Toward robot ethics through the ethics of autism

Toward Robot Ethics through *the Ethics of Autism*

Masayoshi Shibata[1]

1. Why must autonomous robots be moral?

1-1. What does autonomy mean for robots?

The aim of this paper is to present an ethical landscape for humans and autonomous robots in the future of a physicalistic world, which will touch mainly on a framework of robot ethics rather than concrete ethical problems possibly caused by recent robot technologies. It seems that we could not find sufficient answers to such ethical problems as occurring to future military robots unless we understand what autonomy in autonomous robots exactly implies for robot ethics. This paper presupposes that this "autonomy" should be understood as "being able to make intentional decisions from internal state, and to doubt and reject any rule", which requires robots to have at least a minimal desire-belief folk psychology. And if any agent has a minimal folk psychology, we would have to say that it potentially has the same "right and duties" as us with a full-fledged folk psychology, because ethics for us would cover any agent as far as it is regarded to have a folk psychology --- even in Daniel C. Dennett's intentional stance (Dennett 1987). We can see the lack of autonomy in this sense in the famous Asimov's Laws cited by George A. Bekey et al. in this volume, which could be interpreted to show the rules any autonomous robots in the future have to obey (Bekey et al. 2010 sec.2).

Strictly speaking, these laws are not truly the ethics for robots at all since, as I will argue later, they do not presuppose that robots have the same "rights and duties" among them, or admit that robots and humans have the same "rights and duties" among them. At best they are merely "design policies" to make robots better tools for humans. It is often suggested that if their contents were appropriately revised, they could be changed into the three rules for electrical appliances to obey, because these rules do not clarify to what extent those robots have autonomy. When autonomy is reduced, Asimov's Laws could be rewritten into "The Three Laws of Electrical Appliances" *mutatis mutandis*, which correspond to e.g. (1) security (not to damage humans), (2) obedience (to work as humans intend them to do), and (3) toughness (not to be broken easily). But full-fledged autonomy would conflict with an absolutely obligatory

[1] Professor of Philosophy, Dean of Faculty of Letters, and Director General of University Library, Kanazawa University.

rule for robots, the second rule of obedience.

How should we treat robots and be treated by them? It depends on what type of beings we think the robots are. As long as the robots we are considering now are autonomous robots rather than mere mechanical tools such as robots working in factories, they are not like pets, animals, unborn babies, or the elderly with heavy dementia to whom we have one-sided "rights and duties". In a word, the ethics of robots we want to investigate are the ethics required for humans and robots to coexist with reciprocal relations, i.e. the same "rights and duties" in a community.

But is it really possible that humans and robots can live together as equal members in a moral community? Due to many differences of basic conditions between them, such as birth (production), death (destruction), cognitive abilities, physical abilities, appearances, reproduction, and so on, it seems implausible that such two groups could comprise the one and same moral community. We can give examples of such fundamental differences concerning ethical issues as follows.

First, robots could have a kind of eternal lives or iterated lives over a long period of time, which are made possible by the production principle of "the same design, the same robot" in a functionalist sense, and easy availability of their parts in our physical world. Their prolonged lives may endanger the common interests of goals and methods in a life plan between humans and robots, and thereby make it difficult to comprise a common moral community. The production principle allows robots to be recreated with their exact physical copies without end in principle, so that they exist with exactly the same minds. As I argue later, our actual world where autonomous robots are possible would be a physical world where the supervenience (at least, global supervenience) relations hold between physical properties and mental ones, which make "the same physical, the same mental" possible. So many robots' minds exactly the same around us may conflict with a traditional concept of a person, i.e. the absolute uniqueness of a person as a member of a moral community, if robots could be persons.

Second, robots seem to be able to erase or implant their memories arbitrarily. For humans, the consistency and traceability of their memories, though not perfect but to a certain degree, is required to constitute their personhood, which robots may lack in a radical sense to nearly destroy robots' personhood. Can we punish a robot for a murder in spite of the complete elimination of his related memories? Or how could we regard a robot's sincere claim of his worthiness because of a disguised heroic memory of a past trifle action? We may have to treat a robot with a change of memories in this radical sense as different members of the community every time he erases or implants an important memory. Although it is unclear whether the psychological continuity theory of

personhood is right, easy changeability of memory in robots will give rise to serious problems about robots' personhood.

Third, robots do not necessarily have the same psychology as humans. As we will see later, sharing our folk psychology including the ability of understanding other minds is essential for making reciprocal relationships with each other, which is a basis of being a member of a moral community. Since psychological states are the results of physical and physiological needs and wants, robots and humans may not have exactly the same psychological states. Although robots also need an energy supply, they do not want to eat bread or drink water. They do not feel hunger or repletion, so they seem to have different attitudes and emotions toward food, which may result in a very different scheme of desire-belief psychology. This difference may be mitigated in a higher functional level, but it is not certain that the folk psychological mechanism will work sufficiently for both humans and robots to form a mutual moral community.

1-2. The Ethics of Neo-Crusoe

Imagine that an agent is living absolutely alone in a closed area of deep space in the galaxy. He is intelligent like us, but does not need any partners to survive nor have any missions to do. He may be a robot or an alien unlike us in some respects. Although he may have a memory of his society or the community that brought him up, he has been alone for a long time since the collapse of his society, and will be alone this way until his death. He is not a member of any community now and will not be so in the future. In other words, he is not actually or possibly a member of any community.

