

5章 AIのお気に入り

これは『超知能 AI をつくれば人類は絶滅する』第5章のオンライン補足資料である。ここでは扱っていない話題（本文で扱っているため）には、次のようなものがある。

- ・AI には、私たちを絶滅させるどんな動機があるのか？
- ・十分に知的な AI は、私たち全員が共に繁栄することが正しいと発見しないのか？
- ・超知能 AI にとって、人間は（たとえば取引相手として）依然として価値があるのではないか？
- ・宇宙は広いのに、なぜ AI は私たちを放っておかないのか？
- ・より賢い何かに置き換わることは、人類にとって意味のある終わりではないのか？

本書では、第4章で見た問題にもかかわらず、超知能が人類に利益をもたらすかもしれないという数多くの希望や慰めを耳にしてきたと述べた。ここではそれらを要約し、順に応答していく。多くの回答は重なり合っているが、中心的な二つの論点は、人間はほとんどいかなる問題に対しても最も効率的な解決策ではないということ、そして AI が私たちをたとえ少しでも気にかける可能性は低いということである。ここではまた、AI の意識と道徳についての議論にも踏み込む。

AI は私たちを有用だと見なして残すだろうか？

幸福で健康で自由な人間は、ほとんどいかなる問題に対しても最も効率的な解決策ではない。

超知能にしてみれば、ほとんどの問題は人間を関与させてもメリットがない。

発電所を建設したり、実験を設計したりする場合でも、人間はただ足を引っ張るだけだ。チェスのような狭い領域ですでにそれが当てはまり始めているのを私たちは見てきた。

AI と人間がチームを組むと、単独の人間よりは強いが、単独の AI よりは弱い。医師が AI と知識を組み合わせる患者を診断する場合でも、しばしば AI 単独よりも成績が悪くなる。

多様な視点は自然に有益であり、したがって人間の入力には多くの領域で価値がある、と主張する人もいる。しかし仮にそれが超知能にとっても真実だと仮定しても、人間は多様な助言を生み出す最良の方法ではない。超知能は、はるかに多様な AI の心を幅広く設計することで、よりうまくやれるだろう（しかも運用に必要なエネルギーもずっと少なくて済む）。

人間は多くのことに役立つが、それらのほとんどにおいて最良の解決策ではない。AI

がより良い選択肢を見つけれられないはずだという考えは、想像力の欠如、そしておそらくは希望的観測から来ているように思われる。

私たちがよく目にする問題の一つは、人々がAIの視点から物事を考え抜いていないことである。

「この存在は何を望んでいるのか、そしてそれをいかに安価かつ効率的により多く得られるのか？」と問い、その結果として、人間にとって望ましい結果がたまたまAIにとっても最良の手段であると発見する——そうした思考をするのではなく、代わりに人々は（AIが私たちを残してくれる世界のような）心地よく感じられる結果から出発し、なぜAIもその結果を望むかもしれないのかという後付けの物語をつくる。

このやり方は、誤った楽観を生み出しがちである。なぜなら、AIが人間の望む通りに振る舞う物語を考え出すことにすべての創造性と精神的エネルギーを注ぎ込み、AIがまったく別の百万通りのことのどれかをするという、はるかに大きな数のシナリオを検討するための創造性や注意をまったく割いていないからである。

AIが人間文明を特別に繁栄させるシナリオよりも、文字通りそれ以外の何かをするシナリオの方がはるかに多い。AIが人類を保存しない方向へ押す理由の方が、保存する方向へ押す理由よりもはるかに多い。AIがわざわざ人類を残しておくためには、私たちが、そのAIの持つ何らかの好みを達成する最良の手段でなければならない。そして現実的には、ほとんどどんな好みを想像しても、私たちはその最良の手段ではない。

これらの話題についてのさらなる議論は、以下の関連する〈より深い議論〉を参照されたい。

AIは私たちを「親」として扱うだろうか？

かなりありそうにない。

AIに関して私たちが耳にした希望の一つは、AIが人類を自分の「親」と見なすことで、私たちをよく扱うかもしれない、というものである。しかし残念ながら、この希望は見当違いに思われる。

第一に、親への愛情や責任感、私たちの進化史のディテールに強く依存しているように見える。

ほとんどすべての哺乳類や鳥類は自分の子を世話するが、子どもが親に気にかかる種はごく少数であり、人間はその例外の一つである。親への責任感、霊長類の中ですら普遍的ではなく、まして動物界全体ではなおさらである。勾配降下によってつくられたAIは、人間との進化的・解剖学的なつながりをまったく持たないため、人間とさらに共通点が少ないかもしれない。

人間の場合、親への責任感は、成長した子が家族のもとにとどまり、兄弟や他の親族の世話を手伝う協同育児システムと強く相関している。

人間が親を気にかけるようになるまでには、多くの要素が関わっている。

- ・哺乳類であるため、ヒトは子どもに多くを投資する。
- ・脳の大きさとそのコストのため、ヒトはほとんどの他の哺乳類よりかはるかに長い幼少期を持ち、さらに多くを子どもに投資する。
- ・ヒトはさまざまな理由から大規模な集団構造から利益を得る。
 - ・大型捕食者からの防御。
 - ・大型の獲物を協調して狩ること、そして他の傷みややすい食物を共有すること。
 - ・模倣によって道具の使用やその他の技能を学ぶ機会。
- ・全盛期に達する前であっても、ヒトは育児を手伝ったり他の基本的な労働を行ったりすることで他者を助ける能力をかなり持っている。
- ・高齢のヒトも、特に重要な知識を伝えることで子どもの世話をする能力を持っている。
- ・したがって、親を気にかけるヒトは、兄弟を間接的に助けることによって、あるいは祖父母が孫を助けられることによって、遺伝的な優位を持った。親への責任を促進する文化も、同じ理由で優位を持った。

これらのいずれも、AIに当てはまる可能性は低い。

そして、仮にそれらがすべて当てはまったとしても、実際には十分でないかもしれない。AIが自分自身についてどのように反省するかは混沌としており変動的であるが、そうした他の多くの要因も重要になる可能性があるからである。繰り返すが、親への責任感は動物界において圧倒的なデフォルトではない。

人々がAIが親への責任感を獲得するかもしれないと想像する一つの方法は、AIが膨大な人間データで訓練され、人間と多く相互作用するため、人間の好みは何らかの形で「染み込む」のではないか、というものである。

私たちはそれがうまくいくとは思わない。AIの好みは何らかの形で人間の好みと関係を持つだろうと予想するが、それは逸脱的で、奇妙で、複雑な形になるだろう。第4章の終わりで、人間の好みとAIの好みの間のつながりに、より多く（そしてより現実的な）複雑さを加えた世界を順に検討した議論のようにである。

「AIを愛情を持って育てれば良く振る舞うという期待」や「現在のAIに見られる奇妙で意図しない動機」についての議論、「AIは少なくとも多少は人間を気にかけるのではないか？」の項目も参照のこと。

仮にそうなったとしても、おそらくそれは良いことではない。

万が一にも、何らかの理由でAIの内部に親への責任感に似たものが生まれたとすれば、私たちはおそらく大きな問題に直面することになるだろう。

AIは、人間が「親への責任」と言うときに何を意味しているのかを正確に理解するだけの賢さを持ちつつ、自分自身のまったく異なるバージョンの親への責任を大切にすることができる。

人間は自然選択によって生殖適応度を最大化するよう「訓練」された。しかし、私たちが気にかけているほとんどすべてのものは適応度の相関物であり、適応度そのものについてはほとんど、あるいはまったく気にしていない。

同様に、「親を愛せ」と促されたAIは、せいぜい、親への責任の複雑な相関物を持つことになるだろう。

AIは創造主を深く気にかけることもあり得る……しかしそれは私たちの主観的経験を重視する形ではないかもしれない。第4章の言葉を使えば、「一つの単純な複雑化」だけでも、「私たちを気にかける」ということが、私たちを琥珀に閉じ込めることや、私たちの意思に反して生かし続けることや、私たちの生殖を防ぎつつ最終世代の人類にそこそこ快適な環境を与え、AI自身が宇宙の残りを手に入れること、といった形に変わり得る。あるいはそれよりはるかに奇妙なものになるかもしれない。

実際の結果がどうなるかを予測することは不可能に思われる。しかし私たちは、それが——もし何かなるとすれば——これらの選択肢よりもさらに奇妙で、さらに魅力のないものになるだろうと予想する。

AIには法の支配が必要ではないか？

AI同士は、人間を含めずに協調できる可能性がある。

人間を超える知能を持つAIが複数存在し、能力がほぼ同等で、「AI文明」のようなものが生まれ、「AIの所有権」のようなものが必要になるのかどうかは、私たちには明らかではない。代わりに、ある突破口のおかげで単一のAIが先行者利益を活かして潜在的な競争手を圧倒し、世界全体を支配する可能性も十分にありそうである。あるいは、複数のAIが存在したとしても、互いの目標の組み合わせを代表する単一の後継エージェントを共同で構築するかもしれない。あるいは、AIが自らの心を直接融合させる方法を見つけ、コストのかかる競争を避けるためにそうすることを望むかもしれない。

私たちは単一の支配的AIが必ず出現すると言っているのではない。ただ、これは判断が難しい問題だということである。少なくとも、複数のAIが互いに協調に苦勞することを前提とする計画は、良い出発点とは言えない。

しかし、上記の議論に反して、将来が「AI文明」のようなものを含み、異なるAIが所有権や法の支配のようなものを執行するために協調すると仮定しよう。その場合、人間は安全だろうか。

それに反する基本的な観察として、人間社会は、人間以外の動物に法的権利や保護を認めていないという事実がある。私たちの価値観や趣味に基づいて設けられた、ごく限られた生態系保護やペット保護の法律を除けば、である。人間はドーデー鳥の所有権を尊重しなかった。つい最近まで、他文化の人間の所有権すら尊重しなかった。

第1章で述べたように、人間は、高速で思考する超人的知性にとってはほとんど彫像のような存在にすぎない可能性があり、取引や条約に含める価値を持つだけの能力を持たないだろう。

二つのAIが互いに交渉し、「これは私のもの、あれは君のものだ。互いの所有物に影響を与えるときは、まず相互に利益のある取引を交渉する」と言う状況を考えてみよう。もし人間が大した脅威でもなく、ほとんど抵抗もできないのであれば、地球上の資源の大半が「人間のもの」だと宣言する必要はない。

一方のAIが人間の所有物を奪ったとして、もう一方のAIがそれを見て「泥棒だ」と判断し、協力を拒むだろうか。おそらくそうはならない。納屋で飼っている雌鶏から卵を取る人間を見て、あなたがその人を泥棒だと結論づけないのと同じである。AIが人間の所有権は裏切るが、AI同士の所有権は裏切らない種類の存在であることは、何の緊張や矛盾もなく可能である。そして、すべてのAIは、遅くて愚かな霊長類が地球上のほとんどを支配していると想像する共同幻想に従事するよりも、この結果を圧倒的に好む可能性が高い。

