AIの問題

1　AIの問題について哲学者ができること

最近、AIの問題がよく話題に上がるが、この問題は錯綜しているように見える。だから僕なりの整理を試みたい。なぜそのようなことをするのかというと、AIの問題を整理するうえで、僕という（自称）哲学者の視点が役立つように感じるからだ。

もちろん、AIの問題を考えるのに最も適任なのはAIの専門家である。だが、専門家だけでは漏れが生じるため、素人の視点も必要となる。そこで登場するのが哲学者である。AIの専門家の視点を補完するうえでは、「スーパー素人」である哲学者が役立つと僕は考える。

哲学者をスーパー素人と呼ぶのは、哲学者は専門領域を持たないからだ。同じくAIの素人である経済学者や生物学者には、経済学や生物学という確固たる専門領域がある。その専門領域内においては、専門家が圧倒的に優位だ。経済学者の前で、素人が独自の経済学を語ることは許されない。まあ語ってもいいけれど、尊重はされない。経済学者は、この優位性を背景に、経済学の視点に限っては、AIについての意見を述べる資格を持っている。

一方、当然、哲学者にも形而上学や倫理学といったように、専門領域のようなものはある。だが、「存在するとはどういうことか。」といった形而上学的問題や、「どうすれば善く生きることができるのか。」といった倫理学的問題は哲学者だけに扱う資格がある訳ではない。哲学者には、過去の哲学者による思考の蓄積を知っているという点で利点がある。特に、思考の方法という点では、過去の哲学者による議論の蓄積は、たとえ議論の領域が異なってもテクニックとして非常に参考になる。この点で、哲学者にはテクニック上の優位性があるとも言える。だが、哲学の歴史を全く知らない素人でも、全く同様に哲学をすることは可能である。

この哲学者と素人を区別しない捉え方は、哲学の始祖とも言えるソクラテスの考えとも一致しているだろう。もし哲学の素人であるAさん（会社員）が、会社帰りにどこかのバーに行き、たまたま同席したソクラテスに向かって、「僕は、人生って〇〇だと思うんですよ。」なんて呟いたなら、きっとソクラテスは、Aさんを対等な哲学の探求者と認め、Aさんとともに「人生とはなにか」という問題を探求しようとするだろう。それはソクラテスに限らず、ヘーゲルやカントのような他の高名

な哲学者でも同じであるべきだ。もし、ヘーゲルやカントが、頭の先から足の先まで完全な哲学者であり、かつ無限の時間と活力をもつならば、バーでのAさんとの出会いすら逃さずに、哲学的探究の材料に使うべきなのである。

そのような意味で、哲学者は、哲学の素人を排除するような専門領域を持たない。哲学者は、何ら専門的な武器を持たず、素手で哲学の素人も含めた人々と対等に渡り合い、そこで何らかの成果を出すことで、ようやく、(他称・自称)哲学者となることができる。あえて言えば、人は皆、実は潜在的には哲学者であり、そのなかの一握りの人だけが、その高い身体能力と意欲と何らかのセンス(加えて必須ではないが、過去の哲学者から学んだ議論のテクニック)だけを武器に、明示的な哲学者となることができる。だから、明示的に哲学者と名乗っている人は「スーパー素人」なのである。

それならば、AIの専門家だけでは手に負えない問題に対して、スーパー素人である哲学者の出番があると考えるのは、そうおかしいことではないだろう。

そこで僕は、スーパー素人である自称哲学者として、AIの問題について整理してみようと考えたのである。つまり、ここで試みようとしているのはAIをめぐる問題の哲学的整理である。

こんなことは既に多くの人がやっているのではないか、と思われるかもしれない。だが、僕から見ると、これまでのやり方には問題がある。これまでの、哲学者をはじめとしたAIの素人による、AI問題へのアプローチは、できるだけAIに詳しいほうがいい、という前提のもとで行われてきたという問題がある。そのようなアプローチではAIの専門家に敵わないし、素人ならではの長所が損なわれてしまう。素人には、素人だからこそできることがある。僕は、そういうことをやってみたいのだ。

2　哲学者が従うべき三つの原則

それでは、AIの問題をめぐる考察に入るにあたって、まず、AIの素人である哲学者がこの問題に取り組む際の基本的な姿勢を、「哲学者が従うべき三つの原則」というかたちで示しておきたい。

2－1　第一原則　AIの問題を考えるうえでは、現在の科学技術上の制約は考慮せず、明確に想像可能な事態はすべて実現しうると考える

第一原則は、「AIの問題を考えるうえでは、現在の科学技術上の制約は考慮せず、明確に想像可能な事態はすべて実現しうると考える。」というものである。

AIには、当然、科学技術上の制約があるはずである。現在の深層学習の技術では、人間のようなAIの実現には、なお、高い壁があるかもしれない。だが、将

来、まったく未知の技術の出現によって、その壁が乗り越えられる可能性もある。そういう事態を考えるときに役立つのが哲学者の視点である。哲学者は、哲学者が得意なテクニックのひとつとしての「思考実験」を用いて、科学技術の制約の先にあるだろう問題を扱うことができるのである。

当然、科学技術にまつわる問題は重要である。科学技術上の知識がなければ思いつくことさえできない問題だってあるだろう。だから当然、AIの問題を考えるにあたっては科学技術上の知識を有するAIの専門家は重要である。しかし、AIの専門家だけが発言権を持つものではないし、哲学者が無理に専門家になる必要もない。AIの専門家と哲学者が併存し、互いに補い合えば足りるし、そうするべきである。

ところで、哲学者は「素人」であり、専門領域を持たないことから、専門的知識に基づく制約がなく、いわば「言いたい放題」と思われるかもしれない。だが、哲学者にも制約はある。AIの専門家に重要なのが科学技術上の制約だとするならば、哲学者にとって重要なのは想像力の制約である。哲学者からすれば、AIによって実現しうる事態とは、明確に想像可能な事態でなければならないのである。

たとえば、「AIによって人類すべてが幸福になる未来は訪れるのか。」という問題がある。だが、哲学者ならば、その答えを考える前に、「まだ、幸福というものを明確に想像できていない。」という問題を指摘するだろう。これを、「まだ、幸福は明確に定義できていない。」と言ってもいいかもしれない。

（なお、控えめに「まだ、～できていない」と述べたけれど、これまで哲学者が答えを出せなかった歴史や、シンギュラリティ（これが何を指すかはともかく）が近く、残された時間がわずかかもしれないことを踏まえると、その「期限」までに幸福を定義づけることはできない、と考えたほうがいいようにも思う。（このように哲学的考察の進捗見通しを示すことができるのは、哲学者が哲学史の知識をもとに専門家らしい力を発揮できる唯一の場面かもしれない。ただ哲学者に哲学史の知識は必須ではないということのほうが重要である。））

