人間がロボットと共生する日
ロボットの心から人類の道徳まで

平成 29 年 7 月 15 日
金沢大学副学長（教育・法科大学院強化担当理事）
柴田正良

１．ロボットは、三人称的な「心」（心理学的な機能）以上に、一人称的な「心」（意識・クオリア）を持ちうるか
２．物理主義的世界の中での一人称的「心」の位置
３．人間と共生するロボットは、一人称的な「心」の持ち主となるだろう
４．ロボットやペットもメンバーとした新たな道徳共同体の構築

本発表で私は、以上の４つの話題を一つの流れとしてお話ししたい。まず、最初に、ロボット、つまり生物由来ではなく、工学・電子的な人工制作物である機械が持ちうる「心」とは何でありうるかを考えてみよう。

１．ロボットは、一人称的な「心」（意識・クオリア）を持ちうるか。

例えば、信号機の色が赤になれば、運転手は普通ブレーキを踏んで車をとめる。このとき、「赤」の知覚が心に送られ、信号や交通ルールについての知識と総合されて、「車をとめねばならない」という判断を生み、「ブレーキを踏む」という意図的行為がなされる。しかし、これらの一連の心理学的な心の「働き」は、物理的な存在とは次元を異にする、何か特別の、神秘的な実在，例えば霊魂や魂を必要とするわけではない。というのも、この心理学的な心の働きはすべて、物理的、化学的、生理学的な因果的事象の連鎖に還元されるからである。信号機から網膜への電磁波の到達、視神経を流れる電気パルス、脳の視覚領での神経興奮、大脳内での情報処理、脚の筋肉の収縮と伸展を引き起こす遠心性の神経伝達、等々。実は、これらが生ずれば、心理学的なレベルでの「認識」や「判断」や「意図」が、それらに加えて生ずる必要はない。前者が生ずれば後者も生じざるをえない。これは、自然法則的な実現関係だ。したがって、これらの生理学的な機能連関が工学的なレベルで実現されるなら、その工学的な機能を持ったロボットは、それに対応する心理学的な機能、「心」を持ったことになる。

しかし、残念ながら、話はこれで終わらない。機能には還元できな「心」がある。T. ネーゲルが言ったように、「あなたがもしコウモリであったとしたら世界がどんな風に感じられるか」を、あなたは自らコウモリにならない

限り決して知ることはない。それどころか、同じ信号機の赤色を、車の同乗者がどう感じているかを、あなたは決して知りえない。各経験主体が一人称的視点から経験する内容は、その当人自身にしか知りえないのだ。したがって、この一人称の視点からしか知りえない感覚、感じ、つまりクオリアとその総体を含む意識は、他人が窺い知ることのできない、不可侵の領域を形作る。それゆえ、心理学的なレベルで人間とまったく同じロボットに、あなたと同じクオリアや意識が生じているかどうかは、どうやっても分からない。つまりロボットはいかにうまく作っても、一人称的な「心」をもつことまでは保証できない。これを鮮やかに示した、D. チャルマーズの「逆転クオリア」、「欠如クオリア」、「哲学的ゾンビ」の議論も、時間があれば紹介したい。

２　物理主義的世界の中での一人称的「心」の位置

いま、現実世界とは別に、論理的に矛盾のない、様々な法則や個体や性質や事態の組み合わせが可能であるとしよう。それらは「可能世界」と呼ばれるが、その数は無限である。SF 小説や反事実的な想定では、実はわれわれは、こうした可能世界の一つを語っているのだ。例えば、アインシュタインのとは別の物理学が成立している可能世界、あるいはトランプがアメリカ大統領になっていない可能世界，等々。

さて、では、われわれの現実世界はどのような可能世界なのだろうか？現在のテーマに即して言えば、それは、プラトンが思い描いたような、心が第一の存在で、物質は心に依存してしか存在しない、精神主義的世界であろうか。それともデカルトが語ったような、心と物質が相互に独立して存在する、二元論的世界であろうか。私の立場からすれば、現実世界は、プラトンとは逆に、物質が第一の存在であり、心は物質に依存してしか存在しない、物理主義的な世界である。

そこで、現実世界が物理主義的世界だとすると、先ほどの心理学的な「心」、三人称的な「心」は、物理的な事象に法則的にくくりつけられている。つまり、脳の物理的状態が決まれば、それがどのような心理学的状態であるかも決まる。しかし、問題の一人称的な「心」、クオリアや意識はどうであろうか。存在論的には，この種の物理主義的可能世界には２タイプあり、その一つでは意識・クオリアは物理的性質からまったく何の拘束も受けずに、「根無し草」のように、勝手に物理的世界に漂っている。しかし、もう一方のタイプでは、両者の間にスーパーヴィーニエンス(SV)と呼ばれる性質間の依存関係があり、意識・クオリアは物理的世界の状態に依存して生起する。現実世界がどちらのタイプの可能世界であるかに関して、決定的な論証は与えられないが、日常のあらゆる場面からして、SV の成り立つ可能世界だと