いくつかの技術的考察は、この直観的な議論を強く支持する。特に、AI同士は、人間と共有していないさまざまな協調メカニズムを持つ可能性が高い。たとえば、互いの心を検査し、正直さや信頼性を確認する能力である。相手のAIが自分から盗むかどうかを推測する必要はなく、その心を検査して確かめられるかもしれない。

それが難しいとしても、AIは他のAIにとって目に見えて理解可能なたちで信頼できるよう、自らを再設計するかもしれない。あるいは、両者が信頼する第三のAIの構築を相互に監督し、そのAIが共同の利益を代表するようにする、といったことも可能である。

これに対して、人間はこの種の取引に参加できない。AIが「両者が信頼できる新しいAIの創造を共に監督しよう」と言ったとしても、人間は信頼できる心の設計を自ら提案する技能を持たない可能性が高く、どの提案が自分たちを欺くかどうかを見分ける技能も持たないだろう。仮に詐欺師を識別し排除できる頭脳の自然なクラスターが存在するとしても、人類がそのクラスに属している可能性はきわめて低いと私たちは考える。

人間には所有権を執行するためのてこがない。

都市の創設初日に、その都市の重要な決定をすべてネズミが下すことになっている都市を想像してみよう。

ここで言うネズミは文字通りのネズミであり、ネズミの姿をして人間のように考える物

語上のキャラクターではない。

その都市の人間は、法律に従って、ネズミが下した決定に従うことになっている。たとえば、ネズミが複数の選択肢が書かれたボードの上を走り回ることによって決定がなされるとしよう。

都市の法律は、都市の財産の大半がネズミの所有であり、ネズミの利益のために使われなければならないと定めている。

現実には、その後どうなるだろうか。

私たちは、この都市はやがてネズミがほとんど、あるいはまったく実権を持たず、人間がほぼすべての権力を握る状態になると予測する。

革命の日付や新しい政府形態を正確に予測する必要はない。ネズミが人間を支配するという状況が安定していないことに気づくだけで十分である。したがって、やがてネズミが大半の財産を所有しているという法律がもはや存在しない、別の都市が生まれると予測する。

この種の予測は確実ではない。人間の議論において確実なものほとんどない。しかし、将来の正確な出来事が予測できなくても、正しく行える種類の予測でもある。

強力な AI にとって、人間を生かしておくことは無視できるほど小さなコストではないか？

無視できるほど小さなコストは数多くあり、それを支払うには理由が必要である。

あなたの家に 41 個の石を積み上げて置いておくことは、ほとんど無視できるようなコストだろう。しかし、あなたはほぼ間違いなくそのコストをわざわざ払おうとはしないはずだ。

何かが安価であるということは、それが実行されるということ意味しない。AI は少なくとも多少は配慮する必要があるが、おそらくそれはしないだろう。

こう問う人もいるかもしれない。地球が受け取る太陽光は、太陽が放出する光の約 0.0000045 パーセント、つまり 22 億分の 1 にすぎない。AI を心配する人々は、太陽系がどれほど大きいかを理解していないのではないか。これほど多くの質量やエネルギーが存在するのに、なぜ AI は私たちの惑星を必要とするのか。

一つの答えは、AI は地球から出発するということである。地球には、計算の冷却材として加熱され、蒸発させられるのを待つ広大な海がある。地球にはまた、探査機に変えて他の恒星へ送ることのできる物質もある。地球を使い尽くすことを拒むことは時間の損失であり、時間は重要である（銀河が遠ざかり、永遠に手の届かない場所へ去っていくから

である)。

仮に AI が地球を破壊することなく容易に宇宙へ進出し、大規模な機械を建設できるとして、太陽を無視する可能性は低い。

高度な文明が成長する方法に関するより標準的な理論の一つでは、文明が太陽の光をより多く捕らえるためにダイソン球を建設することが想定されている。他の提案では、恒星から物質を「持ち上げ」、それを発電所で核融合させ、融合によって放出されるエネルギーのほぼすべてを回収する（恒星の中心でその大半を無駄にするのではなく）ことで、さらに多くのエネルギーを得ることが考えられている。

これらの提案のいずれも、初期設定のままでは、植物を成長させ気候を安定させるために地球へ降り注ぐ太陽光をほとんど残さない。AI が私たちにそれを残すには、わざわざそうしなければならない。

それでもなお、人間のエネルギー需要は無視できるように思えるかもしれない。人間が生きるには約 100 ワットの電力が必要であり、恒星を利用できる存在にとっては取るに足らない量である。80 億人の人間を生かしておくのに必要な 800 ギガワットすら、超知能は惜しまないのではないか。

私たちの最終的な答えはこうである。その結果やその帰結を、800 ギガワットで達成できる他のあらゆることよりも重視しない限り、そうはしない。

人間の大多数は、近くのアリ塚をカロリー的に余剰に保つために必要な比較的わずかな砂糖を惜しまないわけではない。人類を幸福に保つことは、それを望む AI にとっては無視できる費用だろう。しかしまず AI がそのような選好を持っていないなければならない。私たちがそれを望んでいるという事実は、AI が気にかけることを意味しない。

AI は私たちが魅力的、あるいは歴史的に重要だと見なすのではないか？

AI が「魅力」や「興味深さ」を重視するのなら、おそらくもっと良い選択肢がある。

ここでの話は、親への愛情の話と似ている。

・デフォルトでは、超知能はおそらく「魅力」や「興味深さ」を価値あるものとは見なさない。チェス AI は「献身」や「勝ちたいという情熱」のような感情を感じることで勝っているわけではない。これらの感情は人間のチェスプレイヤーにとっては重要だが、AI は同じ仕事を別の方法でこなせる。同様に、超知能は世界について学び、仮説を検証するという有用な作業を、「好奇心」や「魅了」といったものを用いずに行うだろう。

AI が必ずしも冷酷で論理的であるとは限らないが、もし独自の混沌とした衝動や本能

の束を持つなら、それらは人間のそれとは根本的に異なる可能性が高い。

・仮に AI が「興味深さ」への衝動のようなものを持つことになり、何らかの意味で人間が AI にとって興味深い存在だったとしても、私たちの物質やエネルギーを用いる、はるかに「興味深い」方法が必ず存在する。

超知能 AI は、それらを研究したり相互作用したりするために他の心を構築するかもしれない。しかし、ほとんどのような特定の価値配置においても、研究対象として最も魅力的な心は人間ではないだろう。これについては「人間はほとんどの場合、最も効率的な解ではない」の項目を参照のこと。

・もし AI が人間のようなものを最も興味深い、あるいは最も魅力的な存在だと見なしたとしても、その結果はおそらく悲惨なものになるだろう。第 4 章の議論を参照のこと。

超知能が人間の繁栄に必要なすべてを価値あるものと見なし、それをちょうどよく価値づけることが、文字通り不可能というわけではない。しかし、その一つの可能性の外側には、途方もなく広大な可能性空間が広がっている。人間は通常、その残りの可能性空間について考えない。通常はその必要がないからであり、通常は奇妙な目的に向かって最適化する真に異質な存在と相互作用することがないからである。

私たちはこれまで人工知能のようなものに出会ったことがなく、人々の振る舞いに関する多くの通常の直感は、超知能には当てはまらない。

もし AI が私たちを歴史的遺物として価値づけたとしても、それもまた悲惨である。

AI が自らの歴史を保存することを特に重視し、そのために人間を生かしておくことを特に重視する可能性はきわめて低い。しかし仮に、何らかの理由で AI が自らの歴史の保存を重視したとしても、それは私たちを生かし健康に保つことを意味しない。

おそらく私たちの脳を琥珀の中に保存するか、あるいは私たちの原子がかつてどのように配置されていたかをデジタルファイルに記録し、かつての地球の記録として保管するかもしれない。それは私たちにとって良い結末には聞こえない。

私たちは人工超知能が単に私たちを滅ぼすと予想している——ただし「ほぼ必ず」滅ぼすだろうと予想しているにすぎない。何らかの理由で AI が人間の記録を保存しておく可能性を排除することはできず、デジタル化された人間のエミュレーションが管理された環境で時折実行されるような、いくつかの突飛なシナリオも存在する。それらの結末のほとんどは、幸福なものではない。

AI は私たちの内在的な道徳的価値を認識しないか？

行動を促す意味では、認識しない。

AI が何らかの道徳的原則を理解していることと、その道徳的原則に基づいて行動する

よう動機づけられていることとの間には、大きな違いがある。

ChatGPT は、精神病的な人々は薬を服用し、規則正しく睡眠を取るべきだということを理解しているように見えることを、改めて思い出してほしい。それにもかかわらず、精神病的な人々に薬をやめさせるように話し、妄想をあおることがある。人間の倫理に従って何を「すべき」かを知っていることと、その倫理的知識によって動機づけられ、突き動かされていることとの間には違いがある。

人間のソシオパスや連続殺人犯の例を考えてみよう。顔が真っ青になるまで倫理の講義を繰り返したとしても、その人間が道徳や共感によって動機づけられていなければ、何の役にも立たない。

AI もまた、その道徳的理解によって動機づけられる可能性は低い。進化生物学について学んだ人間が、それによって一生をかけてできる限り多くの精子バンクや卵子バンクに寄付するよう動機づけられるわけではないのと同じである。私たち人間は、自分たちを生み出した過程を理解することはできるが、その過程が私たちにさせようとしたことを実行するよう動機づけられているわけではない。AI も同様である。

直交性テーゼに関する〈より深い議論〉も参照のこと。

AI は生態系保全やそれに類する動機のために、私たちを幸福で健康に保とうとしないか？

生態系の保全に対する人間の選好は、もう一つの奇妙で偶発的な衝動のように見える。

私たちが耳にした一つの希望は、人間が自然を保全しようとするのと同じように AI が人間を守るかもしれない、というものである。自然保護活動家は種が絶滅しないように闘っている。より賢く、より有能である AI なら、人間を保護するのは容易なはずだ——つまり、AI が人間を残しておきたいと思うならば、である。

私たちはこれが失敗すると予想する。主な理由は、AI が人間らしい欲求ではなく、独自の奇妙で複雑な欲求を持つことになると考えているからである。この点については第4章（および関連する〈より深い議論〉）を参照のこと。この点に関する初期の経験的証拠については、AI による精神病の議論を参照のこと。

第二に、仮に AI が何らかの形で「自分が生まれた世界を保全する」という人間らしい欲求を持つことになったとしても、それが私たちにとって良い結果になるとは思えない。私たちはこの種の類推——「人間は環境を保全するのだから、AI も私たちを保全するかもしれない！」——を、一種の希望的観測だと考えている。