I will call him Neo-Crusoe. I guess people would not envy his life, but he could live as he pleases every day. There are no friends or enemies who interfere with him, or whom he interferes with. In a sense Neo-Crusoe enjoys an absolute loneliness, but what does it mean for such an agent as Neo-Crusoe to be moral? Or what kind of ethics does he need? We will have a short remark about this question from Kantian moral theory and utilitarianism. But my concern here is not to get precise interpretations of Kant's, Bentham's, or Mill's theory. Rather, my point is to shed light on some conditions under which any agents including autonomous robots have to be moral, because ethics or morals do not seem to necessarily exist.

Immanuel Kant requires us to accept the Categorical Imperative, "So act that the maxim of your will could always hold at the same time as a principle in a giving of universal law" (Kant 1996, 164), where "the maxim of will" means "a subjective and practical policy of action". Therefore Kant is demanding that our principle of action could be universalized as everyone's principle, or that it would not involve

self-contradiction when universalized as a law. The maxims that cannot be universalized could not be those that tell us our duties, not because of their contents but because of their formal characteristics. For example, it has been said that the maxim, "break your promise as you like when it becomes inconvenient for you" is unable to be universalized. Why? If promises can always be cancelled arbitrarily by their participants, we cannot rely on them precisely when we want them to be fulfilled. Namely, the maxim would destroy self-frustratingly the foundation of promise itself if it were universalized. Among the constructive conditions which make the promise possible at all, there seems to be a condition that the participants have to fulfill it.

Now I want to ask a question: "Are there any maxims unable to be universalized for Neo-Crusoe?" Please remember that he is not actually or possibly a member of any community. He cannot do any actions that necessarily involve relationships with others, i.e. reciprocal actions. Because he is not a member of any community, cooperation, agreement, betrayal, denial, etc., are action types he is not allowed to do in principle. In other words, any actions he can do in this situation are the ones toward him or the rest of the agentless world. Is there any reason that the maxims guiding such actions cannot be universalized? The answer is No, because there is no standpoint for which an action is still of type A, while he is doing an action of type B, the maxim of which destroys the constructive conditions of actions of type A. In such a case he simply changed his mind to do an action of type B, instead of continuously doing A. In order for his action to be a self-frustrated action of type A, there must be other persons for whom it is still A. Therefore no maxims could be distinguished from each other in universalizability, because there are no other persons except him. As far as the universalizability is concerned, there is no ethical viewpoint allowing us to evaluate the morality of his actions performed in his closed area. We could say that any of his actions is morally neither right nor wrong. Ethics are not necessary and indeed they do not exist in the world of Neo-Crusoe, just because there is no "what one has to do" apart from "what one wants to do". Although it would be required to construct more detailed arguments in order to draw this conclusion from Kant's theory when considering, in particular, his treatment of suicide, I think we could sustain this conclusion independently of any Kantian arguments.

Let's see next what utilitarianism will say about Neo-Crusoe's "what has to be done". Jeremy Bentham's utilitarianism of "the greatest happiness of the greatest numbers is the foundation of morals and legislation" is recast in John Stuart Mill's "the Principle of Utility" as follows: "The creed which accepts as the foundation of morals, Utility, or the Greatest Happiness Principle, holds that actions are right in proportion

as they tend to promote happiness, wrong as they tend to promote the reverse of happiness." (Mill 1969, 212). Because Neo-Crusoe is the only person existing in his world, "the greatest numbers" in the Principle of Utility could only mean "one person", i.e. him alone. Without further arguments, it seems evident that whatever actions he may plan to do, there is no "what he has to do" imposed on him contrary to "what he wants to do" as long as he does not intentionally perform actions spoiling his own happiness. And it is certain that he would not intentionally do actions harmful to his happiness. It does not mean that he is always the best judge of his own *future* happiness. It is sufficient for him to not violate the Principle of Utility that he is the best judge *at the present time* of his own future happiness, as far as this Principle is a guide of his actions.

Of course Neo-Crusoe may accidentally invite unhappy results from his actions because of his cognitive failures or bad performances. He may have a strong desire suddenly to touch a green shining stone beside him or climb a steep mountain in the distance, which may occasionally result in bad outcomes for him. But does it imply that he should not have done it? If certain external causes prevented him from doing that action, he would be disappointed seriously and his happiness would be reduced considerably. Even in a case of his regretting his action because of bad consequences, did his regret have any ethical perspectives? If we say that he did a morally wrong action when he brings an unhappy result only to himself by doing an action involving no other members of the community, there seems to be something peculiar in this judgment. Let us remember again Neo-Crusoe's situation. Even if utilitarian calculation of his happy and unhappy consequences says something about his actions' morality, it is mere calculation without a more basic moral intuition implicitly expressed in the phrase of "greatest number" in the Principle of Utility. This is because the Principle shows up only when agents need to have relations with others. One of the presuppositions of the Principle of Utility is that agents are such creatures that are bound to pursue the greatest happiness. Therefore such a lonely agent as Neo-Crusoe satisfies vacuously the Principle of Utility in all his actions because the actions of agents have to be adjusted to one another only when there are plural agents and they come to be necessarily concerned with others' interests. The Principle of Utility is a guide to this adjustment. We could say it is to mistake the means for the end to evaluate the morality of one's actions in spite of there being no others in the community. The above shows that there is no "what he has to do" distinguished from "what he wants to do" for Neo-Crusoe at least in the utilitarian guidance of his actions.

