is comprised of three elements. The first element is “chance (chaos),” and the second is “law.” The universe starts from a state of chaos—a state in which chance is the dominant factor— and then it gradually develops into a state that is governed by laws. Ultimately, when everything has come to be governed by laws, the universe comes to an end. Under Peirce’s theory of the universe, it is the third element, “habit,” that crystalizes the first element, chance, into the second element, law.

According to this theory, chance ultimately comes to be governed by laws, and laws are created through the habit-formation process. The idea that the laws of the universe (laws of physics and chemistry) are created through habit formation may sound nonsensical. However, it makes sense if habit formation is understood to mean the process of learning on the part of observers of the world (humans and artificial intelligence). The habit formation as conceived under the philosophy of pragmatism is a process similar to the identification of features (fixed patterns) by AI systems, particularly through deep learning. If we interpret Peirce’s theory of the universe in that way, it may be said that the theory presents a worldview that is very similar to the theory of strong isomorphism.

The relationship between the habit formation as conceived under the philosophy of pragmatism and deep learning, which has led to the arrival of the theory of strong isomorphism, must be analyzed in detail. By linking the theory of strong isomorphism to the philosophy of pragmatism, it may become possible to conceive a new philosophy suited to the era of AI. Realizing that vision is a challenge for future research.

Reference

- Acemoglu, Daron (2002) “Directed Technical Change” *Review of Economic Studies*, 69 (4), pp. 781-809.

- Arendt, Hannah (1951) *The Origins of Totalitarianism*, Schocken Books

- Harari, Yuval Noah (2015) *Sapiens: A Brief History of Humankind*. Harper

- Harari, Yuval Noah (2016) *Homo Deus: A Brief History of Tomorrow.* Harvill Secker

- Hayek, Friedrich August von (1945) “The Use of Knowledge in the Society” *American Economic Review*, 35 (4), pp. 519-530.

- Hegel, Georg Wilhelm Friedrich (1902) *Lectures on the Philosophy of History.* George Dell and Sons.

- Knight, Frank Hyneman (1921) *Risk, Uncertainty and Profit, series of prize essays on economics*, 31, Hart, Schaffner & Marx; Houghton Mifflin Co.

- Lyotard, Jean-Francois. (1979) The Postmodern Condition: A Report on Knowledge. Les Editions de Minuit

- Orwell, George (1949) *Nineteen Eighty-Four (1984)*, Secker & Warburg

- Pocock, John Greville Agard (1975) *The Machiavellian Moment: Florentine Political Thought and the Atlantic Republican Tradition.* Princeton University Press.

- Rawls, John., Kelly, E. (ed.) (2001) *Justice as Fairness: A Restatement.* Belknap Press

- Rawls, John., (1971) *A Theory of Justice.* Belknap Press

- Sandel, Michael (1996) *Democracy’s Discontent,* Harvard University Press