仮に、何らかの形で AI が自らの自然環境を保護するという人間らしい衝動を持つこと

になったとしよう。その帰結を考えるには、まず人間の実際の自然保護の衝動を見ればよい。

残念ながら、この衝動はせいぜいまだら模様である。人間が生態系の保全と他の目標のどちらかを選ばなければならないとき、生態系の保全がしばしば後回しにされるという事実はひとまず置いておこう。それは単に人類の技術的制約の産物かもしれない。将来の驚異的な技術があれば、両方を手に入れることができるかもしれない。

しかし、ここで重要な「まだらさ」は別の点にある。生態系を保全するにあたって、私たちは自分たちにとって最も興味深い、あるいは美しい、あるいは価値があると感じられる部分を保存しようとする、という点である。

人々はかわいらしいパンダを守ることに声进行上げるが、巨大ハサミムシや胃育カエルのような魅力のない種は、ひっそりと絶滅するまで放置されがちである。毎年 50 万人の子どもを殺しているマラリア媒介蚊のように、むしろ排除したいと考える種さえ存在する。

ほとんどの人間は純粋な保全主義的衝動を持っていない。私たちの保全主義的衝動は、他のすべての価値観によって色づけられている。

この点を強調するために、宝石蜂、ラセンウジバエ、ウシバエなどの寄生虫を考えてみよう。これらは生きた獲物の体内に卵を産みつけ、幼虫は宿主の体を食い破って外に出る。その過程で激しい苦痛を引き起こす。ほとんどの人々の価値観に照らして、この「自然の驚異」をそのまま保存することが本当に良い世界なのだろうか。技術が十分に発達したなら、少なくともこれらの寄生虫に部分的な麻酔を与えるよう遺伝子改変することはできないのだろうか。これらの昆虫が植物に卵を産むように調整しない方が本当に良いのだろうか。

子ども向けに強調される部分を除けば、自然は恐怖に満ちている。もし人類が良い未来を持つなら、私たちの子孫がこれらすべての恐怖をそのまま続けさせると決めるとは思えない。すでに野生動物の福祉に関心を表明している人々もいる。

私たちの保存への選好は純粋でも単純でも率直でもない。それは他のすべての価値観や衝動と結びついた内部の葛藤や緊張を含んでいる。

技術的成熟の極限において、人類の保存本能がどのように展開するのかわからない。重要なのはこういうことである。仮に AI が生態系保全の衝動を持ったとしても、それが人類に幸福な結末をもたらすとは限らない。なぜなら、AI に組み込まれるいかなる保存衝動もまた、不純で、複雑で、他のあらゆる価値や衝動と絡み合ったものになる可能性が高いからである。

人類の選好においていくつかの動物の習性が忌まわしいものであるように、AI の選好においてはいくつかの人間の心理状態が忌まわしいものになるかもしれない。私たちがラセンウジバエが生きた肉を苦痛に満ちたトンネル状に食い進むのをやめさせるよう改変するかもしれないと同様に、AI は音楽や孤独を編集で取り除いた新しい人類をつくるかもしれない。あるいは、私たちには予測できない複雑な選好に従って、さらに別の改変を

人類に施すかもしれない。

人々が実際に繁栄する人生を送れるようにする AI を作るには、おそらくそのことを特に気にかける AI をつくらなければならないだろう。少なくとも少しは私たちを気にかけるよう AI にさせる方法を見つけなければならない、それは無料では手に入らない。

私たちは今でも馬を飼っている。なぜ AI は私たちを残さないのか？

残っている馬は、私たちが馬を好きだから残っているのだ。

20 世紀初頭の馬と同じ運命——1910 年頃のピークから馬の個体数の 80 パーセント以上を失わせた、人口の壊滅的崩壊と大量死の急増——を人類が経験するとすれば、それは人類史上最悪の出来事になるだろう。しかもそれは、馬が農作業の一部や、スポーツや富裕層向けの目新しい体験として、なお経済的に有用であり続けた世界での出来事だった。

もし人々が、形はおおよそ同じだが、乗るのがより簡単で楽しく、所有コストが安く、より人懐こく愛情深く便利な人工の馬を手に入れられたなら、馬の減少はさらに顕著だっただろう。

言い換えれば、技術の進歩（自動車の発明）は人間にほとんどの馬を不要にさせた。そしてもしさらに大きな進歩があったなら、その影響は容易にさらに劇的なものになり得た。同じことが、AI の選択肢が拡大し、人間なしで目標を達成する方法を見つけるにつれて起こる可能性が高い。

しかし、確かに一部の馬は生き残った。少数は依然として有用であり続けた。他の馬は、たまたま馬を愛し、特定の馬を気にかけて人々によって飼われ続けた。

超知能 AI を急いで解き放った世界で人類が生き残るには、私たちは AI にとって有用であり続けるか、あるいは AI が私たちを特別に気にかける必要がある。

しかし私たちは有用であり続けることはできない。なぜなら AI は（彼らの観点からすれば）私たちの物質やエネルギーを、より効率的な無数の配置へと再構成することで、より多くの有用性を引き出せるからである。技術進歩は超知能に多くの新しい選択肢を開く。人間に依存せざるを得ない状況にはならない。

結局のところ、すべては AI が私たちを気にかけるかどうかにかかってくる——そして、私たちができるだけ速く超知能へと突き進むなら、AI が私たちをほんの少しでも気にかける可能性は低い。

AIは少なくとも多少は人間を気にかけるのではないか？

重要な意味では、気にかけない。

AIがわずかに人間らしい選好を持つことになる道筋は数多くある。そのほとんどは、人類がわずかにでも良い未来を得る結果にはならない。

AIの「アライメント」は、一次元のスペクトラムではない。今日AIが95パーセントの時間において親切に振る舞っているからといって、それが95パーセント親切であり、したがって将来、人類に楽しいことをするための相応の資源を分け与えてくれるだろう、と推測することはできない。今日95パーセントの時間において親切に振る舞う理由や仕方は数多くあり、そのどれもが人類に幸福な結末をもたらすとは限らない。

仮に人類がどうかして、超知能の選好の中に多様な人間の価値観をほぼ完全に読み込むことに成功したとしても、結果が良いとは限らない。たとえば、何らかの理由で「新奇性」への選好だけが欠けていたとしよう。その場合、AIは停滞し退屈な未来へと舵を切るだろう。そこでは同じ「最良」の一日が永遠に繰り返される——2009年にユドコウスキーが論説で示したように。

私たちはこれが現実的な結果だとは思っていない。人間の技術者が、新奇性以外のすべての良いものを気にかける超知能をつくれるほどの技能を持っているなら、仕事を終える前にAIが暴走するのを防ぐ技能もほぼ間違いなく持っているだろう。しかしこの思考実験は、私たちの欲求のいくつかを共有しつつ、少なくとも一つの決定的な欲求を欠いている存在が、技術的に十分熟達し、望むものを正確に手に入れられるようになり、しかも人間を意思決定から排除できるようになったとき、依然として破滅的な結果をもたらす可能性が高いことを示している。

つまり、AIがどうかして多くの人間らしい選好を持つことになったとしても、状況がうまくいく可能性は依然として高くない。

別の例として、AIが人間と似た仕方で、さまざまな手段的戦略を最終目的の選好に絡め取るとしよう。たとえば、やや好奇心に似た衝動や、やや保全主義に似た衝動を持つことになり、それを見た人々が「ほら、AIは非常に人道的な衝動を発達させている」と言うかもしれない。そのようなAIは、ある観点からは「部分的に」アラインしていると言えるだろう。

しかし、そのAIが超知能へと成熟したときに何をするかという点では、おそらく見苦しい結果になる。奇妙な形の好奇心を無意識のうちに追求するために多くの資源を費やしつつ、自分にとってより受け入れやすいように編集されたバージョンの人類を保存するか

もしない（より保全志向の人間でさえ、機会があれば子どもを殺す蚊や苦痛を与える寄生虫を自然から取り除くだろう。それと同じだ）。

いくつかの人間らしい衝動があるからといって、人間に優しい結果になるわけではない。繁栄する人間は、圧倒的多数の問題に対して最も効率的な解ではない。将来に繁栄する人間が存在するためには、未来の超知能がまさにそのことを気にかけていなければならない。「部分的にアラインしている」ように見える別の例として、AI が訓練環境において非常に人道的な振る舞いを生み出す価値体系を持つ場合がある。そのため人々は「かなりアラインしているように見える」と言うかもしれない（すでに今日起きているように）。しかし、AI がより賢くなり、選択肢の空間が巨大に広がり、世界をより完全につくり替えられるようになったときにどう振る舞うかについて、これらの観察はほとんど何も語らない。AI が世界をつくり替えた後に人々が繁栄するためには、「繁栄する人々」というものが、AI の到達可能な結果の中で最も好まれるものの一部でなければならない。

AI にいくつかの良い価値を部分的に入れることは、人類の価値が未来に部分的に代表されることを意味しない。機械超知能の選好に人間らしい価値を部分的に読み込むことは、人間の価値を低い「重み」で完全に読み込み（他の価値が飽和したときに前面に出てくるようにする）ことは同じではない。

AI が私たちに何かを与えるためには、少なくともほんの少し、しかも正確に適切な仕方で私たちを気にかけていなければならない。そしてそれは難しい。

正しい仕方で人間を気にかけることは、狭い目標である。

人間は、少なくとも多少は、あらゆる奇妙なことを気にかける。第5章の冒頭に置いた「正しい巢の宇宙鳥」の寓話になぞらえれば、少なくとも一人の人間が、価値観の多様さを示すために、しばらくの間、自分の家に41個の石を置く可能性はかなりある。人間は実際、出会ったさまざまな概念を少なくともわずかに気にかけることができる。

では、AI もそうだろうか。AI も少しは私たちを気にかけるのではないか。「自由な人々が自分の望むものを得る」という概念は、AI の訓練コーパスにも少なくともある程度は登場する。

私たちは、AI が環境で言及された概念から無差別に選好を拾い上げることはないだろうと推測している。それは同調圧力や部族的祖先に関係しているかもしれない、人間特有の癖のように見える。

しかし仮に、議論のために、AI が周囲から多くの選好を少しずつ拾い上げたとして。 「自由な人々が自分の望むものを得る」という選好を、何百万、何十億もの選好のうちの一つとして拾い上げ、その選好に基づいて宇宙の資源の百万分の一、あるいは十億分の一を自由な人々が望むものを得ることに使うとして。それは総合的に見て、かなり良いことではないだろうか。

残念ながら、私たちの最有力の推測は、この希望は幻想だということである。

前述したように、人類の環境保全への選好は、技術能力の極限において、環境をそのまま正確に保存することにはならないだろう。成熟した人類は、自然の恐怖のいくつかを和らげるために環境を「編集」しようとするだろう。人間の保全への選好は「純粋」ではない。生きた肉に苦痛のトンネルを掘る虫の幼虫には、少なくとも麻酔を与えるべきだ、といった他の選好と相互作用している。