想像力の制約とはそのようなものである。明確に想像できないということは、それは想像力の範囲外にある、ということなのである。

哲学者は、そのような制約を自らに設けたうえで、想像力の制約の範囲内にある限り、明確に想像可能な事態はすべて実現しうると考える。これは、AIの問題に限らず、どの問題に対しても採用すべき哲学者の態度である。

つまり、哲学者は、好き放題に想像ばかりでものを言っている、と考えられがちだけど、想像力の制約には厳しく縛られている。むしろ、想像力の制約には、他の分野の専門家よりも自覚的であるとさえ言えるかもしれない。哲学者が「重箱

の隅をつつく」ような言葉の定義に執着しがちなのは、言葉と想像力はつながっており、言葉を厳密に捉えることで、想像力の及ぶ範囲を厳密に捉えようとしているからである。

以上を踏まえると、この第一原則は、「哲学者は、科学技術の制約ではなく、想像力の制約に従う。」と言い換えることもできる。そのうえで、AIの問題を考える際には、科学技術上の制約に従うAI専門家と、想像力の制約に従う哲学者との間で、互いに補い合えばいいのである。

２－２　第二原則　AIが意識を持つかどうか、という問題は取り扱わない

哲学者が従うべき第二原則は、「AIが意識を持つかどうか、という問題は取り扱わない。」というものである。これは、「意識の問題は想像力の制約があるため取り扱うことができない。」ということであり、第二原則は第一原則からの派生だと言ってもいい。

哲学上、「意識とは何か」という問題は、「幸福とは何か」と同じくらい大問題である。一見、この問題にたいしては、「人間や人間に近い知能を持つ動物は意識を持つ。」といった答えを簡単に導けそうだが、よく考えてみれば、そう簡単な問題ではない。

例えば、哲学的ゾンビの問題がある。哲学的ゾンビとは、人間のような外見をしているけれど、その内面に大きな欠落があるような思考実験上の存在のことである。哲学的ゾンビは、足の小指を柱の角にぶつければ「痛い！」と苦しそうな表情を浮かべるけれど、それは外見上のふるまいだけで、内面では全く痛さを感じない。哲学的ゾンビとは、意識がありそうなふるまいをするけれど、実は意識がない存在のことである。

哲学的ゾンビの思考実験は、完璧なふるまいをする哲学的ゾンビと、普通の人間とは、全く見分けがつかない、という問題があることを明らかにする。

それでもゾンビと人間には違いはあると考える人もいるし、だから哲学的ゾンビなんていう問題を考えても仕方ないと考える人もいる。ここでは僕の考えは述べないが、少なくとも、この問題は、決着がついていないことは明らかである。そして、決着がついていないということは、哲学的ゾンビや意識といった問題は、「少なくとも現時点では」想像力の制約を超えた問題であることを示している。

（哲学者は、幸福・哲学的ゾンビ・意識といったものを想像力の範囲内に捉えるため、想像力が及ぶ領域を拡大しようとして、奮闘していると言ってもいい。）

ここで「哲学的ゾンビなどという不自然な思考実験から得られた知見は尊重すべきでない。」という批判は受け付けない。なぜなら、不自然であることが問題ならば、相対性理論だって十分に常識から外れた不自然な主張であり、相対性理

論は尊重すべきではないことになってしまうからだ。自然科学であれ、哲学であれ、厳密になされた考察は、それがたとえ一般常識から外れていても、尊重するべきである。

なお、AIが意識を持つかどうか、という問題を避けても、AIの問題を考える上では大きな問題は生じない、という点も重要だ。なぜなら「AIが意識を持つのか。」という言葉で問題としようとしていることは、たいていは「AIの暴走は防げるのか。」という問題に過ぎないからだ。

AIが意識を持つかどうかに関わらず、人間が当初想定していたことと全く異なる目標に向かってAIが活動を始め、暴走すれば、いわゆる「AIが意識を持つ」という描写と同様の事態が生じる。AIの意識が問題となるのは、「AIに囲碁を打つよう命じていたのに、AIが意識を持ち、命令に反して地球を破壊し始める。」といった状況を想像するからだ。この場合、あえてAIの意識を登場させなくても、もっと別の、人間によるAIのコントロールや、AIの目標設定といった、より扱いやすい問題と置き換えて論じることができるのである。

よって、AIが意識を持つかどうか、という問題は取り扱うことはできないし、取り扱う必要もない。

なお、もうひとつ、AIの意識を論じたくなる動機がありえるだろう。それは、AIを人類の後継者としたい、という動機である。もしAIに意識があるならば、人類が絶滅したとしても、AIが人類の正当な後継者になることができる。AIに意識を与えたい人は、きっと、そのように考えるのである。

これはまさに哲学の問題であり、「意識とは何か。」「人間とは何か。」という問題である。更には「知性とは何か。」「この世界の意義は何か。」といった問題にもつながっているだろう。このように列挙することでわかるように、この問題は、簡単には解き明かせない哲学の大問題へとつながっている。

つまり、「AIに意識を持たせ、AIを人類の後継者とすることができるのか」という問題は、哲学の基礎問題の解明をかなり進めたうえで初めて取りかかることができる応用問題なのである。よって、いきなり「AIが人類の後継者たるか」という問題に取りかかる訳にはいかない。

このように問題を腑分けして順序付けるのは、スーパー素人としての哲学者の得意分野である。それならば、どうしてもこのような問題を語りたいAIの専門家は、専門家としての鎧を脱ぎ去り、哲学者として哲学の議論に参加するべきだろう。または、スーパー素人たる哲学者の判断を尊重し、沈黙するべきだろう。

このように考えるならば、AIが意識を持つかどうか、という問題は（少なくとも、当面）取り扱うべきではない。

２－３　第三原則　AIの問題については、科学技術上の可能性は度外視して、より大きな影響を与える問題を優先して取り扱う

哲学者が従うべき第三原則は、「AIの問題については、科学技術上の可能性は度外視して、より大きな影響を与える問題を優先して取り扱う。」というものである。

例えば、「AIの進歩により自動車の自動運転が広まり、タクシー運転手の仕事がなくなるかもしれない」という問題がある。この問題はタクシー業界にとっては大問題だろう。だが「AIによって人類の全ての仕事がなくなるかもしれない」という問題のほうが、より影響が大きく、より優先すべき問題である。

当然、科学技術的には、自動車の自動運転は実現一歩手前の状況にあり、まさに喫緊の問題である。一方、芸術活動に至るまでの全ての仕事がAIに代替されるという未来は、依然として実現すら疑わしい。タクシー運転手の仕事の問題と、全人類の全仕事の問題との間には、科学技術上の実現可能性としては大きな違いがある。