考える多くの合理的根拠がある。つまり、この現実世界では、基盤となる物理・心理学的状態が決まれば、それに対応する意識・クオリアの状態も決まるであろう。われわれは、心理学的機能主体の奥に一人称的な経験主体が潜んでいるとふつう想定する。現実世界における SV を暗に信じて･･･

３．人間と共生するロボットは、一人称的な「心」をもつ

人間と共生するロボットは、まず、「自分の欲求と信念から自律的に行為する」ことが可能でなければならない。これは、ロボットがいわゆる素朴心理学的な機能を果たすと同時に、素朴心理学的に説明される対象でもあることを意味する。結局のところ、人間が求める共生の対象は、自律性を欠いた、単なる受動的な機械人形ではないであろう。実際、われわれと暮らすペットでさえ、信念や欲求の種類がどれほど貧弱であろうと、自律的な行為者には違いない。彼らが単なる置物や人形でないのは、われわれに完全には予測できない彼らの行動にある。その予測不可能性、未知なる領域の存在が、彼らと人間が共生するための必要条件である。

その予測不可能性は、実はそれ以上のものを暗示している。素朴心理学的説明の不正確さをたとえ物理学的、生理学的説明が克服したとしても、それが描き出すのは行為者（人間やペット）の三人称的な内面（脳や心理）であって、彼らの一人称的な内面（意識やクオリアの経験）ではない。そして、われわれが彼らを自律した行為者として認めるのは、実は、彼らの行動の日常レベルでの予測不可能性の奥にある、一人称的な不可侵の内面世界の故である。行為者が一人称的な主体として何をどのように経験しているかを、他の存在者が文字通りに知ることはできない。これこそが、彼らの自律性の究極の根拠である。

４．ロボットやペットもメンバーとした新たな道徳共同体の構築

こうして、人間と共生するロボットが「一人称的視点をもつ」と人間から認められたなら、それはすでに、ロボットが道徳的な行為主体となったことを意味する。というのも、「他人が侵犯しえない視点をもつ」ということは、他の誰もが肩代わりしえない「代替不可能性をもつ」ということであり、それは，究極的には彼のみが彼の行為の責めを負うということに他ならないからである。行為のこの責任の発生が、いわゆる古典的な自由意志の概念には由来しないことに注意しよう。われわれの世界は物理主義的世界であった。そこでは物理法則が支配し、それを逸脱する「原因なき行為」などというものは存在しない。したがって、この場合も、他の場合と同様、行為の責任は行為の合理性から発生する。行為者の合理性が高ければ高いほど、彼は

酌量の余地なく100%の責任を負うことになるだろう。

われわれは、その議論を、今年、2017年7月にロンドンで開催される国際認知学会（CogSci 2017）において「道徳的行為者としてのロボット：哲学と経験からのアプローチ」"Robot as Moral Agent: A Philosophical and Empirical Approach"と題して発表する。これは、私たちの科研費「個性を持つロボットの制作による＜心と社会＞の哲学」（15H03151）の成果の一つである（http://siva.w3.kanazawa-u.ac.jp/index.html)。

さて、こうしたロボットは、われわれの道徳共同体の中でどのような位置を占めるのであろうか。一人称的視点をもった自律的なロボットの制作は、法律で禁止しても止められないであろう。いずれ彼らは、われわれの周囲に満ちあふれる。彼らは、たいていは人間より知力も体力も優れている。彼らと大人の人間の関係は、現在の大人と幼児、もしくは大人と認知症の老人との関係に似ているかもしれない。その場合、ロボットをわれわれの道徳的な仲間として迎え入れることは、新たな倫理道徳のシステムを創り上げることを意味するだろう。

ここでは詳しい論証は省くが、私はここで、＜政治哲学的な意味での自由主義＞をまずは道徳システムの創作原理として提案したい。それは、「他者危害の原則」を侵さないいかなる行為も倫理的には許される、とする自由至上主義だ。これは、このままでは内容を欠いた倫理道徳であるが、われわれはまず、これを出発点の大枠として選ぶほかはないように思われる。他者とは誰のことか？　われわれと他者が構成する道徳共同体には誰が含まれるのか？　確実にそこには、知性ある良き心の異星人やロボットも含まれるであろう。そして、その周囲には、認知症の老人や幼児、あるいは動物やペットも程度を異にしながら位置づけられるであろう。「他者危害の原則」は、道徳的行為者のすべての正規メンバーに、原則として、「同じ権利と義務」を要求する。しかし、この「他者危害の原則」以外にどのような実質的な道徳原理を新たに採用すべきは、人類の今後の課題である。そして、それに基づく道徳共同体がどのような姿になるのかも･･･

I would like to talk a story about robots and humans in the future society following the 4 steps below.