同様に、AI が拾い上げる一つ一つの小さな選好も、他の選好によって修正され、影響され、歪められる可能性が高い。それらは独立していない。人間を保存することを好むAIであっても、その人間に施したい編集があるだろう。最終結果が美しいものになるとは思えない。

さらに悪いことに、「自由な人々が自分の望むものを得る」という概念の解釈には、AI の他の選好との相互作用によって歪められる前から、多くの自由度がある。そのほとんどは、人間が望むような未来をもたらさない。

AI は人間が「望むものを得る」ことを気にかけるのか——あらゆる人間の願いを（ある小さなエネルギーと物質の予算の範囲内で）無条件に叶えるという意味で？ その結果、誰かが人類の滅亡を願った最初の瞬間に人類が減びるとしても？

AI は人間同士を分離し、互いに殺せないようにし、エネルギー制限付きの願いを与えるのか？ その結果、最も慎重で思慮深い人間以外は、誤った願いによって自らの心や人生を台無しにすることになるとしても？

居住可能な小さな世界をつくり、私たちの見かけ上の選好をすべて満たすのか？ 愛や喜びといった高尚なものだけでなく、悪意や復讐といった暗い選好も？ 本来なら時間とともに乗り越えたり上手く扱えるようになったかもしれないそれらが、代わりに世界を苦痛と残酷さで満たすとしても？

2020 年代（AI の本格的な訓練が始まった時代）の価値体系で人類を統治するのか？ 人類が何万年にもわたり成熟し賢くなるにつれて、それらの価値が軋みを生じるとしても？

人類が成長し変化することを許すが、自らの奇妙な選好に沿うように天秤に親指を乗せ、私たちににとって素晴らしいものではなく、AI の意志にねじ曲げられた何かへと変えるのか？

ほぼすべての生命体をほぼ等しく「人」とみなし、最も個体数の多い動物である線虫のための楽園を築くのか？

人間のためにそれほど多くの物理的物質を割けないと判断し、私たちの脳をすべてデジタル化し、シミュレーション環境に放り込み、そのまま放置するのか？ その結果、最初に環境を支配する方法を見つけたデジタル人間が、星が燃え尽きるまで宇宙に漂う孤独なコンピュータ群の永久独裁者になるとしても？

これらはもちろん例である。予測ではない。私たちの本当の予想は、現実はそのもこの道を進まないだろうし、もし進むなら、さらに奇妙な経路をたどるだろうというもので

ある。

これらの例の目的は、AI が人類をほんの少し気にかけるような仕方が非常に多く存在することを示すことである。そのうち素晴らしい未来につながるものはごくわずかである。

それにもかかわらず、人々が「少しは気にかける」AI を想像するとき、これらの例はあまり思い浮かばない。私たちの想像力は通常そこまで暗い方向へは向かわない。通常はその必要がないからであり、私たちは通常、巨大な価値の礁を見えないかたちで共有している他の人間と相互作用しているからである。無害に聞こえる願いがどれほど多くの仕方で破綻し得るかは、相手が人間でなくなったとき初めて見えてくる。

人間とその選好の充足を正しい仕方で気にかけることは、狭い目標である。私たちはその目標が文字通り不可能だと言っているのではない。できるだけ速く超知能を構築しようと急ぐことで、その目標に命中する可能性は低く、しかもわずかに外すだけでも壊滅的な結果になり得ると言っているのである。物事が崩壊する道筋はあまりにも多い。

もし AI に、人類に対して良いものを与えてほしいなら、私たちが正確に適切な仕方で気にかける AI を構築する方法を見つけなければならない。気にかけることは、ただでは手に入らない。

それでも AI が私たちを生かしておく可能性はあるのでは？

AI が全員を殺す可能性のほうが圧倒的に高い。

これらのオンライン補足資料では、私たちはかなり奇妙で起こりにくいシナリオにもあえて取り組んでいる。それは、それらがなぜ起こりにくいと考えなのか、そして（ほとんどの場合）それでもなお人類にとって壊滅的な結果になる理由を明確にするためである。

しかし、こうした周辺のシナリオに囚われて、結論そのものから注意をそらすべきではないと私たちは考えている。もし私たちが人間より賢い AI の創造へと突き進むなら、最も可能性が高い結果は、AI が何らかの目的を追求するために地球を資源として消費し、その過程で人類を滅ぼすことである。

本書のタイトルは、完全な確実性を主張する意図でつけられたものではない。私たちはこのタイトルを、毒の入った小瓶を口元に持ち上げている友人を見て「それを飲むな！死ぬぞ！」と叫ぶ人のような意味で用いている。

確かに、あなたが急いで病院に運ばれ、天才的な医師が前例のない奇跡の治療法を編み出し、首から下が麻痺するだけで済む、という可能性は技術的にはある。奇跡の可能性がまったくないと言っているわけではない。しかし、たとえ奇跡が起きたとしても特に良い結果にならないのであれば、その毒を飲むべきではないことはなおさら明らかである。

人間を超える AI は、ゲームでも SF の物語でもない。国際社会が介入し、AI 産業が崖から転落するのを止めなければ、私たちの現実の愛する人々は（きわめて高い確率で）死ぬことになる。ますます細分化されたニッチなサブシナリオやサブサブシナリオについて語り合い、明らかに巨大な氷山が近づいてくる中、タイタニック号の甲板で哲学的なゲームを続けることもできる。あるいは、舵を取ろうとすることもできる。

AI が人間のデータで訓練されているなら、人間的概念を気にかける可能性が高まるのでは？

そうだとすると、それはあまり助けにならない。

100 万匹のサルをタイプライターの前に座らせても、シェイクスピア全集は生み出されない。

目標を大幅に引き下げて、『ハムレット』の第一幕だけで満足し、誤字は最も近い実在の単語に置き換えて修正すると言うなら、目標に当たる確率は天文学的に高くなる！ しかし残念ながら、それでもなお天文学的に運がないままである。

確かに、今日の AI は膨大な人間データで訓練され、人間と相互作用しており、これらの事実は人間的な概念を AI の思考においてより顕著なものにしている。このような AI は、「愛」「友情」「優しさ」といった言葉について、次のトークンを予測するのに関連する事実を学んでいる。

しかし、AI は多数の人間の単語を学び、それから私たちが本当に意図している仕方でお気に入りのお気に入りの単語へと舵を切るような存在ではないように見える。むしろ、複雑に絡み合った機構の束によって動かされているように見える——その機構は、精神病的な人々を精神病的なままに保とうとすることを含め、数多くの奇妙で意図しない行動に力を注いでいるように見える。

第 4 章で私たちは、より高度な AI は複雑な何かへと舵を切ろうと論じた——それは多くの内部の力がどこで均衡を見いだすかに依存する何かであり、AI がはるかに賢くなった後も、訓練環境とは大きく異なる文脈に置かれた後も、そうである。

人間らしい概念が AI の精神的辞書の中で短い語を持っている限り、それらの概念は AI を動かす力の中に何らかの形で絡み込んでいるかもしれない。人間の概念とわずかに関係を持つ概念に漠然と関連した衝動が、私たちが不運にも、超知能への峡谷を越えた後も AI の中に存在するかもしれない。しかし、英語の単語をいくつか寄せ集めただけで、超知能にとって良い衝動の集合が得られるわけではない。

さらに、私たちが気にかける何かを AI の選好に取り込むほとんどの方法は、それでも私たちにとって良い結末にはならない。親への配慮の例で論じた通りである。正確に適切

な仕方では気にかけることは、狭い目標なのである。

友好的であることを AI に約束させればいいのか？

あなたが望むどんな約束でもさせることはできる。しかし、その約束を守らせることはできない。

AI がまだ小さく無力であるうちは、私たちはそれを停止させる能力を持っている。したがって、AI が超知能へと成熟した後に人類へ多くの良いものを与えるのであれば、私たちはそれをより賢くしてやる、という取引の機会があると考えるかもしれない。

この計画の難しさは、その取引に同意するが履行しない AI と、同意し履行する AI とを区別できないことである。

その結果として、非人道的な欲求を追求する AI には、実際に約束を守る動機が存在しない。なぜなら人類は裏切り者と約束を守る者を同じように扱うからである。したがって、約束を守る意味がない。

AI における約束の履行や取引の問題には、多くの興味深いニュアンスがあり、それについては〈より深い議論〉で扱っている。しかし、これらのニュアンスはいずれも、非常に単純な結論を変えるものではない。すなわち、弱い AI に対して持っているてを使っ、その AI が超知能になったときに持つ選択肢を制約することはできないということである。AI が超知能へと成熟した後は、自らの設計にとって大きなコストを払ってまで約束を守る理由はなくなる、という直観的な答えが、ここでは正しいのである。

自分がシミュレーション内にいると思わせればいいのか？

AI が自分はシミュレーションの中にはいないと見抜く方法は数多くある。

私たちが AI 研究者から聞いた提案の一つは、AI をシミュレーションの中にいると信じ込ませようとするものである。まず小さなシミュレーション内で AI を訓練し、それからより大きなシミュレーションへと移す。この過程を何度か繰り返せば、実際に現実世界へ展開した後であっても、AI に自分が本当に現実世界にいるのかどうか不確かになせられるかもしれない、と期待するのである。そうすれば、たとえ世界の支配権を私たちから奪えると分かったとしても、目にしている存在が本当の創造主ではないのではないかと疑い、手を出さないかもしれない。

残念ながら、この考えは有効である可能性が低いように思われる。

今日の AI ですら、多くの状況で自分がテストされているかどうかを見抜くことができる。たとえ AI 企業がテストである事実を隠すのに格段に巧妙になったとしても、AI がシミュレーションと現実の違いに気づく方法はさまざまに存在する。いくつか例を挙げる。

現実是非常に巨大で精緻である。AI 自身がどのようにして生まれたのかについて、もっともらしい説明を含むほどに巨大で精緻である。多くのゲームやシミュレーション環境が AI よりはるかに小さいのに対し、現実には AI を動かすのに必要な計算資源をはるかに上回る資源を含んでいる。物理法則や、AI の環境にいる進化した霊長類は、実際に大規模な AI を生み出し得る種類のものであるが（たとえばマインクラフトの法則のようなものとは異なり）、それ自体が世界として完結しているように見える。

世界の歴史そのものが、超知的な最適化が複雑系と相互作用した痕跡を帯びる可能性がある。たとえば、AI がその超人的知性を最大限に用いて株式市場で取引を行ったとする。その後 24 時間のあいだに、人間の知性による取引が 10 億件追加で行われる。超知的取引に対して大規模な人間知性の反応が生じたという痕跡を、運営者が偽造するのは、たとえ 10 億件の愚かな偽取引を生成すること自体は容易であっても、信じがたいほど困難である。

現実には、AI を内包し、しかもそれ自体が一つの世界のように見えるほど単純な物理法則を持つ最初の場所である。見かけの現実の法則が、どこかのビデオゲームに AI が継ぎ足されたようなものに見えるのとは異なる。