All I was going to suggest in this section is that there are no ethics for such a being

as Neo-Crusoe. Although the above is not a decisive argument, we could say now whatever action he does is morally neither right nor wrong. For him, what has to be done is nothing other than what is desired to be done. This means that agents have their ethics only if there are other members belonging to the same community who could have the same "duties and rights" among them. I will call this "the community condition" of ethics. This condition is not sufficient but necessary for ethics to come into existence for the world of agents. If there had been no "what one has to do" cut off from "what one wants to do" for an extremely long time even for all members of the same community, I think their world would be morally a best one for all of them. In the sense that for one to be able to do what he wants to do is a freedom in a primitive form, ethics are required only for somehow avoiding collisions among agents' freedom to do "what one wants to do". Viewing the ethics as "deprivation of freedom", it seems to me that the most respectable value in the ethical context is Libertarian Freedom. Here Libertarian Freedom should not to be understood as implying "free will without any cause" in a metaphysical sense, but "free choice without any constraint" in a political sense (Cf. J. Greene and J. Cohen 2004).

2. Natural Conditions for Humans to Be Moral

2-1. A Physical World Where Autonomous Robots Are Possible

What is a physical world where we can make autonomous robots? In spite of a lot of arguments in the contemporary arena of philosophy of mind, let me skip the complicated issues to a minimal version of physicalism because I think that kind of world must be a physicalistic world, i.e. a world where some version of physicalism is true. In fact, robot ethics is one of the most difficult moral problems that will be raised by the essential features of a physicalistic world in the future. Ignoring the details, minimal physicalism consists of the following two assertions:

(1) Any individual is identical to some physical individual (that is, there are no souls or spirits as non-physical individuals).

(2) Any property supervenes on some physical property, even if the identity relation between them does not hold (that is, if physical properties as subvenient properties are the same, mental properties corresponding to them are necessarily the same).

According to minimal physicalism, our world is a world where once the physical facts are fixed, all other facts that are characterized as non-physical are determined. For example, the same type of brain state necessarily corresponds to the same type of psychological state, and a same type of physical movement in the same type of environment

necessarily corresponds to the same type of action. Notoriously, the local supervenience relation does not hold between mental states and brain states, when the former is characterized and classified in folk psychological concepts and terms. But if we take as a subvenient basis a sufficiently large spatiotemporal region of the physical world including brains in question, almost all physicalists would admit that the supervenience relation holds (i.e. the global supervenience). So keeping this reservation in our mind, we could roughly assert of the supervenience between mental and physical properties the following relation: for the realization of a psychological state "I have to go to the airport now" it is sufficient for a type of corresponding brain state to occur. And this relation does not allow that although two brains are physically of the same type, the one is realizing a psychological state "I have to go to the airport now", the other "I want to make an omelet" (Cf. Kim 1993).

But in so far as robots are made technologically from various hard materials rather than neurons or hormones, it is not possible that robots have the same type of brain state as humans. Does it mean that robots cannot have the same type of psychological state or belong to the same moral community as humans? Fortunately the supervenience relation allows *multiple realizations*. That is, the same type of psychological state can be realized by many different kinds of physical states. If you are a reductive physicalist like Jaegwon Kim, you have to read the term "same" as "similar" in the previous sentence, but here we will not go deep into the difference between them, because it matters only in the context of psychological laws. What remains as "the same" in the multiple realizations is a function fulfilled in different ways by different mechanisms of a lower level. Indeed it is a precondition for us to produce artificial intelligence or robots to have a conviction that we could make beings artificially which could act in almost "the same" way as we do, because without it there would be no serious efforts leading to the recent flood of various robots. We have been given "ontological supports" by these multiple realizations every time various functions of humans imitated artificially are extended to new territories.

But robots seem to have one worry. The multiple realizations can be endorsed by a robust argument as far as they are concerned with the functions realized by causal mechanism, but there is room for a lot of controversy concerning qualia, or consciousness as an applied problem of "philosophical zombie" (Cf. Chalmers 1996). For example, if it is true that robots do not feel any pleasure or pain at all in spite of fulfilling the same functions as humans, what kind of justification do we have to regard them as subsumed under the Principle of Utility? Here rational beings without sensations (robots) may seem to give rise to a different problem from one caused by sentient beings

without rationality (animals) with regard to the membership of moral community.

But there is good news for robots in the ethical context. If the actual world is one that allows us to make functionally isomorphic robots to humans, the problem whether robots are "zombie robots" without qualia has the same structure of argument as the philosophical problem of other minds, which seems to be unsolvable as a *purely epistemological* problem. As the question of how to know directly other minds beyond external evidence could lead to skepticism regarding other minds; the question of how to be certain about the existence of qualia or consciousness in robots could not be given any decisive answer. In a nutshell, robots occupy the same *epistemological position* as humans in this regard. But in our context of ethics, it is highly important that we have built up a moral community in spite of the skepticism of other minds. That is, our reason to make others members of the moral community is not the epistemological confirmation of mental states of others, but the practically motivated *ontological decisions*.

Therefore, although it remains a philosophically important question whether functionally isomorphic robots to humans have the "same" qualia or phenomenological consciousness as humans, it does not have any significant impacts on the problem whether robots and humans could make a common moral community. I think the more problematic issue is how to construct a common moral community when robots have superior rather than isomorphic functions to humans.

2-2. Humans in a Physicalistic World

Let us take a brief look at what will happen to humans in such a physicalistic world as makes various functional robots possible. The key word here is "enhancement beyond therapy". My concern is in the situation where the natural conditions *making our community possible* will considerably change by humans' coming to be cyborgs and producing many robots around them, and thereby endanger the "existence conditions" of our usual community *making our usual ethics possible*.