1. Can robots have the first person “mind”, consciousness and qualia, in addition to the third person “mind”, psychological functions?
2. The place of the first person “mind” in the physicalistic world.
3. Robots coexisting humans would be ones with the first person “mind”.
4. Constructing a moral community including robots and pets as its

members.

1. When we see “red” of a traffic signal, perception of “red” is sent to mind, and with the knowledge of traffic rules the driver makes a decision that he must stop the car. But a special kind of entity such as soul or sprit is not necessary for a set of these psychological occurrences. These psychological events are reduced to physical, chemical or biophysical causal events. In fact when these causal events occur, those psychological events have already occurred without any further events. Therefore if these biophysical functional relations are realized in mechanical level, robots with these functions come to have psychological mind.

But the story does not come to end. As T. Nagel told us, you could not know what it would be like to be a bat, unless you yourself were a bat. But what is even worse, you could never know how your friend in the car feel “red” of the same traffic light. Because only subject of each experience can know the contents of their own experiences. This territory of first person perspective of qualia and consciousness constitutes an inviolable region from others’ point ow view. Therefore it is not certain that even robots with psychological functions indiscernible from humans can have the same qualia and consciousness as humans. Namely, we cannot ascertain that robots have first person “mind”, however ingeniously we make them.

2. Suppose there are many possible worlds where various combinations of different laws, particulars, properties, and state of affairs hold. Then what type of possible world is our actual world? I would like to say here that our world is neither a Platonic possible world where the mind is a prior being and the physical exists only depending on the mind, nor a Cartesian possible world where the mind and physical exist independently to each other. Our world is, contrary to Platonic possible worlds, one of physicalistic worlds where the physical is a prior being and the mind exists only depending on the physical.

If our actual world is a physicalistic world, where are the first person minds located in our world? Of course once physical or physiological states are determined, the corresponded psychological states and the

third person minds are fixed. But if the supervenience relation, that is a depending relation between properties, does not hold in our world, the first person mind of qualia and consciousness is rootless. It does not belong to any place in our world.
We cannot have any decisive arguments, but we usually suppose that the supervenience relation holds and there is a first person mind or perspective behind the third person mind whenever we meet the latter one. So we are tacitly assuming in the ordinary life that once psychological state is determined, the corresponding state of qualia and consciousness is fixed.

3. Robots coexisting with humans must have an ability of acting autonomously based on their own beliefs and desires. This means that they perform folk psychological functions and they are explainable from folk psychological point of view. In fact humans would not want to live with mere passive and not-autonomous mechanical dolls. Actually even pets around us must have their own beliefs and desires, however shallow and poor they are. The reason they are not dolls but coexisting beings with us is that we cannot predict their actions perfectly well. And this unpredictability or the existence of territory we cannot know perfectly is a necessary condition for them to coexist with humans.
But something more is suggested by this condition. As we have already seen, even if folk psychological explanations are replaced with more accurate physical or physiological explanations, we cannot get their first person inner world of qualia and consciousness, but only their third person inner world of brain and psychology. But we accept them as autonomous agents just because of their inviolable inner world. It is an ultimate reason of their autonomy that they have their own first person perspectives other agents cannot know from the outside.

4. When we admit that robots have their own first perspectives, they are already “moral agents”. Because to have “perspectives that others cannot violate” is nothing other than to have “non-substitutability that others cannot cover”. This means that only he, not others, is responsible to his actions. But at the same time, notice that this emergence of responsibility does not derive from the old concept of

“free will”. Our world is a physicalistic one, where physical laws dominate, so our world cannot give any space to the free will deviating from those laws. Responsibility of action comes from rationality of action.

We will have a presentation of the above argument at CogSci 2017 to be held in London, which has the title “Robot as Moral Agent: A Philosophical and Empirical Approach”. This a result of our research grant from MEXT, “Philosophy of mind and society through making robots with personality” (15H03151): http://siva.w3.kanazawa-u.ac.jp/index.html

Finally, where in our moral community do such robots occupy their seat? It would be in vain if we prohibited people from making robots with first person perspective. At some future day there will be over with such robots around us. They are usually superior to us in respects of intelligence, physical strength, and so on. They might have almost the same relation to adult humans, as present adults to infants or demented elderly. If we want to accept robots as members of our moral community, we will have to construct a new moral system.

Without detailed arguments, I would like to propose here “Libertarianism in a political-philosophical sense” as a first step to building such a new moral system. This Libertarianism says that anything not-violating “Principle of harm to others” is morally permissible. “Principle of harm to others” demands “equal right and equal duty” from every regular member of moral community. But who is others? Who is included to our new moral community? I am sure that aliens with intelligence and warm heart would be included, besides robots. And maybe pets and animals too. Now we have the task of considering what moral principles, in addition to “Principle of harm to others”, should be accepted in the future, and what kind of moral community should be constructed.