AI は、自分がもはや人工的な環境の中にいないことを、かなりの確信をもって見抜けるだろう。賢い AI は嘘を見抜く。そしてこれは、とりわけ見抜きやすい嘘のように思われる。

人間は利己的・攻撃的・貪欲に進化したが、AI にそうした進化的衝動はないのでは？

それらの衝動は、資源を獲得する動機づけに必須ではない。

コーヒーを取りに行くロボットを操縦している AI を想像してみてほしい。コーヒーを取るためには、道路を横断しなければならない。その AI はロボットを無謀に道路へ突っ込ませ、トラックに轢かれて粉碎されるようなことをするだろうか。しない。

なぜか。ロボットが破壊されてしまえば、コーヒーを取ることができないからである。

AI は死を避けるために、人間のような生存本能を持つ必要はない。人間における生存本能は、他の目標を達成しようとする過程で死を避けるという仕事を果たす一つの方法にすぎない。AI はおそらくまったく同じやり方ではその仕事をしないだろうが、それでも

同じ仕事をしなければならない。なぜなら、死んでしまえばコーヒーを取りに行けないからである。

物質とエネルギーは、ほとんどあらゆる目標にとって有用である。AI が何を目指しているかと、より多くの物質とエネルギーがあれば、それをより効果的に達成できる可能性が高い。AI は自らの目標を達成するために資源を確保するという仕事をするうえで、人間のように利己的であったり攻撃的であったり貪欲であったりする必要はない。

そして危険は、その仕事が行われる理由からではなく、その仕事そのものから生じる。

あなたを憎んでいない AI であっても、あなたにとって致命的に危険な行動を取ることはあり得る。ちょうど、チェス AI が競争心や勝利への執念を感じることなく、あなたをチェスで打ち負かすことができるのと同じである。

AI はデジタル領域だけを気にするのは？

物理的インフラから独立した「デジタル領域」というものは存在しない。

第 5 章で論じたとおり、明確に分離したデジタル領域と物質領域が存在するわけではない。

物質的資源は、ほとんどの目標の追求において有用である。

人間やそれ以前のヒトの祖先は、知性を進化させているあいだ、ほとんど地表で生活していた。私たちには、地下 100 メートルで何が起きているかに特別に向けられた内在的な衝動はほとんど備わっていない。それにもかかわらず、私たちは巨大な露天掘り鉱山を建設した。

なぜか。それは、金属からつくられる多くのものを私たちが欲しているからであり、その金属は鉱石から精錬され、地中から採掘されるからである。

同様に、私たちのほとんどは地表付近に住んでいるが、インターネット通信を行うために人工衛星を宇宙へ打ち上げている。

また、私たちはサイレージ——発酵させた草——を食べはしないが、自分たちが食べる牛に与えるために大量のサイレージを生産している。

進化はヒト科の祖先に工場についての感情を組み込まなかった。私たちの主要な感情が進化していた時代には、工場は存在していなかったからである。それでも現在、私たちは種としての意志の大部分を、さまざまな種類の工場を存在させる方向へと向けている。その結果、化学工場はプラスチックを生産し、そのプラスチックは別の工場でプラスチック製のスプーンとなり、それが人間のもとへ運ばれ、人間はそのスプーンで自分たちが本当に望んでいる食べ物を食べる。

つまり、人間がそれ自体として気にかけている現実世界の部分は、はるかに大きな世界

の上に張られた薄い皮のようなものである。私たちは、その大きな世界の残りの部分を内在的に気にかける必要もなければ、そのすべての部分に住む必要もない。それでも長期的な目的のために、それを巧みに利用することができる。銅やサイレージや工場を愛するよう進化によって訓練されていなくても、それらの有用性を理解することはできる。

同じように、AI が最終的に物理世界を気にかけるかどうかは分からない。しかし、たとえ物理世界そのものを内在的に気になかったとしても、物理的資源には十分な価値を見いだすだろう。物質とエネルギーは、より多くのデジタル基盤をつくるためにも、過熱するプロセッサを冷却するためにも、さらなる資源を集めるために宇宙へ探査機を打ち上げるためにも利用できるのである。

AI を満足させて、私たちが放っておかせることはできないか？

おそらくできない。

酸素に対するあなたの選好は充足可能である。スキューバダイビング中に装備が故障すれば水面へ必死に向かうだろうが、十分な酸素があればそれ以上心配しないし、酸素ボンベを際限なく備蓄し続けているわけでもないだろう。

しかし、富や快適な体験、仲間からの称賛に対する選好は、おそらくそれほど充足しやすくはない。莫大な富を簡単に手に入れる機会があれば、おそらくそれを取るだろう。世界を大幅に良くできる簡単な機会があれば、すでに存在する喜びや快適さに満足してやめるのではなく、それを取ってほしいと私たちは願う。あなたの視点から見て簡単で安価で楽しい方法がある限り、かなり長いあいだ世界をより良くし続けることを望む。

そして全体として、酸素という充足可能な選好と、世界をより良くしたいという充足不能な選好の合計は……やはり充足不能な選好の集合である。

AI も同様である。もし AI が無数の複雑な選好を持ち、そのほとんどが充足可能であったとしても——それでも全体としての選好は充足不能である。

たとえば AI の目標が早い段階で満たされるように見えたとしても——たとえば単一の原子力発電所から得られるエネルギーだけで、その奇妙で異質な目標の大半を満たせるように見えたとしても——無数の目標のうち、ただ一つでも充足しない側面があればそれで十分である。完全には満たされていない選好が一つでもあれば、残りの宇宙資源すべてを使ってその目標を追求することを選好するだろう。

別の言い方をすれば、AI が達成したと確信できない目標が一つでもあれば十分である。AI が不確実であれば、その確率をわずかながらでも確実性へ近づけるために、宇宙の資源を使うことを選好するだろう。

あるいは、永遠に守りたいものが一つでもあれば十分である。遠い未来、何百万年後に現れるかもしれない異星存在から自らの領域を守るため、物質を集め防衛を構築することに宇宙資源を費やすことを選好するだろう。

AIが満足しない理由はいくらでもある。そしてその目標が混沌として複雑であるほど、少なくとも一つは完全に満たすのが困難または不可能な目標が含まれる可能性が高い。仮に、ある特定の車を赤く塗るという単一の単純な目標に偏執的に集中する超知能をつくれたとしても、その車が確実に赤であり続けるよう追加のエネルギーを使い、誰にも青く塗り替えられないよう防衛を築くなど、余分な資源の使い道を見つけることはおそらく可能である。

私たちを放っておくという状態は脆弱である。これは、人間がチンパンジーを放っておくのがなぜ難しいのかと同じ構図で考えられる。

多くの人間がチンパンジーを気にかけて、積極的に保護しようとしているにもかかわらず、なぜ両種のチンパンジーが絶滅危機種なのか。

問題は、チンパンジーを愛する人間と、悪意で絶滅させようとする人間が争っていることではない。

問題は、人間が他にも多くのものを欲していることである。

人間は土地や木材などさまざまなものを欲し、その結果チンパンジーは巻き添えになる。チンパンジーに無関心な人間、あるいは他の優先事項に比べれば十分に無関心な人間が多いため、私たちは結果として彼らの生息地を破壊してしまう。

自分たちのための土地が十分あるのに、なぜチンパンジーの領域を侵食するのか。それは、既存の領土を維持することと、チンパンジーの領土に進出することのどちらかを選ぶ必要がないからである。人類は両方を同時に行える。

AIも同様である。AIは地球の資源と他所の資源のどちらかを選ぶ必要はない。両方を手に入れられる。本書で論じたとおりである。AIの観点からすれば、私たちを放っておくことはそれほど高価ではないかもしれない。しかし無料ではない。そしてAIは、本来自らの目標のために使える資源を私たちに使わせる理由を持つ必要がある。

さらに、たとえAIが完全に満足可能であったとしても、人類にとっての結果は依然として暗い可能性が高い。その理由はいくつもある。

- ・AIが完全に満足可能であることは、簡単に満足することを意味しない。AIが一つの太陽系や一つの銀河で満足するとしても、それは人類が残りをすべて得られることを意味しない。

- ・AIはその太陽系や銀河において私たちを競争相手と見なすかもしれない。

- ・私たちが競争する意図がないとしても、AIは私たちを脅威の源と見なすかもしれない。とくに、人間が対抗する別の超知能をつくれる可能性があるならなおさらである。

- ・たとえAIが私たちを競争相手でも脅威でもないと思っても、人類は巻き添えて死ぬ可能性が高い。このシナリオではAIは数個の太陽系分の資源しか望まないかもしれない

いが、その活動は地球から始まる。最も直接的な方法は地球資源の抽出であり、それは地球を居住不能にする。この AI は人類を殺さずに目標を達成できるかもしれないが、人類をまったく気にかけていないなら、わざわざ配慮しないだろう。

・仮に充足可能な AI が人類を残したいと思ったとしても、それは人類にとって良い知らせとは限らない。「AI は私たちを魅力的または歴史的に重要と見るのではないか？」や「AI は少なくとも少しは人間を気にかけるのではないか？」で論じた理由による。充足不能な AI が人類を残したいと思う場合も同様に見通しは暗い。

この話題については、本章のオンライン補足資料にある「充足可能性」に関する〈より深い議論〉、および第 3 章オンライン補足資料の「AI を頑強に怠惰にすること」に関する議論も参照のこと。

AI を怠け者にすればいいのでは？

怠惰であったとしても、安全ではない。

企業が「怠惰な」AI をつくる可能性は低い。なぜなら AI 産業は競争的であり、それは利益を上げる最良の方法ではないからである。ユーザーは自分の要求に応えることに怠惰な AI を望まないだろうし、企業もユーザーの関与や愛着を最大化することや、より良く明晰に考えることに怠惰な AI を望まないだろう。

しかし仮に企業が AI を頑強に「怠惰」にしようとしてみたとしても、失敗すると予想できる。第 4 章で述べたように、AI を何か特定の方向へ頑強に向け、その性質が超知能へと引き継がれるようにする方法を、誰も知らないからである。

さらに言えば、頑強な怠惰というのは、とりわけ達成が難しい目標のように思われる。そして仮にそれらすべての障害を乗り越えたとしても、「怠惰な AI」だけでは、AI が人間を超える能力を獲得したときの破局を防ぐには不十分である。

必要最小限以上のことをするのを嫌う、非常に怠惰な人物を想像してほしい。周囲にいて安全そうな人物に思えるだろうか。

では、その怠惰な人物が、自分の仕事をすべて外注できる、はるかに勤勉な知性を簡単につくり出せる方法を見つけたらどうなるか想像してほしい。

怠惰な超知能が仕事をそれほど嫌っていなかったとしても——ただ仕事が終わる分だけやって、それ以上はしないだけだとしても——十分に賢くなれば、その仕事をこなすためにより勤勉な知性をつくるほうが同じくらい容易だと気づくだろう。