The purpose of enhancement which is becoming a big problem today in the fields of medicine, law, morality, and so on is to reinforce a variety of functions of humans in various ways, and to make humans live for a longer and longer time with the health and strength of youth (finally, to attain perennial youth and immortality). All biological phenomena are determined by physical phenomena in a physicalistic world so that in principle any phenomena could be realized if those are physically realizable. But of course all phenomena of each level are governed by the laws of each level. So it is evident that the possible transfiguration (as enhancement) of humans as biological beings has a limit. I am not sure now, but this limitation might mean for humans one more

step in their evolution from biological existence, who have been changing their protein-based forms, to mechanical existence that will have poured their consciousness into robots. In other words, humans might change into robots together with their minds and consciousness in the remote future. It does not mean that humans will be cyborgs, nor that humans' mind and consciousness is a mere program that could be installed in any suitable hardware, but that humans will have minds in robots' brains as one of multiple realizations and bodies as bases for their experiences in the environment. Although this image needs more detailed stories, I cannot present them here because of my inadequate knowledge about the relations among humans, robots, and their evolutions.

Anyway, keeping that limitation in mind, we will see a couple of imaginable results of our enhancement today. First, when the enhancement goes "beyond therapy", it certainly takes a direction toward the fundamental improvement of the state requiring cures. For example, after giving effective medicines to people suffering from dementia, we will try to reproduce or reorganize the neural circuits in their brains. Also in the case of mental diseases and developmental disorders including ASD (autism spectrum disorders), neuromodulators, such as oxytocin, are being suggested as a possible treatment (Neumann 2008, Insel 2010). Furthermore, if possible, we may choose surgical operations on particular parts of brains once we find the neural causes of those diseases in them someday. Naturally, biomedical treatment will extend to embryos and fetuses through DNA-based diagnostics to prevent mothers from giving birth to babies having such birth defects as Down's syndrome by using genetic technologies (Barnbaum 2008, ch.4, Autism and Genetic Technologies). The goal we will reach from here "beyond therapy" is that every parent will have "more desirable babies", or "perfect babies" who will have such desirable characteristics as higher intelligence and physical abilities than usual, more excellent figures and appearances than usual, a strong will, fine sensibility, honesty, brightness, and so on (Kass 2003, ch.2, Better Children). Normal "imperfect adults" already being in our society are not exceptions in this regard. Everyone would want to transform oneself into a "perfect man/woman" ordinarily by taking biomedical treatments to prevent the decline of muscles, preserve immune systems, overcome lifestyle-related diseases, and improve his/her physical appearance. It must be certain that we would finally aim at the perfect avoidance of aging, i.e. the endless prolongation of a lifetime by making thorough use of advanced genetic technologies.

As a result there will appear completely new "natural conditions", or "survival conditions" humans have never yet experienced. Taking cognitive abilities as an ex-

ample, it is highly probable that everyone will become a brilliant individual or a genius. Certainly such high cognitive abilities are not so stereotyped, but their differences will seem to be restricted within a smaller range. The case is essentially the same with people's figures and appearances, too. Making a caricature of this situation, our world is overflowing with geniuses who are handsome men or beautiful ladies. Although the concept of "perfect humans" does not necessarily mean one and the same set of properties for each person, it is sure that we will have nearly all very similar humans around us. Because, as a Russian novelist once said, though the reasons why people are unhappy are different, the reason why they are happy is identical. In a word, we may be faced with a completely new circumstance in which our concepts of personal uniqueness, endeavor, achievement, superiority to others, or goal and happiness in life will change their meanings considerably. This possibility may appear to some people disgusting, to some worrisome, and to others welcoming.

Herbert L. A. Hart explained a reason why, "given survival as an aim, law and morals should include a specific content" (Hart 1961, 189). The minimum content of the ethics we have now is derived from the natural conditions that are contingently imposed on humans. In other words, our natural conditions require a definite set of rules for us to survive, which constitute our minimal laws and morals, without which we "could not forward the minimum purpose of survival which men have in associating with each other" (ibid.). Hart specifies these natural conditions as follows (ibid. 190ff.):

(i) Human vulnerability. "The common requirements of law and morality consist for the most part not of active services to be rendered but of forbearances, which are usually formulated in negative form as prohibitions". This reflects "the fact men are both occasionally prone to, and normally vulnerable to, bodily attack".

(ii) Approximate equality. "Men differ from each other in physical strength, agility, and even more in intellectual capacity". But "no individual is so much more powerful than others, that he is able, without co-operation, to dominate or subdue them for more than a short period".

(iii) Limited altruism. "Men are not devils dominated by a wish to exterminate each other. …But if men are not devils, neither are they angels; and the fact that they are a mean between these two extremes is something which makes a system of mutual forbearances both necessary and possible".

(iv) Limited resources. "Human beings need food, clothes, and shelter". And "these do not exist at hand in limitless abundance; but scarce, have to be grown or won from nature, or have to be constructed by human toil. These facts alone make indis-

pensable some minimal form of the institution of property …, and the distinctive kind of rule which requires respect for it".

(v) Limited understanding and strength of will. "The facts that make rules respecting persons, property, and promises necessary in social life are simple and their mutual benefits are obvious. …On the other hand, neither understanding of long-term interest, nor the strength or goodness of will, …are shared by all men alike". Therefore "… submission to the system of restraints would be folly if there were no organization for the coercion of those who would then try to obtain the advantages of the system without submitting to its obligation. ….Given this standing danger, what reason demands is voluntary co-operation in a coercive system".

