1/FEB/2023

哲学倫理学特殊Ⅱ Ⅰ期末課題

人間が必要とする人工知能の条件

所属　　文学部哲学専攻 3 年
学籍番号　　12000555
氏名　　荒金　彰

[1]　AI が信頼できない理由としてあげることができるものと、それに対する応答。

[関連する設問]　2.「機械学習に基づく AI システムは不透明であり、従って信頼できない」という主張について論じてください。「信頼できる AI」の条件とは何でしょうか。
[参考資料]：哲学倫理学特殊Ⅱ Ⅰ授業内で使用されたスライド、および授業内での自分のコメント。

（１）　AI はそもそも信頼の対象ではない（**道徳的行為者性を持たない**）から。
- a. 反例：①道徳的行為者性を持たない盲導犬を、彼が道案内をしてくれることについて、我々は信頼する。②また、道徳的行為者性を持たない梯子を、それが私の梯子を登る間に私の体重を支えてくれることについて、我々は信頼する。ゆえに、AI が道徳的行為者性を持たないことは、AI が信頼の対象ではないことの理由にはならない。

（２）　AI のシステムは不透明（**ブラックボックス**）であるから。
- a. 反例：①その行動原理やメカニズムが不透明である盲導犬を、彼が道案内をしてくれることについて、我々は信頼する。②また、その行動原理やメカニズムが不透明である人間を、彼が業務を遂行してくれることについて、我々は信頼する。ゆえに、AI がブラックボックスであることは、AI が信頼に値しない理由ではない。

（３）　AI に**任される仕事の難易度が高く**、かつ、その仕事を引き受けるだけの**資質・能力を AI が満たしていない**ために、その仕事の達成に失敗するであろうから。
- a. AI に任せる仕事を低リスクなものにするか、あるいは AI の資質を向上させるかのいずれかの方法によってのみ、この問題は解決される。例えば国家の政策や共同体の議決について、その判断が難易度の高いものであり、それらの事柄の判断に関して熟練した人間のみが行いうるような案件であるならば、そうである間は AI にその判断を任せないのがよい。

（４）　AI の**学習データの偏り**によって、その判断にバイアスがかかっていると言え、その判断は適切ではないから。
- a. ①その判断にかかるバイアスが適切である間は、問題がない（ちょうど、人間の視覚は錯視に惑わされるが、状況が通常である限りにおいて、人間の視覚はものを見ることについて信頼されるように）。②しかしその判断にかかるバイアスが不適切である間は、学習データを偏らないものにすることによって、あるいは AI のバイアスが問題となる場合にのみ人間が判断することによって、問題が解決される。

（５）　AI は生み出されてから間もなく、潜在的に重大な欠陥を抱えたままそれが露見していない可能性があるから。（**十分に時間が経たなければ露見しない問題があるという懸念**）
- a. これは、時間の経過を待つことによってしか解決されない。新技術を、単にそれが新しいからという理由で警戒する人々は、その技術が立派に普及するのを見て、その技術が時間の経過によって新しいものでなくなるのでなければ、警戒を解かないであろう。

[2]　**人間に必要とされる人工知能は、人間の利益を目的とするものでなければならず、他のもの（例えば技術それ自体）を目的とするものであってはならない。**

[関連する設問]　4.強い AI には X が必要である(しかし、現状の AI には X が欠けている) の X に入るものを挙げて、なぜそう考えるのか説明してください。「AI」と呼びうるシステムやソフトウェアの具体例に基づいて論じてください。「強い AI」の代わりに、「知性をもつ AI」「言葉を理解する AI」などとしてもかまいません。
[参考資料] : なし　「技術が何かを目的とする」という考えは、プラトン『国家』によって得た。

あるもの X が何者であるかを理解するには二つの方法がある。第一の方法は、理解したいそのもの X だけを単独で取り上げて分析する方法であり、例えば人体や機械など考察の対象になるものそれ自体を解剖なり分解なりして調べ上げる方法がこれに当たる。第二の方法は、理解したいそのものが周囲のものとどのような関係において在るのかを調べる方法であり、例えば人体や機械が自然環境や外界においてどのように振る舞うのかを調べ上げる方法がこれに当たる。本稿では後者第二の道を辿る。