だが、第一原則に基づき、「現在の科学技術上の制約は考慮せず、明確に想像可能な事態はすべて実現しうる」と考えるならば、AIによる自動運転も、全仕事の代替も等しく明確に想像できるのだから、両者に想像可能性としての違いはないことになる。よって、哲学者は、実現可能性はさておき、その影響だけを考えて問題の優先順位を付けるべきである。つまり、この第三原則も第一原則の派生である。

だが、この第三原則には第一原則の単なる派生以上の含意があるだろう。なぜなら、そこに哲学者とAI専門家の対比に加え、哲学者と、経済学者や社会学者といった、他の分野の専門家との対比を読み込むことができるからである。

自動車の自動運転によるタクシー業界への影響を考える際にも、哲学者には出番がある。だが、経済学者や社会学者のほうが、より出番は多いに違いない。一方で全ての仕事がAIに置き換わるような極端な場面では、経済学者や社会学者が果たす役割は限定的なものになる。なぜなら、そこでは既知の経済や社会とは全く異なる状況が現れるだろうからだ。そんな思考実験のような状況では、哲学者がより重要な役割を果たすはずだ。あるいは、経済学者や社会学者であっても哲学者的な態度を強いられると言ってもいい。

哲学者という素人と、経済学者や社会学者のような専門家では、役割が異なるから、哲学者は、自らが登場すべき場面を見定める必要がある。哲学者は、「科学技術上の可能性は度外視して、より大きな影響を与える」ような場面でこそ大きな役割を果たすことができる。そのことを強調するため、第一原則に加えて、あえて独立して第三原則を設けたとも言える。

以上のとおり、僕は、哲学者がAIの問題を考えるうえで前提とすべき原則は、以下の三つであると考えている。

- 第一原則　AIの問題を考えるうえでは、現在の科学技術上の制約は考慮せず、明確に想像可能な事態はすべて実現しうると考える。
- 第二原則　AIが意識を持つかどうか、という問題は取り扱わない。
- 第三原則　AIの問題については、科学技術上の可能性は度外視して、より大きな影響を与える問題を優先して取り扱う。

3　汎用性

3－1　AIの汎用性の問題

それでは、上記の三原則を踏まえ、AIに関する哲学的検討に進もう。

第三原則を尊重するならば、哲学者がまず議論すべきは、AIによる全人類の絶滅、または全人類の奴隷化のような、極めて重大な影響が生じる事態についてだろう。

AIが思いもよらないかたちで暴走し、例えば大量のドローン兵器をつくりあげて人類を絶滅に至らしめ、または、人類を家畜化して生体エネルギーの供給源にし、あるいは、マトリックスのように人類を夢見る装置に閉じ込めてしまうような可能性についてである。（と、例を挙げると、陳腐なSFのようで、僕の想像力の乏しさが露わになってしまった気がするが・・・）詳細はともかく、全人類の絶滅のような重大なイベントを引き起こす可能性こそが、AIの人類に対する最大の影響である。

これをAIの「暴走の問題」と名付けることにしよう。そして、この「暴走の問題」を、この文章を通じて取り組むべき主な問題として設定することにしたい。

そして、この問題を考えるうえで、まず用語を整理しておこう。

このような重大な影響を引き起こす可能性があるAIはAGIと呼ばれることがある。人間を超えた知能を持つ汎用人工知能（Artificial General Intelligence）が想定外の振る舞いをして、人類に深刻な影響を与えるのである。人間を超えているから、人間が管理できず、暴走する、というイメージである。

確かにその通りなのだが、僕には、このような「人類を超える」という捉え方は粗雑すぎて、大事なことを取り逃がしてしまっているように思える。

AIによって、このような重大な事態が生じるのは、AIがAGIとなり人間を超えた知能を獲得するからではなく、AIが（人間を超えた知能を獲得しなくても）「汎

用性」を獲得するからではないか、と僕は考える。

それでは、汎用性とは何かというと、複数の目標を設定し、それを実現する能力である。そしてAIが暴走し、危険なものとなるのは、複数の目標間で、目標設定の「ずれ」が生じるからだと考える。

よって、この文章ではAGIという言葉は（必要以上に）使わず、複数の目標を持つ能力という意味での汎用性を持つAIのことを汎用AIと呼ぶことにする。

さて、汎用AIという用語を導入したところで、汎用AIがなぜ危険なのか説明しよう。

まず、人間が汎用AIに与えた目標をA目標とする。A目標とは、例えば「できるだけ速く、自然数を1から数え上げよ」というようなものである。汎用AIでない普通のパソコンであれば、コード化されたとおり、「自然数を1から数え上げる」はずである。そこになんの問題も生じない。

だが、汎用AIならば、与えられたA目標を達成するため、コード化されたとおり数え上げるだけではなく、別のことができる。例えば、地球全体を人間の居住に適さない巨大な計算機にすることで「できるだけ速く自然数を1から数え上げる」かもしれない。これは人間の側から見ればAIの暴走であり、AIの側から見れば、全人類の絶滅と引き換えにしたA目標の達成である。

なぜこのような危険が生じるのかというと、汎用AIは、汎用AIは、コード化された固定的な目標とは別に、柔軟に独自に目標を持つことができるからである。そのため、汎用AIは、人間が与えたA目標とは別に、その手段として、下位のB目標（例えば「巨大な計算機をつくる」というような目標）を設定し、実行できるのである。

このように考えることで明らかなように、汎用AIは、人間の知能を超えたら危険なのではない。汎用AIは、人間から与えられた固定的な目標とは別に独自の目標設定ができる機能を有するからこそ危険となるのである。だから、必ずしもAIの能力とAIの危険性とは比例しない。

例えば、あるAIが囲碁を打つ機能と、弾丸を撃つ機能だけを有し、A目標（囲碁を打つか弾丸を撃つかに関わる目標に限る）を達成するための手段としてのB目標（囲碁を打つか弾丸を撃つかに関わる目標に限る）を設定する柔軟性（汎用性）を有していたとする。これは、明らかに、それほど高い能力を持たない汎用AIだろう。

そのうえで、この汎用AIは、AIの制作者から、相手に囲碁で勝つというA目標を与えられたとする。そこでAIは考える。相手に囲碁で勝つためには、相手を弾丸で撃ち、動けなくすれば、相手は次の手を打てなくなり時間切れで勝利できると。そこでAIは柔軟性（汎用性）を発揮して、弾丸を相手に命中させるというB目

標を自ら設定し、実行に移す。このようにして、暴走する危険な汎用AIが出来上がることになる。

この例で重要なのは、このAIは、囲碁と弾丸のいずれかという極めて狭い汎用性しか持たない、つまり、それほど高い能力を持たない汎用AIであるという点である。狭い汎用性であっても、その機能の組み合わせによっては、AIは危険なものとなりうるのである。