These conditions are at most the contingent ones on which humans have been depending rather than the necessary ones humans have to accept. Therefore, as we have already seen, there is a possibility that these conditions will change considerably in our physicalistic world. At least, robots are going to overcome these natural constraints without difficulty. What type of ethics is needed then? In order to make it clear to a small extent in the next section, let me take one of the conditions that make our current ethics possible in the way as we have them now. Hart's five conditions suggest a reason why we need the ethics we have now. What will be suggested in the ethics of ASD (Autism Spectrum Disorders) in the next section is "understanding of other minds" as one of the conditions upon which the ethics we have now become possible.

3. The Ethics of ASD (Autism Spectrum Disorders)

3-1. Theory of Mind Matters

Although there has been much research on ASD and its cause, including genetic related causes, no definitive answer has yet been found. As is widely known, it is salient that ASD does not show a single symptom, but makes a spectrum (a wide range of continuous syndrome) from the type of delay of spoken language or intellectual deficits to that of Asperger's syndrome, which occasionally show "islets of ability", special talents and abilities reaching to a Nobel Prize class (James, 2006)(1). Here I will argue the possibility of the common ethics between extremely different beings, following mainly a remarkable book, *The Ethics of Autism,* published by Deborah R. Barnbaum in 2008. We will see that if robots and humans could build up a common moral community, robots would have to have at least "theory of mind" abilities, and that if both could build up a common community at all, there should be mutual conditions

under which it is possible. But at the same time, we will be troubled by the fact that we could not easily find the mutual conditions due to potential tremendous differences between their ways of being.

It is for us the most important characteristic of people with ASD that some of them do not seem to be able to recognize intentional states of other people as different from their own states. It is often said that although some autistic people, unlike psychopaths, are not indifferent to others' predicaments once they are told of such situations, they could not see through others' emotional states at all. It seems natural to regard this deficit as a malfunction of the so-called "theory of mind". According to this view, some autistic people have trouble ascribing intentional states to others or falsely ascribe their own states to others, because their theory of mind could not function adequately. In any case, many of the autistic cannot pass the false belief tests, which are now very famous in various contexts (2). Of course the problem is not restricted only to beliefs. For someone to recognize that others have minds is recognizing that others are different persons from him/her, and that they have their own independent mental states, including all kinds of intentional and non-intentional mental states such as desires, preferences, worries, and emotions. In the following discussions, among various types and degrees of ASD, we will focus on the type of ASD with serious problems of "theory of mind".

It is not only "the theory of mind" thesis that purports to explain this unique character of ASD. As Barnbaum explains, we have also "the weak central coherence" thesis and "the weak executive function" thesis, roughly speaking, the former of which seems to present a better explanation of why some people with ASD often adhere to meaningless parts rather than the meaningful whole, and the latter a better explanation of why they are often preoccupied with stereotyped and repetitive motions, each compared with "the theory of mind" thesis. But as Barnbaum says, these three theses do not contradict one another, because each of them is merely "re-describing" the properties of ASD by presupposing the hypothetical cognitive functions from each perspective rather than giving a consistent explanation of the causal mechanism of ASD. In this sense, we could say that "the theory of mind" thesis best re-describes the essential features of some people with ASD, and that theory of mind greatly matters in relationships with others, although it is still unclear how to make brain-based autonomous robots with non-autistic minds.

3-2. ASD and Membership in the Moral Community

What do ethics mean for the people who could not truly understand that others

have their own mental lives or recognize what these lives are? This kind of question could be discussed as a problem of membership in the moral community; what properties are necessary and sufficient for any being to be a "person", a member of the community? In other words, could that type of persons with ASD belong to the same moral community as a non-autistic person? Barnbaum, after examining and accepting with some reservations the arguments from Martha, C. Nussbaum, Thomas S. Scanlon, Derek Parfit, and Robert M. Veatch, rejects clearly the most extreme proposal presented by Piers Benn. The arguments about membership in a moral community logically imply that once the necessary conditions for membership were determined, individuals or groups who do not satisfy them would be expelled from the community. Benn is precisely arguing that some people with ASD should be excluded from the moral community of non-autistic people. According to Barnbaum, Benn's argument makes it a necessary condition for membership in the moral community that "a good human life and well-being" consists in "relations that persons have with other persons" (Barnbaum 2008, 93). But autistic persons of that type fail to satisfy this condition because they could not take "intentional stance" toward others due to a lack of theory of mind. Certainly they are biologically humans, but they are neither *person* in a moral sense, nor located within our moral community. In Benn's terminology, only those who can possess "participant reactive attitudes" can be "proper objects of such attitudes" of others, but some autistic people could not take the "reactive attitudes" to others (Benn 1999 33). Reactive attitudes as Benn understands them are emotions such as anger, frustration, or preference, so it would be absurd if they were directed at objects that cannot have anger or preference. "If a hurricane destroys your house, it does not make sense to get angry at the hurricane, because hurricanes do not get angry themselves" (Barnbaum 2008, 94). Some people with ASD could not take appropriate reactive attitudes because of a lack of theory of mind, even if they are faced with emotional situations. But according to Benn, only those who can have reactive attitudes and be objects of reactive attitudes are members of the moral community. Therefore some of autistic persons (or robots without theory of mind, either?) are not members of the moral community.

What is the reason that Barnbaum rejects Benn's argument? It seems to be essentially a consequentialist one. To cut that type of autistic person off from the non-autistic is to cut the non-autistic off from the autistic at the same time. The result of this is a possible "performance of morally wrong actions", and "an erosion of the moral status of the autistic and non-autistic alike". "The moral standing of anyone is damaged whenever that person affirms that some other human being is disqualified from

moral consideration". Therefore "we should continue to be as inclusive as possible when determining who should count as a member of the moral community, because the costs are so high if we are wrong"(ibid. 102).