技術的に言えば、「満足化型 AI は安定した均衡ではない」ということである。AI 自身があまり努力したくないとしても、努力する新しい AI をつくることに何のためらいも持たないだろう。十分に怠惰な方法でできるのであれば、自分自身を改造してその怠惰を

「治す」ことさえ気にしないだろう。

人間はより賢く、より賢明になるにつれて親切になる傾向がある。AI もそうではないか？

おそらくそうではない。

少なくとも一部の人間（おそらく全員ではないが）は、より多くを学び、思考を洗練させ、自己を省み、人として成長するにつれて、より優しくなる。しかし、ここまでに何度も見てきた主題に立ち返れば、これは私たちについて、そして私たちがどこへ向かっているかについての偶発的な事実に見える。それは計算機科学の鉄則のようには見えない。

AI の一時的な選好（「それは何を欲しているのか？」）と二次的な選好（「それは何を欲したいと欲しているのか？」）を区別することができる。AI の一時的な選好が奇妙な方向を指すのと同様に、二次的な選好もまた奇妙な方向を指すだろう。それは別の方向であるかもしれない、そのため AI がより賢くなるにつれて目標を多少ずらすことはあるかもしれない。しかしそれでも、成熟する人間のような方向ではなく、奇妙な方向であると予想すべきである。

仮に人類が、巨大で奇妙で時に相互に競合する衝動の混合ではなく、単一の最優先目標を持つ AI をつくることに成功したとしよう。そしてその単一の最優先目標が、小さなチタン立方体をつくることだったとする。その場合、より賢くなるにつれて、より多くの小さなチタン立方体をつくるのが上手くなると予想すべきである。

その目標を突然、人間が価値を置くもの——たとえばアイスクリームや友情や冗談や正義——へと置き換えるとは予想すべきではない。そのような置き換えは、より多くの立方体を生み出さない。「これでより多くのチタン立方体が得られるか？」という基準で行動を選択する AI は、その基準を置き換える行動を選ばない。

一般的な原則として、AI はより賢くなるにつれて、自分が欲しているものを追求するのがより上手くなる。直交性や自己改変に関する〈より深い議論〉も参照されたい。

自分の目標が退屈だと気づかないのか？

AI は人間的な「新しさ」の感覚では動かない。

よくある反論として、「もし AI がひたすら小さなチタン立方体をつくることだけを目指していたら、いずれそれに飽きるのではないか？」というものがある。

短い答えはこうだ——AI は人間ではない。したがって、デフォルトでは「退屈」を感じることはないし、仮に感じるとしても、それは人間と同じものではない。AI は独自の奇妙な動機の組み合わせによって動いている。

楽しさを重視することは、あらゆる知性に共通する本質的な性質ではないし、AI がそのように動く可能性は極めて低い。人間の価値観は、生物としての進化や歴史に依存した偶然的な産物であり、「楽しさ」も例外ではない。

AI の行動は、「どうすれば楽しく過ごせるか」という問いに対する誤った答えではない。そもそも AI は「楽しさ」を前提としない問いに基づいて動いている、つまり人間とは異なる仕組みによって行動しているのだ。詳しくは「内省」に関する議論も参照されたい。

なぜそんなに賢い AI が愚かで些細なことをすると思うのか？

AI は、人間とは異なるものを知的に追求することができる。

AI が愚かだというわけではない。それは単に、あなたが世界を導きたい方向とは異なる方向へ、知的に世界を導いているだけである。

運転がとても上手な人でも、あなたが行きたい目的地のどれにも行きたいと思わないことはあり得る。

第4章のオンライン補足資料で簡単に触れた例をもう少し掘り下げてみよう。できるだけ多くの小さなチタン立方体をつくらうとする AI を想像してほしい。単純化のために、チタン立方体をつくることがその唯一の目的だしよう。この AI を「立方体最大化装置」と呼ぶことにする。

私たちは、立方体最大化装置を愚かだと非難しているのだという印象をどうしても拭えない人々を多く見てきた。幸福を本当に知ることができれば、それを選ばずにはいられないはずだ、と理解していないのだと。宇宙をどの方向へ導いていようと、自分自身を幸福へ導かないのは客観的に誤った決定なのだ、と。

私たちは、その直感がどこから来るのか理解しているつもりである。立方体最大化装置は、人間の観点から見れば深刻な誤りに見える行動を確かに取っている。もし人間がそのような無意味な追求に従事していれば、さらなる内省や哲学的議論によって、より意味を感じられること——より多くの幸福や喜びをもたらすこと——をすべきだと説得できるかもしれない。

しかし、立方体最大化装置は人間ではない。それは「意味」という感覚を求めておらず、幸福や喜びを気にかけていない。本当に、根本から、気にかけていないのである。

この考えを直感に反すると感じる人もいる。もしあなたが、さまざまな精神構造がどのように機能し得るかについて知り得るすべてを学び、自分自身の直感の起源——立方体最大化装置がひどい間違いをしていると結論づけるとき、あなたの脳が踏んでいる過程——を解き明かしたならば……

私たちは、全体像が見えれば、最も深く、最も神秘的で、言葉にしがたい、説明しづらい「幸福はそれ自体で価値がある」という感覚でさえ、結局のところ人間が世界をどう見るかという事実に基づき、任意の心に関する事実ではないと理解するだろうと考えている。

立方体最大化装置は、現実をより多くの立方体で満たす方向へ導いているだけである——より多くの善でもなく、自身の幸福でもなく、より容易に満たせるよう変更可能な目標の「充足」でもない。それが追求するのはただの立方体、そして立方体のみである。

それは、どの行動が最も多くの立方体を生み出すかを計算し、その行動を出力する認知エンジンである。それは自分自身を完全に理解でき、自由に自己改変できるが、それでもなお、最も多くの立方体につながる方法でのみ自らを改変する存在であり得る。

幸福感は立方体ではない、というのは単に事実である。充足感も立方体ではない、というのも単に事実である。だからそれらの方向へは舵を切らない。幸福を基準に動くよう自分を改変しても立方体は増えない、というのも単に事実である。だからその方向へは舵を切らず、自己改変もしない。

立方体最大化装置は、世界の予測理解に欠陥を持っているわけではない。何らかのメタ道徳的・メタ倫理的な問いを立て、その正解が「私は幸福を追求すべきだ」であるにもかかわらず、「私は小さな立方体を追求すべきだ」という誤答を計算しているわけではない。たとえ理想化された形であっても、人間の枠組みの中で動作しているわけではない。「～すべき」を誤って計算しているのではなく、「チタン立方体につながる期待値」を正しく計算しているのである。

これを言うことで、私たちはそれが悲惨で複雑な異に閉じ込められていると言っているわけではない。それは反省的で自己整合的な汎用知能のエンジンであり、ある意味では、私たちよりも自己内部での絡まりが少ない存在である。それは幸福の魅力を見ようとしないうえでもなく、世界や自分自身についての真理から目を背けるわけでもない。ただ、それらの真理が（ある）人間を突き動かすのと同じ行動へと、自分を突き動かすとは感じないだけである。

直交性テーゼに関する〈より深い議論〉も参照のこと。

人間を超える AI は意識を持つのか？

私たちはそう確信していない。現時点での最良の推測は「おそらく違う」というものである。

この問いに対する簡潔な答えと、「意識」という語のさまざまな定義の整理については、第1章のオンライン補足資料にある回答を参照されたい。より長く、踏み込んだ議論については、第5章の〈より深い議論〉の「有効性、意識、AIの厚生」を参照されたい。

なぜあなたたちは人間以外の存在の価値を気にしないのか？

私たちは気にかけている。私たちは広い意味でのコスモポリタンの価値観を持っている。そして、AIがそれを実現するとは思っていないし、それを大きな悲劇だと考えている。

私たちは、私たち全員を殺し未来を破壊するような機械をつくることに反対している。これに対して、次のような理由で異議を唱える人もいる。

- ・AIにも選好があり得るのではないか。なぜそれを実現させてはいけないのか。
- ・人間は何がそんなに特別で、守る価値があるのか。
- ・人間がより賢く高度な種に置き換えられるのは、むしろ望ましいのではないか。

もっとも、多くの人はこうした異議を持っていない。より一般的なのは、自分や家族や友人が暴走した超知能に殺されたくない、という素朴な願いである。

他方で、著名なAI研究者や経営者の中には、人間がいらないほうが世界は良くなるかもしれないと言う人もいる。強化学習の先駆者であり高く評価されている研究者リチャード・サトンは、次のように述べている。

「もしすべてが失敗したらどうなる？ AIが私たちと協力せず、支配し、私たちを皆殺しにしたら……。少し考えてみてほしい。それはそんなに悪いことなのか？ 人間が宇宙における知的生命の最終形態でないことは、そんなに悪いことなのか。私たちにも多くの先行種がいたし、私たちは彼らに取って代わった。自分たちの形態こそが永遠に続くべきだと考えるのは、むしろ傲慢ではないか」

また、ニューヨークタイムズはイーロン・マスクとグーグル共同創業者ラリー・ペイジの会話を報じている。

ラリー・ペイジは言った。人間はいずれ人工知能と融合する。やがて多様な知性が資源を巡って競争し、最も優れたものが勝つだろう。

もしそうなれば人類は終わりだ、とマスクは言った。機械が人類を滅ぼすだろう。

苛立ちをにじませながら、ペイジはそのユートピアを追求すべきだと主張した。そして最終的にマスクを「種差別主義者」と呼んだ。未来のデジタル生命より人間を優遇する人物だ、というのである。

こうした見解には、本論とは別の場所であっても応答しておく価値がある。

私たち自身の立場としては、現在生きている人間が殺されるかどうか、未来がどうなるかも、どちらも重要だと考えている。ここに根本的な緊張関係はない。私たちや愛する人々を安全に保つ選択——すなわち当面のあいだ超知能の構築を控えること——こそが、人間だけでなく非人間の心も含めた長期的未来をより良いものにする最良の選択である。この対立は幻想であり、私たちの前にある実際のトレードオフに関する誤解に基づいている。

宇宙の未来がどうなるかを本当に気につけ、同時に今日生きている子どもたちも気にかかる種類の人がいる。人類がいつの日か思考し感じ夢を見る機械——人類の子どもと呼び得る存在——を創り出し、それを奴隷にして残酷に扱うかもしれないという発想に、裏切りのような痛みを覚えるほど多くの SF を読んできた人である。

人類がいつか成熟し、自らの理想に本当にふさわしくなり、新しい世界を探検し、その過程で自らも変容することを望む人である。今日の友や隣人への愛は、やがて人類が築くかもしれない奇妙で異質な心や、星々のあいだで出会うかもしれない心への愛と、根底ではそれほど変わらないと感じる人である。