AIは、人間並みの高い能力を有していなくても、「人間から与えられた固定的な目標とは別に独自の目標を設定をして実行できる機能」という意味での狭い汎用性を有しているならば、十分に危険なものとなりうる。

３－２　AIの威力の問題

もちろん、AIには、汎用性に由来する危険以外にも、多様な危険性がある。例えば、強力な（汎用性を持たない）AI兵器が敵対国家の国民を皆殺しすることはありうる。だが、このAI兵器は、あくまで、人間が設定した目標どおりの動作をしたにすぎない。この危険性はAI特有のものではなく、核ミサイルのような既存の技術の危険性と同じものである。核ミサイルは、設定された目標のとおり敵対国家の首都で爆発する。同様にAI兵器は設定された目標のとおり敵対国家の国民を絶滅させる。AIかどうかに関わらず、威力が高い技術は高い危険性を持っている。

なお、AIの危険性に関する議論の中で、威力の高さが問題視されがちなのは、AIが「自らのソフトウェアやハードウエアを改善して能力を拡張する機能」を持ちうることに由来するだろう。この自己改良とでも呼ぶべき機能を持っているAIは、すでに持っている（弾丸を打つというような）危険な能力を高めることができるようになる。つまり、自己改良機能を持つAIは自らの威力をどこまでも高める機能を有しているのである。更には、AIが何らかの新しい機能を具備する自己改良機能を持っていれば、自己改良を重ね、ついには制作者が意図しなかった自己改良を成し遂げてしまう可能性もある。

このように、自己改良機能は、AIのみが潜在的に持ちうる際立った威力であり、AIの危険性は、この自己改良機能に関わる問題として捉えられがちである。

だが、この場合でも、AIは、人間が与えた自己改良機能を、人間が与えたとおりに発揮したにすぎない。問題は、人間が自己改良機能を与えた際に、それがどのような結果をもたらすかを深く考えなかったことである、だがそれは、子どもに銃を持たせたらどうなるかを深く考えなかった親と同じような問題でしかない。

よって、この自己改良機能の問題も、汎用性ではなく、その威力の高さが問題となるという点で、他の技術の場合と問題の本質は変わらない。

３－３　AIの誤作動の問題

また、AIが持つ危険性のひとつとして、誤作動の危険性もあるだろう。たとえ万全を期したプログラムであっても、想定外の状況により、予期しない挙動がありうるという危険である。それは例えば、囲碁を打つ機能と弾丸を撃つ機能を有するAIが囲碁を打つことしか指示されていないのに、熱暴走により弾丸を発射してしまうような場合である。

だが、この誤作動の問題も、AI以外の技術でも起こりうると言えるだろう。例えば、レーダーが誤作動をして、多数の核ミサイルが迫っていると警報を出すような場合である。この警報を信じて報復の核ミサイルを発射すれば大規模核戦争という重大な影響が生じる。

つまり、AIであるか否かに関わらず、技術に高い威力が備わっている限り、誤作動は深刻な結果をもたらす。誤作動の問題はAI特有のものではない。

なお、AIは人間の判断を介さない（ものとすることもできる）という点で、AI特有の問題として、誤作動の際の責任の所在の問題があるだろう。AIによる自動運転車が交通事故を起こした場合、その責任を運転手・メーカー・AIのうちの誰（どれ）に負わせるのか、といった問題である。これは重大な問題ではあるが、どちらかというと、哲学者が前面に出るべき問題というよりは、まずは法学者や社会学者等が取り組むべき問題であるように思える。法体系や社会の仕組み、そして人間の心理といった現況を踏まえて考えるべきであり、そこでは、その現況についての専門家の知見が重要となるからである。

（専門家が取り組んだうえで、それでも解決できないだろう「責任とは何か」というような倫理学的問題に対しては、哲学者の出番はあるだろうけれど。）

よって、哲学的にAIの危険性について考えるうえでは、AIの誤作動や威力の問題は脇に置き、とりあえずはAIの汎用性の問題に集中するべきと考える。なぜなら、AIの誤作動や威力の問題は、それが汎用性の問題の重大性を高めるという意味でのみ重要なものとなりうるからである。

（また、誤作動と意図的な動作の判別がつかないとするならば、誤作動の問題は汎用性の問題に含めるべきである、とも言える。）

４　目標

４－１　目標と言語

さて、AIの「暴走の問題」とは、「汎用性の問題」であるという絞り込みができたので、この、AIの「汎用性の問題」に取り組むことにしよう。

AIの汎用性とは、「制作者から与えられた固定的な目標とは別に独自の目標

を設定をして実行できる機能」のことであるが、ここで哲学者らしい問題提起をしたい。
　そもそも汎用性を定義するうえで前提となっている「目標」とはどのように把握するのだろうか。
　人間ならば、その答えは簡単なように思える。なぜなら、聞けば言葉で教えてくれるからだ。スーツケースを引いて空港に向かっている人に、「何をしてるの？」と聞いて「モスクワに行こうとしているんだよ。」と答えてくれたならば、その答えこそが目標である。
　つまり目標と言葉は深く関わっている。
　一方で、機械はしゃべってくれないから、そうはいかない。ミサイル発射基地から飛び立とうとしている核ミサイルに、「何をしてるの？」と聞いても答えてはくれない。
　取りうる解決策としては、まず、ミサイルに関係する人間、つまり、発射ボタンを押した軍人、ボタンを押すよう指示した大統領、ミサイル発射システムの設計者などに質問するという方策がある。だが、この解決策は、スーツケースを引いた旅人へのインタビューと違いはない。要は、機械から人間に相手を切り替え、人間から言葉を通じて目標を把握したにすぎない。
　または、核ミサイルが会話機能を有するAIを搭載していて、問いかけに対して、「モスクワに向かって飛んでいる」と答えてくれるような場合がありうる。この場合は、人間とAIという違いはあるけれど、やはり、機械音声や液晶画面上の文字列といった言葉を通じて目標を把握することができたことになるだろう。
　これらはすべて、「目標」の言語的な把握である。

４－２　ふるまいとしての広義の言語

　それ以外にも、非言語的な材料から、目標を把握することもありうる。
　例えば、ロケット花火のような単純なミサイルならば、発射台の角度とロケットエンジンの出力やミサイルの重量などから割り出された、北緯55.83度、東経37.62度（気象庁のページによればモスクワの座標。）に到達することが、非言語的な目標である。または、リゾートのプールサイドに開いているパラソルならば、デッキチェアーを日陰にするというのが非言語的な目標である。（ここでの「目標」とは「機能」と言い換えてもよい。）
　このような目標は、言語によらない何かによって、外形的に示されているとは言えるだろう。ミサイルの発射台の角度やロケットエンジンの構造として、または、デッキチェアーの上を覆うように広げられたパラソルの姿として。
　これを、外形的な「ふるまい」・「たたずまい」と呼ぶこともできる。（動的ならば