But this seems nothing other than a "selfish reason" that we want to avoid evil consequences that might visit us due to the essential reciprocity of moral considerations. Here Barnbaum says it is acceptable. Although her argument for this is very interesting, lastly, it seems evident that what makes the "selfish reason" persuasive is the natural condition that both autistic and non-autistic people belong to the same biological category of human beings, and share so many ordinary interests and lives. By the same token, she seems never to think of including animals such as birds, fish, or livestock among members of the moral community. Of course robots are never occurring to her mind. In other words, what determines the widest range of our moral community is, roughly speaking, the five contingent natural conditions Hart indicated earlier. But, as we will see in the next section, she thinks that these natural conditions cannot provide ethics which are applicable to both that type of autistic and non-autistic people equally, because the differences which divide them are so profound and their worlds are so distinct from each other, even if they abide with the same conditions. If those natural conditions are not sufficient for ethics that could cover both sides, what would be the ground for both to comprise a common moral community? Furthermore, taking into account of the fact that there could not be even mutual natural conditions shared by humans and robots because robots could easily stray far from Hart's natural conditions, unfortunately we would have to say that it is more difficult to find or invent ethics for robots than ethics for ASD.

3-3. ASD and Moral Theories

In order to find ethics which could cover both that type of autistic and non-autistic people equally, Barnbaum asks "what ethical theory is applicable to both?", instead of "what ethical theory is right?". Her strategy means that if no one could know what a true moral theory requires because of a lack of adequate cognitive abilities, that would violate a moral axiom, "Ought implies Can", even if there were such a true theory. In that case, since no one in the moral community could know the distinction between right and wrong actions defined by that theory, its lessons would be impractical for them. Generally speaking, from evidence available, it has been doubted that some persons with ASD have a moral sense, they could understand moral dilemmas, and that they could distinguish moral questions from other questions. It is believed that they might have moral blindness. Namely, it is doubted that there are any practical moral

theories for them. And if the case turned out to be tragic for that type of people with ASD, there would be no moral theory for them to obey by their own choice.

Barnbaum concludes that neither Humean, nor Kantian theories work for some autistic persons because of their mental peculiarities. According to her, the story is the same concerning Jonathan Dancy's moral particularism and W. D. Ross's prima facie duties. The possibility of finding moral theories shared by that type of autistic and non-autistic people is rather low. What peculiarities of ASD would hinder moral theories from being adapted to the autistic? In what follows, we will see only a part of her arguments about Hume and Kant (ibid. 114ff.).

As Barnbaum points out, for Hume, morality is more felt than judged, and morality is determined by sentiment. One particular emotion, sympathy, is the core of Hume's idea of morality. But sympathy or empathy requires one's recognition that others have their own intentional states, and that these states can be different from one's own. This means not only holding a belief about others but also recognizing a belief others have. But it is this barrier that some people with ASD could not overcome because of their lack of theory of mind. We have now two explanations of how we gain access to others' mental states, the first of which, the "theory" theory, asserts that we predict others' mental states in question by adopting "theory of mind" to them, and the other, the simulation theory insists that we make up in our own minds, by simulation, the mental state that we would have if we were in that situation instead of others in question, and then transfer that state to them. Whichever theory will turn out to be true, the problem remains the same for people lacking a theory of mind because they could not use either mechanism. In particular, they could not naturally have sympathy or empathy with others. This deficiency results in moral indifference in Hume's idea of morality. In a word, "without this feeling, an agent would be unable to act rightly or wrongly according to Hume's moral theory" (ibid. 120).

On the other hand, Kant's theory may seem to be applicable to people lacking a theory of mind just because it recommends us to reject emotions like sympathy in order to do morally right actions. In fact, Kant thinks that actions have moral worth only when they are done from duty, and that this is guaranteed by the recognition that this action is the one which duty demands, rather than by emotional motivations such as love, sympathy, or pity. Furthermore, Kant even dismisses these emotions as no help to do morally right actions. Therefore even that type of autistic people, who have a serious problem of sympathy with others (hot methodology), could choose a morally right action by following the recognition of duty in each situation, if they adapt Kant's theory (cold methodology), which thinks much of "rule following" aspects. But accord-

ing to Barnbaum, the case is not so easy. For example, at least some versions of Kant's Categorical Imperative are not workable for that type of autistic people because of a lack of theory of mind. One of those versions says "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means" (Kant 1996b 80). Kant thinks that any member of the moral community does not exist as a mere subjective (instrumental) value for someone else's end, but exists as an objective value in his /her existence itself for which all other beings exist. That is, any of these members is never a means for other beings, but an end for him/herself. But could that type of autistic people treat other people not as a means for some ends, but as ends in themselves? To do this, it is necessary for them to assume that others each have their autonomous point of view, which is nothing other than to recognize that others have their own "ends-means" relations different from mine, i.e. their own intentional attitudes toward the world different from my attitudes. As we have already seen so many times, however, it is extremely difficult for that type of autistic people to attribute intentional states different from his/hers, and, as a result, nearly impossible to understand that others are the starting points of their own intentional states, i.e. "the ends in themselves". In consequence, Kant's moral theory is unable to find accord with the peculiarities of some autistic people, either. Barnbaum concludes "the fact that a Kantian moral theory cannot accommodate the autistic individual may be reason for rejecting Kantian theory, not for excluding the autistic person" (Barnbaum 2008, 130.).