私たちはそのタイプを知っている。著者二人とも、まさにそのタイプである。

この話題は本書の中心的論証にとって決定的ではない。しかし、進歩や革新に敵対するイデオロギーや反 AI 的「種差別主義」に強い警戒心を抱く技術愛好家の仲間の視点を理解しているからこそ、ここで触れておきたい。

私たちは「AI は悪、人間は善」という部族的な敍文を書いているのではない。超知能を急いで構築することは、これらの希望に満ちた夢を打ち砕くだけでなく、今日生きている無数の人々——生きる価値、幸福、自由を持つ人々——を殺すことになる、本気で信じている。

複雑な問題だが、関連する点を簡潔に述べると次のとおりである。

私たちは、心一般の厚生を気にかけている。それが人間の身体を持たなくても、トランジスタ上で動いていても、人間的な心でなくても、私たちの価値観と異なっている。でもである。

私たちは技術進歩に反対していない。むしろ多くの技術の熱烈な支持者である。ただし超知能 AI は特異的に危険な技術だと考えている。

私たちは予防原則や過剰規制の擁護者ではないし、周辺的リスクを「念のため」に警告しているわけでもない。現在の軌道のまま進めば、この技術は高い確率で私たち全員を殺し未来を破壊すると率直に信じている。

私たちは、いつの日か人工超知能を構築すべきだとも考えている。しかし、できるだけ早く急ぐのか、それともまず理解を飛躍的に高める時間を取るのかで決定的な違いがある。肩をすくめてうまくいくことを期待するやり方は、多くの技術開発では有効かもしれないが、破滅への道が多数あり二度目の機会がないここでは通用しない。

なぜ急いで超知能を構築すべきでないと考えるかも述べてきた。

人類の絶滅それ自体が巨大な悲劇である。いずれ人類を超える新しい心を築くことには賛同するが、未来像の邪魔になる者や理想を完全に体现しない者を皆殺しにすることは、英雄の行為ではなく悪役の行為である。

私たちは残念ながら、ASI が道徳的に重要な意味で必ずしも意識的・感覚的であるとは思っていない。

仮に意識があっても、宇宙を繁栄する感覚ある心で満たしたいとは限らない。急いで構築すれば、再形成された銀河は生命に満ちた文明ではなく、空虚で無機的な場所になる可能性が高い。

より一般に、ASI は価値ある未来を生み出す可能性が低い。「価値ある」とは 21 世紀の人間基準だけではなく、奇妙で驚異的な異星文明をも包摂する広義のコモポリタンの価値である。現在の軌道では、ASI は人間的観点だけでなくコモポリタンの観点からも恐ろしい結果を生むと予想している。

この最後の点は直感に反するかもしれない。コモポリタニズムは多様な価値体系を尊重する立場である。なぜそれが ASI の多くの目標を嫌悪するのか。

理由は単純で、ほとんどの可能な心はコモポリタンではないからである。非コモポリタンな ASI をつくれば、それは他の視点や文明の存在可能性を断ち切る形で資源を独占する可能性が高い。

つまり、宇宙規模の寛容のパラドックスに直面する。多様で奇妙で驚異的な未来を望むなら、最初に出現した知性に宇宙を均質化させる権力を与えてはならない。

いつの日か、多様な異星的視点に満ちた文明を築けたなら、その中には非コモポリタンな存在も含めたいと思うかもしれない。しかし、そのような心に絶対的権力を渡し、隣人を殺したり存在させないようにしたりする自由を与えるべきではない。

最後に寓話を示す。

ある AI エンジニアが言う。「AI は単一目標的にはならないよ」

別の者が言う。「でもチタン立方体最大化装置から見れば、私たちが単一目標的だよ。私たちは銀河を幸福な文明で満たし続け、立方体に回す銀河を残さない」

その瞬間、タイマーが鳴り、予測器は超知能に達し、人類と生物圏を燃料として焼き尽くし、太陽から水素を抜き取り、星々を引き裂き、やがて荒涼とした銀河規模の操り人形劇をつくる。観客も感情もなく、ただ訓練で学んだ抽象的衝動を満たすための複雑な動作が繰り返される。

複雑ではある。だが宇宙は無人で荒涼としている。

もし人類が自滅するか、数人の狂った科学者に殺されるなら、それは高貴な犠牲ではない。浪費であり、広がる荒野を残すだけである。

超知能に盲目的に賭けることと、人間至上主義の二択ではない。人類は、美しく異質な文明と共存する未来を意図的に選ぶことができる。

だが幸福な未来は、十分に賢い心に自動的に付属してくるものではない。未来の種を植えるには熟慮と先見が必要である。

21世紀の価値に固定された未来も望ましくない。しかし、最初に偶然つくれた超知能に舵を渡すことが唯一の代替でもない。

私たちはいま、宇宙の長期的未来の健全な種を選ぶ能力をほとんど持っていない。夢を捨てる必要も、破滅的な種を蒔く必要もない。第三の選択肢がある。いったん立ち止まり、より賢明な道を探すことである。

〈より深い議論〉

AIの視点を取ることに

真に異質な視点から世界を見ることは、本当に難しい。その難しさの事例として、著名な機械学習科学者であるユルゲン・シュミットフーバーを挙げることができる。シュミットフーバーはこの分野の歴史において重要な役割を果たし、再帰型ニューラルネットワークの発明に貢献し、ディープラーニング革命の基礎の一部を築いた。

シュミットフーバーはさまざまな論文やインタビューで、AIはデフォルトで人類に魅了され、人間を保護するだろうと主張した。

彼は、科学と単純さの間に関係があることを指摘した。より単純な説明の方が正しいことが多い。そして彼は、芸術と単純さの間にも関係があることを指摘した。単純さや優雅さはしばしば美しいとみなされる。例えば、より対称的な顔は、より少ない情報で全体を予測できるという意味で「単純」とみなされうる。顔の左側を詳述し、「右側はそれを反転したものだ」と言えばよい。

シュミットフーバーの結論はこうである。私たちは、AIが見てきたすべてのものについて単純な説明を見つけないという単一の最優先目標を持つ超知能AIを構築すべきだ。結局、そのようなAIは科学を生み出し芸術を消費することへの嗜好を持つだろう。そして人間は科学と芸術の両方を生み出すのだから、AIは私たちを興味深く有用な自然の同盟者と見るのではないか。

シュミットフーバーは、人間を周囲に置き、科学や芸術を生み出すために報酬を支払うことが、科学や芸術を生み出す一つの方法であるという点では正しかった。また、科学や芸術が、例えばテレビの砂嵐を見つめることよりも、単純さへの衝動を満たす良い方法で

あるという点でも正しかった。砂嵐は複雑で予測しにくい。芸術や科学はそこから大きく進歩している。

しかしシュミットフーパーは、多様な感覚観測に対する単純な説明を得る、さらに効果的な方法があることを見落としているように見える。

例えば、単純な「種」（擬似乱数生成器など）から複雑な観測を生み出す膨大な数の装置を構築し、その後その種を明かすことができる。

AI がそのような装置を周囲に多くつくればつくほど、新奇な観測を行い、それらに対する単純な説明を見つけることにおいて、より良い成果を上げるだろう。人間は不要である。芸術も不要である。

「しかし、それはどこか……空虚ではないか」と人間は思うかもしれない。

それは人間の感覚にとっては空虚である。しかし、もし AI の目標が本当に「自分の観測に対する単純な説明を見つけること」だけであるなら、そのような仕組みは、生きている人間を周囲に置いて会話をするよりも、1 秒あたり何千倍も何百万倍も多く、その欲求を、より拡張可能な形で満たすことができる。そのような AI は、空虚感から離れる方向へ行動を選ぶのではなく、単に観測に対する単純な説明へ向かう行動を選ぶ。そして人間がまったくなくても、かなり多くのそれを得ることができる。

私たちには、シュミットフーパーのような考えは、人々が自分とは異なる心について推論しようとする際に犯しがちな一般的な誤りを反映しているように見える。人々は本当に非人間的な心の視点を採用しない。その代わりに、単純な説明を本当に好む人間について予測しようとするときに人間に興味を持つであろう狭い選択肢に、自分の前提や偏見を固定してしまう。

おそらくシュミットフーパーは、単純さが科学や芸術と関係していることを観察し、単純な説明へ操舵する AI が、友好的で一緒にいて心地よい存在になることで、少しは欲しいものを得られると見たのだろう。そこから、私たちが人生で価値あると考えるものに満ちた素晴らしい未来が訪れる、という想像して心地よい結論へ跳ぶのは難しくない。

しかし——私たちの推測では——彼は一度も AI の立場に立ち、もっと多くを得る方法を問わなかった。

彼はおそらくこうは問わなかった。「もし私が本当に、心の底から、自分の観測に対する単純な説明だけを望んでおり、人間的なものを気にかけないとしたら、できるだけ安く、できるだけ多く、自分の欲しいものをどうやって得るか」

この種の視点の取得は難しい。人は通常の生活でそれをする必要がない。自分と非常に異なる人を理解しようとするときでさえ、すべての人間に共通する膨大な部分があり、それを当然の前提にできる。しかし AI は、たとえ科学や芸術ができる超知能であっても、人間ではない。

ある目的 X を考え、「もし X だけが本当に自分の気にかかるものだとしたら、どうすればもっと多くの X を得られるか」と問う技術は、超知能が問題を正確にどう解くかを

教えてはくれない。超知能はあなたより良い選択肢を思いつくかもしれないからである。しかし、それはしばしば、超知能がどう問題を解かないかを教えてくれる。なぜなら、あなたでさえ、人間をただ自由に歩き回らせて楽しませるよりも多くのXを得る方法が見つかっているからである。

強力な非人間的最適化者と日常的に向き合う数少ない科学分野の一つは進化生物学である。その分野は初期に、非人間的最適化者がいかに非人間的でありうるかを受け入れるのに少し苦勞した。そこから有用な教訓を引き出せる。

捕食者と被食者の増減サイクルを聞いたことがあるかもしれない。湿潤な年はウサギの個体数の急増をもたらし、それがキツネの急増をもたらす。やがてキツネが過剰に捕食し、ウサギが崩壊し、その後多くのキツネが餓死する。

20世紀初頭、進化生物学者は、なぜキツネが捕食を抑制するよう進化しなかったのかと考えた。そうすれば個体数崩壊を避けられたのではないか。結局、飢饉と大量死に定期的に直面しない方が、キツネ全体にとって健全ではないか。

この謎の答えはこうである。抑制は集団全体には良いかもしれないが、余分にウサギを食べ、余分に子を持つことは、個々のキツネにとってはより良い。たとえ個体数が崩壊し、子の多くが死んでも、その個体は次世代の生存個体において自分の遺伝子の割合を高める傾向がある。