社会という集合には、様々な部分が含まれている。例えば社会は、人間、制度、食料、建物、道路、道具、書物、天候、といったものを含んでいる。これら様々なものが複雑なネットワークをなして、全体のなかでそれぞれの位置を占めていると考えられる。人工知能という部分は、社会という全体のなかで、とりわけ人間に対して、どのような位置を占めるだろうか。

質問者　「人工知能は、一種の技術であろうか、それとも技術が配慮するところのものであろうか。」
回答者　「その問いは、どういう意味のことを言っているのですか。」
質問者　「医術は、人間の健康を配慮する技術である。このとき人間の健康は、医術によって配慮されるところのものである。また、算術は、人間の数に関する認識の正しさを配慮する技術である。このとき人間の数に関する認識の正しさは、算術によって配慮されるところのものである。」
回答者　「はい。」
質問者　「これらのことから、少なくとも、技術が達成しようとするものと、技術そのものは異なる、ということは我々の間で同意されるのではないか。」
回答者　「はい、そのように同意できます。」
質問者　「それでは、技術は常に、何らかのものを達成しようとする目的を持っていることにも同意できるだろうか。それとも君は、目的を欠いた何らかの技術を挙げることができるか。」
回答者　「いいえ、挙げることはできません。造船術は船を作る目的を持っており、狩猟術は動物を狩るという目的持っており、また芸術家の技術は美術作品を生み出す目的を持っていますから。」

質問者　「技術が何らかの目的を持つ限り技術であるならば、達成しようとする目的に応じて、技術には絶えず変更が加えられるのではないか。」
回答者　「それはどのような意味ですか。」
質問者　「技術が常に目的を持つならば、その目的を達成しない技術は劣ったものであり、不要なものであって、なるべくこれを避けなければならない。また、その目的を達成する技術は優れたものであり、必要なものであって、なるべくこれを求めなければならない。このように、技術が何かを避けたり求めたりする動きを、技術に変更が加えられ技術が進歩すると言う。」
質問者　「例えば、人間の健康を配慮しない医術は、真正の意味での医術ではない。医術が患者の健康（利益）を考慮するものではなくなり、医術そのものの利益や、医術を司る医者の利益を考慮するものとなるとき、それは医術とは呼べなくなるのではないか。」
回答者　「医者を利する技術は、患者を利する医術とは別の名前で呼ばれ、区別されなければなりません。」

質問者　「ここに来て、人工知能は一種の技術であって、技術が配慮するところのものではないことに君は同意できるか。」
回答者　「はい、そのように同意できます。」

質問者　「ところで、あらゆる技術が配慮するところのものとは、一種の人間の利益ではないか。」

回答者　「はい、先に言われたことによれば、医術は健康という人間の利益を、算術は数に関する正しい認識という人間の利益を、それぞれ配慮するものでしたし、他の技術についても同様のことが言えましょうから。」
質問者　「ならば、人工知能が一種の技術である以上、それが配慮するところのものも、人間にとっての諸々の利益のうちどれかでなければならないのではないか。」
回答者　「はい、そうでなければなりません。」

質問者　「また、先に言われたように、技術が何らかの目的を持ち、その目的に資するように技術には変更が加えられるものであるならば、人工知能が一種である技術である以上、人工知能のありかたもその目的に資するように変更されなければならないのではないか。」
回答者　「はい、先の議論によればそうなると思われます。」

質問者　「さて、人間の便益を達成するものであるなら、なんでもこれを求めるべきであって、方法は問われないのではないか。」
回答者　「はい。」
質問者　「また、人間の便益を害するものであるなら、なんでもこれを避けるべきであって、方法は問われないのではないか。」
回答者　「はい。」
質問者　「ただし、ここで言われた『人間の利益』『人間の便益』といった言葉は、文字通りの真正の意味での利益や便益であって、決して短期的には有益だが長期的には有害であるような”利益”、一部の人間を益し一部の人間を害するような”利益”、人間の住む環境を害すゆえに近い将来人間にもその害が及ぶような”利益”といった、文字面だけの偽りの利益を指すものではないのだね。そのようなものを我々は利益とは呼ばないのだ。」
回答者　「はい、真正の利益について、以上のことが語られました。」