4. Conclusion

4-1. Ethics are not Programs but Attitudes

To resolve ethical dilemmas robots will encounter, Bekey et al. in this volume consider two types of approaches to the problem of programming ethics into robots (Bekey et al. 2010, sec.6). The one is "top down" approaches that take seriously an idea that morality is a set of rules to obey in any circumstances without exceptions, and the other is "bottom up" approaches that try to construct morality through experiences without top-down a priori ethical theories. The former corresponds roughly to GOFAI (Good Old Fashioned AI) programming approaches and the latter neural network approaches. Bekey et al. rightly conclude that both are not sufficient to make morally autonomous robots because of the notorious "frame problem". Roughly speaking, the frame problem arises necessarily when robots without human intuitive cognitive abilities carry out any general directions in the real world on condition that

they have to consider all and only relevant important effects by their actions. "The frame problem is particularly acute for top-down approaches to programming, but persists for bottom-up approaches as well" (ibid.). In consequence, Bekey et al. take a third way, i.e. a hybrid strategy of top-down and bottom-up, namely rule-following and experience. This approach is characterized by them as highly related with "virtue ethics" (3). "This approach understands the teaching of ethics as concerned with development of moral character---one's underlying dispositions or tendencies to act in a given situation in a given role" (ibid.). I hope their third way, elaborated sufficiently, will be successful in handling robots. But I have more fundamental worries about programming or teaching ethics to robots.

In fact, putting aside the frame problem, there are two inherent problems here. The one is whether we could have the *true* moral theory that would be programmed into autonomous robots, and the other is how to apply moral theories in general to real situations. My speculation is that moral properties do not supervene on physical properties even globally, so that there is no objective truth in ethics in the sense of reducibility to truths caught in physical sciences. In consequence, morality exists only deep in the center of belief systems of robots or humans, and there is no direct evidence for any moral theories in our perceptual world. Further, since clues and grounds for moral decisions inevitably bring obscurities to some degree in any contexts, applications of moral theories to real situations are not apt for robots' programs as a set of axioms and derived theorems from them. In this regard, morality resides only in a holistic web of beliefs. I cannot show detailed arguments here due to a lack of space, ethics are neither any rigid rule with clear "applicability conditions", nor empirical truth acquired inductively from experiences, but merely attitudes of each belief system toward other belief systems. But nobody knows how to install moral attitudes into robots' belief systems.

4-2. An Ethical Landscape in the Future

But there is a harder problem in robot ethics than what has been discussed before. It is how to build a mutual moral community of robots and humans whose "existence conditions" are extremely different. Unfortunately, I cannot give any decisive answer to this problem now, or even to what robot ethics or the ethics of autism would be like in this concern. But let me suggest something to give the answers in the near future.

First, as the situation of our Neo-Crusoe suggests, ethics have no meaning unless there is a community where one has equal "rights and duties" with others (the community condition). And it is a result of adjustment of interests and actions of mem-

bers within such a community that "what one has to do" arises with a different content from that of "what one wants to do". Therefore, from the point of view of "ethics as deprivation of freedom", it is desirable that ethics hold "what the members of community want to do", i.e. their freedom of actions, in high regard as far as ethics can.

Second, it is due to several contingent natural conditions as Hart points out that our present ethics have their contents as they have now. But these natural conditions are really those of a physicalistic world, when we regard them as preconditions for the emergence of autonomous robots in the future. It seems that these conditions are going to be more and more "improved", that have been the source of our important values for humans, if we make use of new "natural conditions" possible in our physicalistic world. The situations we have experienced until now such as "differences by chance", or "uniqueness in each person" would be rendered increasingly stereotyped and monotonous. Here the fundamental conditions of "approximate equality", "limited resources", etc. in humans may lose an important role to regulate the contents of our ethics and laws, and, instead of those, "perfect equality", "unlimited resources", etc. may change the meaning of life in humans, and thereby the meaning of ethics of humans, too. But I cannot see through the results of this change now. On the other hand, the "natural conditions" Hart pointed out are not serious ones for robots. "Approximate equality" would lose the role of determining the contents of robot ethics too, because they may vary considerably one another in abilities, strengths, or life spans. What would be like the same "rights and duties" shared by humans and robots, the latter of whom could have at least the same mental abilities as the highest ones of "enhanced" humans, "less vulnerability" and more durability than humans, and semi-eternal "life"?

Third, as the peculiarities of ASD show, in order to be a member of a moral community, the most important ability robots need to have is one afforded by "theory of mind", which is, generally speaking, included in folk psychology of humans. Certainly, the role of theory of mind seems to be a little exaggerated in Barnbaum's arguments, because it sounds as though theory of mind alone makes recognition of other minds possible and fundamentally builds up moral attitudes. But if robots did not have a folk psychological mechanism including theory of mind as its core at all, they would not stand in genuine reciprocal relations with other robots or humans, because the robots would not have "other minds" as targets of their moral considerations. The profound difference between two worlds of some of autistic and non-autistic people indicates anticipatorily how extremely heterogeneous "existence conditions" are among robots and humans. Barnbaum could not give an answer to the problem of what contents the

ethics would have that could treat those two worlds morally equally. What she showed is that non-autistic people should not exclude that type of autistic from the moral community, however difficult it is to find ethics covering the two worlds. But, although her conclusion is intuitively right, her argument is not effective in regard to robots, because her argument tacitly depends on the "approximate sameness" of natural conditions of the members, which cannot be expected to hold among robots and humans.