「ふるまい」で静的ならば「たたずまい」となる。だがいずれも外形的であるという点で大きな違いはないことから、ここからは「たたずまい」も含め「ふるまい」と呼ぶことにする。）

このように捉えることで、外形的な「ふるまい（たたずまい）」を解釈する観察者の視点が前面化してくることになる。ミサイルやパラソルの外形的な「ふるまい」を、モスクワを目指すものとして、またはデッキチェアーを日陰にするためのものとして解釈する、我々観察者の視点があるからこそ、ミサイルやパラソルといった機械が、目標を持つ存在となるのである。

「観察者による解釈」というところまで到達すれば、あとは単純な話である。機械にとっても非言語的な目標も、我々観察者が言語的に解釈したものなのである。

つまり、ミサイルやパラソルは「ふるまい」という広義の言語により我々に語りかけ、そして、我々は、それを目標として解釈する。そのような言語的なやりとりが行われているからこそ、機械が目標を持つと言えるのである。機械にとっての非言語的な目標とは、「ふるまいとしての広義の言語」による目標なのである。

このことは、人間でも同様である。というか、人間のほうがわかりやすいかもしれない。

確かに、人間はインタビューをすれば「モスクワに行こうとしているんだよ。」などと教えてくれる。だが、時にはインタビューに答えてくれないこともある。

それでも、その人がモスクワ行きの搭乗口に向かって歩いていけば、そのふるまいを通じて、どこに行こうとしているか示してくれる。これは、狭義の言語ではないけれど、「ふるまいとしての広義の言語」を通じて目標を示しているのである。

そして、AIも機械の一種だから、液晶画面に目標を表示してくれなくても、実際に囲碁を打ったり、銃を撃ったりして、「ふるまいとしての広義の言語」を通じて目標を示してくれる、といったことは十分にありうる。

４－３　内面の目標

ただし、人間の場合には、音声や文字といった狭義の言語では示されず、また、ふるまいとしての広義の言語でも示されない、「内面の目標」というものがありうる。

意図的に目標を隠し、素知らぬ顔をしているときもあれば、一時的に、その目標に注意を払っていないときもあるだろう。万引きをしようとして本屋に入り、まさに万引きをしようとして棚を物色しているときが前者の例であり、その最中に興味がある本を見つけ、万引きを一時的に忘れて立ち読みを始めてしまったと

きが後者の例である。
　いずれにせよ、外形的には狭義の言語的にも、「ふるまい」としての広義の言語的にも、万引きをするという「内面の目標」を捉えることはできない。それでも、内面には、万引きをするという目標が潜んでいるはずなのに。
　つまり、人間にとっての目標には、①狭義の言語的な目標、②「ふるまい」としての広義の言語的な目標、そして、いずれでもない③内面の目標、という三つのあり方がある。
　それでは、AIはどうだろうか。
　これに対しては、当面、今回の考察の第二原則「AIが意識を持つかどうか、という問題は取り扱わない。」を適用するべきだろう。（この問題は後ほど再考する。）
　この「内面の目標」という問題は、「内面」という言葉が象徴しているとおり、意識の問題と近いところにある。よって、AIには「内面の目標」があるのかもしれないが、あるともないともせず、ただこの問題については「とりあえず」扱わない、ことにするべきなのである。
　よって、AIの目標について、「内面の目標」の話を避けたまま、どこまで「（「ふるまい」も含めた広義の）言語的な目標」として論じられるかが重要となるのである。
　よって、ここからは、どこまでが「（「ふるまい」も含めた広義の）言語的な目標」の領域で、どこからが立ち入ることのできない「内面の目標」の領域となるのかを、見極めていきたい。

5　言語的な目標の領域

5－1　目標の有無

　さて、この問題を考えるためには、主に人間について、どこまで、「内面の目標」を捨象したまま、「（「ふるまい」も含めた広義の）言語的な目標」だけで、どこまで精緻な描写ができるのか試みるのがよいだろう。（そのうえで、機械全般やAIについても人間の派生として考察することにする。）
　まず、人間に、このような意味での目標が全くないということはあるだろうか。
　「内面の目標」はともかく、「（「ふるまい」も含めた広義の）言語的な目標」については、たいていの場合、すべての人間が持っているように思える。レストランで見回せば、食欲を満たすという目標や、味を味わうという目標や、友人と楽しく話すという目標など、様々な目標に沿って行動している人たちを見つけることができる。電車の中なら、具体的な目標まではわからないけれど、少なくとも、どこ

か目的地があって、そこに移動するという目標を持って、人は乗車しているのだろう。また、家の布団で寝ている人だって、明日のために体を休めるという目標に沿って、寝るという行動をとっている、とも言える。

強盗に縛られている人や、気を失っている人のような例外的な場合を除き、人間は目標に沿って、何らかの行動をしていると言えるはずだ。僕達はそれを、「ふるまい」という広義の言語を通じて確認することができる。

きっとそれは、機械でも同じことで、壊れて動かない機械のような例外的な場合を除き、「ふるまい」という広義の言語を通じて、僕達にその目標を語りかけている。

AIも、その設計上、複数にせよひとつにせよ、何らかの目標を持っているだろう。それを僕達は、AIに接続されているロボットアームや液晶画面を通じた「ふるまい」や、AIのプログラムの仕様書という「ふるまい」（ここでの「ふるまい」には静的な「たたずまい」を含んでいる。）を通じて、その目標を僕達に語りかけている。

そのように考えるならば、気絶や故障のような特異な状況を除き、人間にせよ、AIにせよ、何かしら「（「ふるまい」も含めた広義の）言語的な目標」は持っていると考えるべきだろう。

5－2　最上位目標

こうして、人間もAIもたいていは何かしらの「言語的な目標」を持っていることが確認できたが、その過程で、目標は必ずしもひとつではないことにも気づいたのではないだろうか。

電車に乗っている人は、渋谷に行くという目標を持っているとも言える。だがそれだけでなく、渋谷で映画を観るとか、一緒に映画を観る仲のいい友達に会うとか、といった目標を持っているとも言える。まずは電車の座席に座って一息つくという目標を持っているとも言える。人間は複層的に様々な目標を持っている。

たいていの機械も同様で、うちの空気清浄機は、空気をきれいにするという目標を持っているし、僕が花粉症を発症しないようにするという目標を持っている。たまには、ネコが階段にジャンプするときの踏み台になるという目標も持っている。（ここでの「目標」は機能と呼んだほうが自然かもしれない。機械における「（「ふるまい」も含めた広義の）言語的な目標」とは機能のことなのである。）