質問者　「また、人間の利益が具体的に何であるかは、まだ以上の議論だけでは明らかになっていないのではないか。つまり、人工知能が信頼に値すること、人工知能が人間と見分けがつかないほど流暢な会話をなすこと、言葉を理解すること、優れた知性を持つこと、意図を持つこと、生物的有機体であること、主体性を持つこと、これらが人間の利益であることは、確実なことは一切論じられていないのではないか。」
回答者　「それらのことは論じられませんでした。」
質問者　「それでは、それらのものが人間の利益である限りにおいて、それらのものを持つことが人工知能に求められるが、それらのものが人間に有害である限りは、それらのものを持たないことが人工知能に求められるのではないか。少なくとも、今の議論の流れ、つまり人工知能が人間の利益を目的とする技術である間は。」
回答者　「はい、そのようになります。」

質問者　「それではまた同様に、人工知能を限りなく人間に近づけることは、人間の便益を達成するものであるなら何でもこれを求めるべきであるが、それが人間の便益を害するものであるなら何でもこれを避けるべきではないか。」
回答者　「今までの議論が示すところに従えば、そのようになります。」
質問者　「したがって、強い AI が何であるか、それはあまり本質的な議論ではないのであって、**（１）それら機械の持つ諸能力が人間の利益に資するか否か**、また**（２）人間にとって最大の、あるいは真正の利益とは何であるか**、これらを考えることこそが本質的な議論ではないか。」
回答者　「そうであるかもしれません。」
質問者　「むしろ、これら本質に関わる議論をすることによって、強い AI が何であるかという議論が開始された当初の目的にあるところのものも、自ずと達成されるのではないか。というのも、強い AI が何であるかという結論が明らかになったとて、その最も強い AI とされるものが人間を利するなら優れた技術であり求められ、人間を害するなら劣った技術であり避けられるからである。」
回答者　「そのように思われます。」

質問者　「今言われたことと同様のことを、他の語り方で試してみよう。」
質問者　「ぬいぐるみの目的は、人を癒すことにあるのではないか。」
回答者　「そうです。」
質問者　「ところで、ぬいぐるみは生物を模倣して形作られたのではないか。」
回答者　「はい。」

質問者　「しかし、ぬいぐるみが生物の模倣を完全に達成して、実際に生物になること、つまり、生物の持つ諸特性のなかで、人間を不快にするものまでもがぬいぐるみに付け加わることは、望まれるだろうか。」
回答者　「いいえ、望まれません。」
質問者　「それならば、ぬいぐるみが生物を模倣して作られたのは、人を癒すという目的を達成する限りにおいてであった。そして、人を癒すという目的が達成されなくなるならば、生物の模倣は放棄されるのではないか。」
回答者　「はい。」

質問者　「また、もし生物の完全模倣を放棄しなかったとしても、そのようなぬいぐるみは、少なくともぬいぐるみとしては、人間の必要を満たさないものであり、人間の支持を失って廃れてゆくだろう。また、もし人間の支持を失わないとしても、それはもはやぬいぐるみとしてではなく、ペットなり家畜なり野生の動物なり、他のものとして、である。」
回答者　「はい。」

質問者　「したがって、人間に近いことができる人間の模倣としての人工知能は、人間の利益を満たすという目的を達成する限りで求められる。しかし断じて、人間に近づけることが人工知能の最終目的ではない、少なくとも人工知能が何らかの人間の利益を最終目的としている間は。そしてそのような人間の利益を達成する AI は、人間によって好まれ、人間たち自身の手によって自ずと拡散するであろうが、人間に似るが人間の利益を達成しない AI は、いくら政府や公的機関が推奨しあるいは強制しようとも人々の間では拡散しないであろう。」