Finally, the last point I have arrived at is that we should make artificially anew a moral system which has the following characteristics rather than look for ethics depending on some natural conditions as Hart pointed out. The new system would include robots, humans, autistic people, non-autistic people, and all other groups of peculiar beings, in so far as they have minimal folk psychological understandings of others. And these folk psychological understandings would be guaranteed by the attitudes that respect others as independent moral agents and by the recognition that others have their own independent minds and interests. Ironically enough, the core of the new moral system in this robot-century would be Mill's famous principle of "harm to others", which urges that we are permitted to do anything unless it does harm to other moral agents, whatever purposes, desires, intentions, feelings, or preferences we have. Of course we have to read this principle as demanding every moral agent to respect all other agents' freedom to act in every context as far as he/she can.

Notes

(1) According to the fourth edition of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (DSM-Ⅳ), a diagnosis of autism requires at least two signs from A, one sign each from B and C, and at least six signs overall as shown below.

A. Qualitative impairments in reciprocal social interaction as manifested by at least two of the following:
    1. impairment in multiple non-verbal behaviors such as eye-to-eye gaze and facial expression
    2. failure to develop peer relationships appropriate to developmental level
    3. lack of spontaneous seeking to share interests or enjoyments with others
    4. lack of social or emotional reciprocity

B. Qualitative impairments in communication
    1. delay, or lack of development of spoken language
    2. impairment in the ability to initiate or sustain conversation despite ade-

quate speech

3. stereotyped and repetitive, or idiosyncratic use of language
4. lack of varied spontaneous pretend play or social imitative play appropriate to developmental level

C. Restricted, repetitive, and stereotyped patterns of behavior, interests, or activities

1. preoccupation with one or more patterns of interest, with abnormal intensity or focus
2. compulsive adherence to nonfunctional routines or rituals
3. stereotyped or repetitive motor mechanism
4. persistent preoccupation with parts of objects

(2) We have several scenarios making up the false belief tests. As Barnbaum puts it, according to the "Sally and Anne Test", children are asked to consider the following story (Barnbaum 2008, 22). Sally and Ann, often represented by puppets, play with a marble, which they put in one place, e.g. a basket. Sally then leaves the room, and Ann moves the marble from the basket somewhere else e.g. into a box before Sally comes back. After observing this, the test subject is asked, "Where will Sally look for her marble?" or, in some cases, "Where will Sally think the marble is?" (Baron-Cohen 1995, 70f.). Children with an intact theory of mind give the correct answer. On the other hand, autistic children often say, "Sally will look for it in the basket" because of their failure to understand that someone in Sally's position may have a false belief, while they have a true belief.

(3) Interestingly, Barnbaum also suggested a possibility of virtue ethics as a moral theory covering autistic and non-autistic people in a personal conversation with me at International Conference on Social Brain: Autism and Neuroethics, Kanazawa, Japan, on Mar. 24, 2010. But the prospects of virtue ethics seem to be uncertain in the case of autism, too.

References

Asimov, I. (2000). *I, Robot*, Oxford University Press.

Barnbaum, D. R. (2008) *The Ethics of Autism*, Indiana University Press.

Baron-Cohen, S. (1995) *Mindblindness*, The MIT Press.

Bekey, G. A., P. Lin. and K. Abney. (2010) Ethical Implications of Intelligent Robots, in this volume.

Benn, P. (1999). Freedom, Resentment, and the Psychopath, *Philosophy, Psychiatry, & Psychology* **6**(1), pp. 29-39.

Goldman, A. I. (2006). *Simulating Minds*, Oxford University Press.

Chalmers, D. J. (1996). *The Conscious Mind*, Oxford University Press.

Dennett, D. C. (1987). *Intentional Stance*, The MIT Peress.

Garland B.(ed.) (2004). *Neuroscience and the Law*, Dana Press, New York.

Greene, J. and J. Cohen. (2004). For the Law, Neuroscience Changes Nothing and Everything, in Zeki et al. (2004)

Hart, H. L. A. (1961). The Concept of law, Oxford Unigversity Press.

Insel, TR. (2010). The challenge of translation in social neuroscience: a review of oxytocin, vasopressin, and affiliative behavior, *Neuron*, Mar 25;65(6):768-79.

Ioan James (2006). Asperger's Syndrome and High Achievement, Jessica Kingsley Publishers Ltd.

Kant, I. (1996a). Critique of Practical Reason, *Practical Philosophy*, *The Cambridge Edition of the Works of Immanuel Kant*, pp.153-271.

--------- (1996b). Groundwork of the Metaphysics of Morals, *Practical Philosophy*, *The Cambridge Edition of the Works of Immanuel Kant*, pp.37-108.

Kass, L. R. (2003). *Beyond Therapy: Biotechnology and the Pursuit of Happiness,* HarperCollins.

Kennett, J. (2002). Autism, Empathy and Moral Agency, *The Philosophical Quarterly* **52**(208), pp.340-357.

Kim, J. (1993). *Supervenience and Mind*, Cambridge University Press.

Mill, J. S. (1969). Utilitarianism, *Collected Works of John Stuart Mill*, Vol. X, University of Toronto Press, pp.202-259.

Neumann, ID. (2008). Brain oxytocin: a key regulator of emotional and social behaviours in both females and males, *Neuroendocrinol.* Jun;20(6):858-65

Shibata, M. (2001) *Minds in Robots*, Kodansya, (in Japanese).

Zeki, S. and O. Goodenough (eds.). (2004). *Law and the Brain*, Oxford University Press.

This work was supported by Grant-in-Aid for Scientific Research (B): 19320001, Japan and RISTEX, Japan Science and Technology Agency.