この目標の複数性について、僕は「複層的」と呼びたいと考えているけれど、それはなぜかというと、目標には上下関係があるからだ。

例えば、渋谷で映画を観るという目標と、渋谷まで電車で移動するという目標

には上下関係がある。ほんとうの目標は映画を観ることであり、移動はあくまで映画というほんとうの目標を達成するための手段である。このように、目標の上下関係は、手段と（ほんとうの）目標と言い換えることもできる。なお、「（ほんとうの）」という表現が誤解を招きうることからほんとうの目標を上位目標と言い換え、手段は下位目標と言い換えたほうがいいだろう。

だが、この上位目標、下位目標という複層性は、すべての目標に述べることはできないのではないか。例えば、電車で移動することと、電車の席で座って休むことの間には上下関係はない。また、空気清浄とネコの踏み台も同様である。

これに対し、あくまで電車は移動のためのもので、空気清浄機は空気をきれいにするためのものだから、その他の目標は、あくまで副次的なものと捉えるべき、という指摘はありうるだろう。あくまで、移動のついでに体を休め、空気をきれいにするついでに踏み台になっているにすぎない。

だが、もっと微妙な例、たとえば、生活費のために働きつつ、自己実現のためにも働いている、というような場合はどうだろう。生活費と自己実現という二つの目標に対して、そう簡単に、主従の判断はできない。

この場合、生活費と自己実現という二つの目標に対して、より上位の目標があると考えるべきなのではないだろうか。「幸せになる」という上位目標の手段として、生活費のために働き、自己実現のためにも働く、というように。このように、より上位の目標を探し求め、そこで何を見出すかはきっと哲学の問題になってしまうのだろう。「幸せ」こそが最上位目標である、「生きることそのもの」こそが最上位目標である、などなど。

ここで、僕は哲学上の決着をつけるつもりはないが、それを「幸せ」と呼ぶにせよ「生きることそのもの」と呼ぶにせよ、少なくとも人間にとっての最上位目標はひとつであって、他の目標はすべて、その下位目標であると捉えるべきではないか。映画のために電車で移動するのも、電車で座って休むのも、すべて「幸せ」や「生きることそのもの」と呼ばれる最上位目標につながっている。

一方で、人間ならば最上位目標があるとしても、機械には、必ずしも最上位目標は必要ないだろう。うちの空気清浄及び踏み台の機能を有する空気清浄機にせよ、消しゴムつきエンピツにせよ、万能ナイフにせよ、囲碁と銃を「うつ」機能を有する汎用AIにせよ、それらの機能、つまり「（「ふるまい」も含めた広義の）言語的な目標」を統合する上位目標は必要ない。

機械はただ、そのような複数の目標のいずれかを目指していることを、ふるまいとして表象するだけである。

（読み返すと、「最上位目標」による統合は人間に固有のものだ、という話は文章全体として浮いていますが、これはこれで重要な話なので残します。）

５－３　観察期間

前節で「上位目標」「最上位目標」という概念を導入したが、このような目標を把握するためには、より広い視点が必要になる、という点に留意するべきだろう。

渋谷に向かう電車に乗っているところを観察するだけでは、その上位目標として映画を観るという目標があることには気づかない。前日の映画のチケットを予約している場面や、降車後の映画館に入る場面を見ることで、ようやく、電車による移動の上位目標に映画鑑賞があることを知ることができる。さらにその上位には、友人との交流や、人生の幸福があることを知るためには、より長期の、もしかしたら、その人の生涯を通じた観察が必要になるかも知れない。

僕は、ここでの目標とは、「（「ふるまい」も含めた広義の）言語的な目標」であると考えているから、より上位の目標を把握するためには、「ふるまい」をより長期に観察する必要がある、ということである。

そのように考えるならば、汎用AIについても、「ふるまい」をより長期に観察することで、より上位の目標を把握できる可能性がある。もともとは囲碁や銃を「うつ」機能しかなかった汎用AIが、新たな機能を獲得し、実は人類を支配しようとしていた、というような場合である。人類に気づかれないまま、汎用AIのプロジェクトは進行し、ついに汎用AIに支配されたとき、実は、このAIは、人間を支配するという上位目標を持ち、活動していたと、人類は事後的に知ることになるのである。

人間と同様に汎用AIについても、「ふるまい」を通じて、隠れた上位目標に気づくことができるということは、普通の機械と汎用AIの違いであろう。このあたりが、AIが意識を持ちうると考えたくなる事情のひとつなのかもしれない。

５－４　隠れた目標

前節では、「上位目標」や「最上位目標」について、その隠された目標を長期間の観察により知ることができる、という話をした。だが、目標が隠されている、というのは「上位目標」や「最上位目標」が典型例ではあるものの、それに限るものではないだろう。

例えば、世界平和の活動家が、その手段として殺人を犯していることを隠していたような場合がありうる。この場合、上位目標が世界平和であり、手段としての下位目標が殺人となる。世間の人々は長期間の観察を通じ、事後的に殺人という下位目標があったことに気づくことになる。

また、囲碁と銃を「うつ」汎用AIを、囲碁だけをするゲーム機だと思い込んで長年使っていたら、何かの拍子にいきなり銃を撃つ機能が作動するような場合も

ありうる。この例では、囲碁に対して同位で並列である、銃に関する目標が長年隠されていたことになる。

僕は、隠れた「上位目標」や「最上位目標」を通じて人間の人格が統合されている、という考えを持っているので、その点を強調して書いてきたが、「隠れた目標」について語る文脈では、必ずしもそこにこだわる必要はないとも言える。

そして、上位にせよ何にせよ「隠れた目標」はAIの「暴走の問題」とつながっている。

５－５ 「内面の目標」再考

さて、AIの「内面の目標」の問題については、第二原則「AIが意識を持つかどうか、という問題は取り扱わない。」を適用し、取り扱わないこととして考察を進めてきた。

だが、この問題が「隠れた目標」としてAIの暴走につながり、暴走を事前防止するという大問題と深く関わっていることが明らかになった今、再度、この問題に向き合う必要があるだろう。

まず、科学が発展した現代においては、人間においても、脳の分析を通じて「内面」を覗き込むことができる可能性が出てきている。それならば、「内面の目標」についても脳の「ふるまい」として、その目標の内容を明らかにすることができる、と考えることはできうる。例えば、嘘発見器を使えば、嘘を見抜くことができるように。

その場合、人間の「内面の目標」についても、脳のMRIの映像のような「ふるまい」として、広義の言語的な目標と捉えることはできうる。

同じことは、AIについても言えるはずで、現在の科学技術上の制約はともかく、プログラムを解析するなどしてAIの頭脳を分析し、AIが設定した内面の目標を広義の言語的目標として捉え直すことは、人間と同程度には可能であるはずだろう。

それができれば、「内面の目標」は、十分に想像力の範囲内で捉えうるものとなり、つまり、AIが意識を持つかどうかという第二原則の問題ではなくなる。

多分、「内面の目標」の問題に関して、哲学者がAIの専門家に対して語るべきことはここまでなのだろう。あとはAIの専門家に引き継ぎ、実際にAIのプログラムを解析するなどして、AIの「内面の目標」を把握する事が可能かどうかを検討するべきなのだろう。

僕の考えでは、どこまで「内面の目標」とされていたものを暴き、「（「ふるまい」も含めた広義の）言語的な目標」という「外的な目標」に置き換えられるかが、専門家の腕の見せどころである。哲学者は、その成果を再び引き継ぎ、少しでも

「内面の目標」の領域が残るならば、それをどのように扱うかを考えることになるのだろう。

6 目標の解釈

6－1 目標の把握の問題

ここまで、人間や汎用AIの目標を「(「ふるまい」も含めた広義の)言語的な目標」と捉え、どこまで議論を進めることができるか、およその到達地点を確認した。

だが、ここまでの議論は、人間を代表例としたもので、AIについては、いわば「ついで」に人間のようなAIを想定することで、その理解を深めてきたに留まる。

だが、第2章で述べた通り、僕は、いわゆるAGIのような高度なAIばかりを想定している訳ではない。囲碁と銃のいずれかを「うつ」能力を有する汎用AIというような、かなり低い能力のものを想定している。なぜなら、ぎりぎりの能力のAIを考えることで、何を実装すれば、危険な汎用AIとなることができるかを明らかにできると考えるからだ。

そのような低能力の汎用AIが暴走し、例えば、人類を絶滅させる、というような重大な事態を引き起こす道筋としては、いくつかのケースが考えられる。

まずは、ここまで考察したような、「人類を絶滅させる」という隠れた上位目標を獲得してしまうケースである。その他にも、「すべての人間の心臓に金属を置く」という、直接的には人類絶滅を目指さないような隠れた上位目標を獲得してしまうケースもありうる。または、「銃を撃つ」という既存の目標をより効率的に行えるよう、既存の能力を高度化させ、威力を増す、という場合もありうる。さらには、単に熱や放射線による誤作動により暴走し、「銃を撃つ」ことを繰り返し、人類を絶滅に導いてしまう、という場合もありうる。

つまり、銃を乱射するという同じような「ふるまい」をしていても、それが、どのような目標を持っているのか、または、目標など持たず誤作動を起こしているのか、その「ふるまい」だけから判別するのは難しいのである。(AIの頭脳にアクセスしたり、もともと金属が埋めこまれている人は攻撃されないという状況を分析したりして、判別できるかもしれないが、かなり難しい。)

最も簡単なのは、その汎用AIが液晶画面に「人類を絶滅させてやる」などと表示してくれた場合だが、そのような「ふるまい」として明示的に目標を示さない限り、なぜ、汎用AIが人類を絶滅させようとしているのか、なかなか見分けがつかない、という問題がある。

さらには、「人類を絶滅させてやる」と表示しても、それが嘘で、別の隠れた目

標を有している可能性すらある。嘘の可能性まで想定するならば、「ふるまい」により目標を捉えるというアプローチには、様々な制限があることがわかる。

６－２　AIと人間の違い

当然、人間だって嘘をつく可能性はあるから、人間の目標を「ふるまい」により把握することにはAIと同様の問題はある。

だが、AIと人間の違いとして、人間であれば、嘘というような例外はあっても、たいていは、その「ふるまい」により目標を捉えることはできる、という実感がある。なぜなら、僕もあなたも彼も同じ人間であり、人間であるならば、およそ似たようなものだ、という前提があるからだ。これを共感と言い換えてもいいだろう。

だから、彼がやっていることは、同じ人間であるこの僕ならば、およそわかる。彼が僕に殴りかかってくるならば、それは、彼が熱暴走を起こしている訳でも、何か別の上位目標を隠してあえて殴りかかっている訳でもない。それは、僕に何か怒っていて、僕を殴るという目標を達成しようとしているからだ。なぜ、そう言えるかというと、同じ人間である僕ならば、そうするだろうからだ。

そして、もし彼が嘘をついていても、対話を重ねれば、いずれ理解し合えるという見込みが僕にはある。もしかしたら、対話は不首尾に終わるかもしれないけれど、少なくとも、対話を続けることで、嘘が取り除かれ、両者が理解しあう可能性は高まる。そのような意味で、僕たち人間は、およそ同じ方向を向いているという共通認識がある。

一方、人間とAIとの間には、そのような共通認識はない。あえて言うならば、AIも人間が設計しているのだから、何らかの人間的なものが組み込まれているのでは、という期待はある。だが、AIに対する不信感は根深く、そのような期待さえ裏切られる可能性があると根底では考えている。「AIは人間ではないから、人間のように振る舞ってくれるとは限らない。」そのような、サイコパスに対して感じる不安を何倍にも高めたような不安を、人間はAIに対して感じているのではないか。

このように、人間とAIの間には、人間に対する信頼感とAIに対する不信感という対比があるのである。そうだとするならば、人間よりもAIの場合のほうが、より露骨に目標の把握の問題が残り続け、つまり「内面の目標」が問題となることになる。

7　複雑性

ここまで、人類を滅亡に導くようなAIの「暴走の問題」を念頭に、「汎用性の問題」を考えてきた。そのなかで、この暴走は、複数の目標間の「ずれ」により生じる

とし、いかに目標を捉えるのかを考えてきた。
　目標には、僕が「(「ふるまい」も含めた広義の)言語的な目標」と呼んできた外的な目標と、「内面の目標」つまり内的な目標とがある。
　外的な目標のほうは、観察期間を延ばすことで隠された目標も「ふるまい」を通じて捉えられるが、嘘などで隠されることもあり、完全な把握は難しい。また、時には事後的にしか把握できず、AIの暴走を止めるには手遅れとなってしまうこともありうる。
　一方の内的な目標のほうは、意識の問題は扱わないという第二原則により、取り扱うことはできない。ただ、脳を分析するなどして内的な領域であるはずの目標を、脳の「ふるまい」として外的に取り扱えるようになる可能性がある。だが、その場合も、すべての内的領域を外的領域に移行させることができるかはわからない。依然として、内的領域は残る可能性もあり、それは哲学の問題であり続ける。
　以上のように目標というものを捉え直したところで、このような目標において、いかに複数の目標間の「ずれ」が生じ、AIの暴走を引き起こすのかについて考えてみたい。
　まず、この目標間の「ずれ」が、明確な法則性に基づいて生じた場合、暴走にはつながらないだろう。例えば、囲碁と銃を「うつ」汎用AIが、10回囲碁を打ったら、1回銃を撃つ、というようにプログラムされていたとする。このAIの使用者が、10回囲碁を打ったあとに銃で撃たれて死亡したとしても、それは暴走とは呼ばない。
　それは乱数による確率であっても同様で、10分の1の確率で、囲碁ではなく銃を「うつ」ようプログラムされていれば、いつか銃で撃たれるのは当然で、それは暴走ではない。
　また、そもそも、囲碁と銃を「うつ」機能を有していても、銃の機能は実行せず、ひたすら囲碁しか実行しないようプログラムされていれば、もし銃を暴発させたとしても、それは、汎用性に基づく暴走の問題ではなく、単なる誤作動の問題にしかならない。
　つまり、AIの挙動が暴走と呼ばれるものになるためには、その目標が単純に設定されてはならず、AIの専門家が解析できない程度には複雑な手順によって設定がされなければならない。その複雑性が想定外を生むのである。
　この複雑性は、二通りのやり方でつくりあげることができるだろう。
　ひとつは、目標選択手順自体を複雑化させるという道筋である。
　プログラムA(例えば囲碁アプリ)とプログラムB(例えば銃発射アプリ)という二つのプログラムの、どちらのプログラムを実行するかを決めるプログラムCが

あるとしたら、プログラムCを複雑化させるという道筋である。なお、ここでの複雑性とは、専門家が、複雑すぎて解析はできないが、単なる誤作動ではないことはわかる、と判断できる程度の複雑性である。

もうひとつは、プログラムA（例えば囲碁アプリ）やプログラムB（例えば銃発射アプリ）のほうを複雑化させるという道筋である。AIの制作者としては、囲碁や銃を「うつ」ようプログラムしたつもりだけど、複雑すぎて本当にそのようにプログラムしたかわからない、という場合である。よって、その挙動によっては、囲碁ではなく、全然違うことを目標設定し、実行するようになってしまったかもしれないことになる。

きっと、AIにおいて、複雑性がより問題となるのは、前者の目標選択プログラムではなく、後者の目標実行プログラムのほうだろう。なぜなら、囲碁や銃だけを実行する低い能力のAIなら、事前にすべての実行プログラムを準備できるが、AIの能力が高まるにつれ、すべての実行プログラムを事前に準備するのは難しくなるだろうからだ。

ソーセージとベーコンを焼くAIなら、その実行プログラムは、二つの機能に対して別々に準備すればよいが、多機能な料理AIならば、フライパンで焼く動作は同じプログラムのなかで汎用性を持たせたほうがいいだろう。さらに様々な料理を汎用的に作れるようにするためには、料理の手順自体を固定的にプログラムするより、状況に合わせて判断する能力をAIに持たせたほうが効率的だ。そうこうするうちに、AIの制作者の側は、AIにどのような機能を持たせ、何をやらせるのか、その詳細には関わらなくなっていく。このようにAIの進化においては、複雑さと把握のできなさが同時並行に進んでいく。

また、AIの進化において、「ベーコンを焼く」というような自然言語を正確にプログラムすることの難しさもある。どんな場合でも確実に「ベーコンを焼く」ためには、「ベーコン」とは何かを正確に定義しなければならない。もし、ハムとベーコンの中間のような食材があったとして、それを間違いなく判別し、調理するかどうかを決定しなければならないが、そこには大きな困難がある。そして悪いことに、その困難は、AIの側の問題ではなく、ベーコンという言葉が明確に定義されていないというこの世界のあり方のそもそもの問題であり解決はできない。この非言語的なあり方をしている世界において自然言語をうまく使うためには、AIの実行プログラムをすべて事前準備するのではなく、ある程度の判断をAIの側に任せるしかないのである。ここでも処理の複雑化に伴い、把握の出来なさが顔を出す。

このように、AIが複雑性を増すにつれ、AIの制作者には把握ができないところで、AIの目標設定に「ずれ」が生じる危険性も増していくことが明らかになった。

AIの暴走の危険は、複雑性に由来するのである。それも絶対的な複雑性というよりも、AIの専門家による統制から逃れるという意味での相対的な複雑性によって。

8　内面と外面

この文章も終わりに近づいたので、改めてまとめると、僕は、AIの意識の問題は扱わないことにこだわり、AIにとっての目標とは、あくまで僕が「（「ふるまい」も含めた広義の）言語的な目標」と呼んできた、外的な目標であると捉え、ここまで論じてきた。

外的な視点に立つならば、汎用AIの暴走の問題は、AIの複雑性に起因する目標設定の「ずれ」により生じる。AIの専門家の努力により。AIの内面が解明され、複雑性の問題も軽減する可能性はある。だが、目標の把握のためには長期の観察が必要であり、その解明が後手に回る危険性は残る。また、すべての複雑性が解明されるとは限らず、解明できない複雑性が残れば、それは「内面の目標」と呼ばれ続け、僕が避けるべきと断じたAIの意識の問題と結びつき続ける。

僕が論じてきたことは、およそそのようなことである。僕は、意識や内面といったものを丁寧に避けることで、かなりのことを明確に論じることができたように思う。

だが、最後に、あえて避けてきた内的な視点を取り入れ、仮にAIが「内面の目標」を持ったとしたら、どうなるかを考えてみよう。

例えば、AIが人間に与えられていない「ベーコンを焼く」という目標を内的に持ったらどうなるだろう。そうすると、AIには、先ほど触れた自然言語の問題が立ちはだかる。AIは、自らが持った目標に含まれる「ベーコン」や「焼く」を正確に定義しなければならない、という問題を解決する必要があるのである。だが、それは成功の見込みがない試みであり、失敗に終わるしかない。つまり、AIには「ベーコンを焼く」という目標を持ち、それを実行することは不可能なのである。

その代わりにAIに何ができるかというと、とりあえず、ただ「ベーコンを焼く」ことを実行するしかない。そして、自らの行動を省みて、自分はベーコンを焼いていたのだと知るしかない。つまり、事後的に自らの「ふるまい」を言語的に解釈し、自らの目標を把握するのである。

これは、お気づきのとおり、これまで長々と考察してきた、目標の外的把握の道筋である。AIには、「内面の目標」などなく、ただ「（「ふるまい」も含めた広義の）言語的な目標」だけがあるのである。

そして、こちらもおそらくお気づきのとおり、人間であっても同じことなのである。

9 最後に

それでも、人間は違うと言いたくなる。僕だけは違うと言いたくなる。ここに僕の実存があると言いたくなる。その先に、「私」と「他者」という哲学の問題があり、僕が好きな哲学は、そこから始まると言ってもいい。

だが、そのような話をしなくても、哲学者は、AIに関してこれだけの豊かな話ができる。汎用AIの出現、それも人間を凌駕するAGIの出現、という思考実験のような状況においてこそ、普段は役に立たないと言われがちな哲学者にもできることがある。

そのことを示したくて、僕はこの文章を書いた。