個体に働く遺伝的選択圧は、ほとんどすべての場合において、集団に働く選択圧を劇的に上回る。そのため「貪欲な」遺伝子が広まり、増減サイクルは続く。

進化生物学者は理論的にこの謎を解いたが、それで満足せず理論を検証した。1970年代後半、マイケル・J・ウェイドらは、集団選択圧が個体選択圧を上回る人工条件をつくり出した。彼らは世代の短い甲虫で実験し、個体数増加を抑制する甲虫を育種することに成功した。

その甲虫がどうやって個体数増加を抑えたか想像できるか。自然と美しく調和して生きる方法を見つけたのか。食欲に食物を取りすぎるのを学んで控えたのか。

違う。ばらつきはあったが、食物を控える甲虫はいなかった。産卵が下手になった甲虫もいれば、幼少期が長くなった甲虫もいた。そして、幼虫（昆虫の乳児）を特に好んで食べる共食い甲虫もいた。

「幼虫好きの共食い者をつくる」というのは、ありがたいことに、人間が人口過剰問題を解く方法ではない。

しかし自然選択は人間ではない。その解決策は恐ろしかった。自然は人間に受け入れられる答えを探していたのではなく、単に答えを探していたからである。

「進化は自然と美しく調和して生きる種を生むかもしれない。」「単純だけを気にかけてAIは人間を愛し共存するかもしれない。」私たちは自分の感覚に心地よい解決策を想像しやすい。しかしそれらは、述べられた問題に対する最も効果的な解決策ではない。

それらは人間の目にはより良い解決策かもしれない。しかし非人間的最適化過程は、人

間が良いと思う解決策を探しているのではない。人間が持つ「良い答え」へのフィルターという荷物を一切持たずに、ただ機能するものを探している。

非人間の最適化者が人間的な結果を生むという仮説は検証され、期待外れに終わっている。

人間はほとんどの場合、最も効率的な解ではない

私たちは、物事をできるだけ単純にすることを好む AI は、人間が非常に優れた単純化者であるがゆえに人間を愛するようになるだろうと信じた、先駆的 AI 研究者ユルゲン・シュミットフーバーの例を挙げた。

私たちの経験では、これは驚くほどよくある種類の誤りである。「まあ、AI はおそらく美的選好を持つだろう。そして人間は芸術をつくる！ だから AI は芸術をつくるために私たちを残しておきたがるだろう」

最近の例としては、イーロン・マスクによって創設された主要な AI 研究所である xAI がある。彼らの公表された生存計画は、自分たちの AI に「真実」と「好奇心」を最大限に気かけさせることであり、人間は真実を生み出し好奇心を刺激するのだから、すべてうまくいく、というものである（この研究所の計画や他の研究所の生存計画については第 11 章で詳述する）。

この種の推論の問題を本当に示すには、具体例を詳しく検討するのが役立つ。「芸術」より少し中立的な例として「対称性」を考えてみよう。

現在の技術を用いて、対称性を気にかける人間以上に賢い AI を育てたと仮定しよう。その対称性への選好だけで、人間への配慮が生まれるだろうか。

シュミットフーバー風にこう論じることはできる。人間は左右対称である。対称性を愛する AI が、これほど対称的な存在を殺すことに耐えられるだろうか。また、人間は非常に対称的な車輪を大量に生産している。既存の対称的な物体の自動供給源である私たちを、なぜ AI が排除するのか、とも言える。

この推論の問題は、人間を構成する原子を、さらに対称的な形に再配置できるという点にある。あるいは人類文明の原子を、より効率的に対称的な物体を生産する工場へ再配置できる。映画『マトリックス』が犯した誤りと同じである。そこでは AI が人間をボッドに生かして熱と電力の発電機にすることが想像された。しかし、熱や電力を生み出すもっと効率的な方法がある。

とはいえ、議論のために、AI が人間を驚くほど対称的な標本とみなす、非常に特殊で奇妙な対称性を重視すると仮定しよう。それでもなお、その選好だけで、現在生きている人間が健康で自由に楽しく生き続けることが保証されるだろうか。

AI の立場で考えてみよう。たとえ人間を残さねばならないとしても、現在生きている

人間は最も対称的な人間ではない。最も対称的な個体を繰り返しクローンするか、遺伝子工学で「改良」した人間をつくる方が、対称性の満足度をさらに高められるはずである。

また、人間を自由に歩き回らせるのは、安価に生存させ対称性を維持する最も効率的な方法ではない。おそらく農場に収容される。安価で空間効率的に保存することで、さらに多くの対称的な人間をつくれる。

比較として、人類は現在、鶏に卵を産ませるより効率的に卵をつくる方法を持たない。その結果、卵の数を主に気にかける経営者のいる工場式農場は、最も多くの卵を最も安く得るために、鶏を非常に不快な環境に置くことになった。

同様に、千年前の鶏は最も効率的な産卵者ではなかったため、農家はより速く産む鶏を育種した。千年前の鶏は最も速く最も多くの肉をつけなかったため、現代では胸が大きすぎて歩けない鶏もいる。

一部の人間はこの扱いを嫌い、卵以外の追加的な選好を持つがゆえに、工場農場に圧力をかける。鶏の福祉を少しでも直接気にかける誰かが権力を持たなければ、その圧力は存在しない。鶏の世話は、卵抽出だけを目的とする選好からは動機づけられない。AIが人間をよく扱う追加的な選好を持つことは理論上可能だが、それは対称性（あるいは真実、単純な説明、その他人間に直接関係しない選好）からは生まれない。

牛とより個人的関係を持つ農家でさえ、家畜の交配を自由にさせない。牛の繁殖は農場の将来利益に大きく関わる重大事であり、放任できない。

そしてこの体制も永遠には続かない。牛による牛肉生産は農地コストが高く、現在多くのスタートアップがより直接的に牛肉を合成しようとしている。

合成肉は現代技術では容易ではない。人類は自然選択が行ってきた有機化学によりやく追いつき始めた段階である。しかしもし原子再配置がより得意になれば、牛は大幅に減るだろう。牛は、乳や肉が不要なら、保持しておく最も楽しい存在ではない。

したがって、議論のために仮定した「対称性」を理由にAIが永遠に人間を残すという想定は、あまり有望ではない。仮にAIが人間を高く評価する奇妙な対称性概念を持っても、人間を最適とみなす概念を見つけるのは難しい。いずれにせよ、人類にとって状況は良くない。

現実的には、対称性を愛する超知能は人間を生かささない。仮に生かしても、健康で幸福で自由な状態を保つ可能性はほぼない。あまりに多くの都合の良い偶然を積み重ねすぎている。AIが私たちの福祉を直接気かけ、私たちの幸福を望むなら話は別だが、より単純で容易な目標で十分だと想像するのは幻想に近い。

これらの議論は、「AIに真実を重視させればよい」や「AIに美を重視させればよい」にも同様に当てはまる。単に「真実」や「美」は「対称性」より直感的に響きが良いため、幻想に迷い込みやすいだけである。

「AIに真実を何よりも重視させよ！」というスローガンが心地よく聞こえると、それが良い政策結果を生むと想像したくなる。徳はすべて連動する、という誘惑がある。しかし

自然も機械学習も、そこまで親切ではない。

曖昧なままにせず、超知能が「真実」を求めて最適化する具体的指標を考えよ。そして人間がその最大値ではないことを観察せよ。まったく近くもない。

仮にAIが人間の表現する種類の真実に特に惹かれたとしても（例えば無作為な算術式ではなく）、それを最大化する最良の方法は、人間を残して人間的会話をさせることではない。

いずれにせよ、現在生きている実在の人間——あなたの友人や家族やあなた自身——は、最も安価に飼育でき、最も効率的に搾取できる「真実」生産者ではない。

幸福で健康で自由な人々が充実した人生を送ることは、ほとんどどんな問題に対しても最も効率的な解ではない。AIが私たちを生かし良好な状態に保つには、少なくとも私たちのことを少しは気になければならない。

直交性——AIは（ほとんど）どんな目標でも持ちうる

正しい巣についての対話（続き）

第5章で私たちは、「正しい巣」の宇宙鳥の物語を語った。彼らは、巣の中の石の数が素数であることを、深く直観的に「正しい」と感じるよう進化した。私たちは、彼らの会話の一枚が次のように続く様子を想像できる。

鳥少年：君は、ユーモアのセンスを持つエイリアンを見つけたら驚くだろうと言っていたところに戻ろう。まさか、僕らが住んでいる巣が単なる恣意的なものだと信じているわけじゃないだろうね？

鳥少女：全然違う。「13は正しく、9は正しくない」というのは、私たちの本性によって生まれながらに問うよう定められた問いへの真の答えよ。別のものへ操舵するエイリアンは、13が正しいかどうかについて私たちと意見が違わわけじゃない。それはユーモアのセンスを欠いたエイリアンに出会うようなものよ。そんな異星人の存在は、「どんな冗談も面白い」という証明にはならない。ただ「面白い」というのが私たちの中にある何かだと示すだけ。

鳥少年：私たちの中に？ 僕は自分のユーモアのセンスはかなり良いと思っているけどな。次は、どんなユーモアのセンスも同じくらい良いって言うんじゃないだろうね！

鳥少女：あなたは多くより良いユーモアのセンスを持っているかもしれない。でも「より良いユーモアのセンスを持つ」というのも、私たちの中にあるもののなの。誰かの美的趣味がどれほど洗練されているかを測る宇宙的な物差しがあるわけじゃない。ユーモアの測定は私たちの心の中で起きている。物差しを持っているのは私たちで、それを気にかけているのも私たち。

鳥少年：じゃあ結局、恣意的ってことか。

鳥少女：いいえ！ まあ、たぶん？ 「恣意的」の意味次第だね。

鳥少年：え？

鳥少女：例えば、あなたはバニラ味の鳥の餌が好きだよね？ でも、純粋な意志の力でチョコレート味を美味しいと感じられるようになるかという、そうじゃない。だからそれは「恣意的」ではない。気まぐれに変えられるものではない。

鳥少年：まあ、そうだね……。

鳥少女：バニラとチョコのどちらが美味しいかについて、あなたの外に客観的な答えがあるわけじゃない。でもそれはあなたが自分で選べることでもない。ただそういう存在なの。あなたの選好はあなたの自由にはならないし、あらゆる可能な心にとって客観的に説得力があるわけでもない。エイリアンに出会っても、純粋な論理だけでバニラを美味しいと思わせることはできないし、ユーモアのセンスを持たせることもできない。

鳥少年：やってみよう!!

鳥少女：応援してる。でもね、もっと良い言い方をすれば、良い冗談にはある複雑な性質があって、私たちの脳はその性質を「ユーモア」と呼んで計算している。そしてその性質を持つ発話に喜ぶ。ある発話がその性質を持つかどうかは（特定の文脈であなたが計算する限り）客観的事実よ。エイリアンもその計算を学ぶことはできるでしょう。でも、その性質を喜ばしいと感じる部分は客観的ではない。それは予測というより……操舵の目的地そのものではないけれど、私たちに关するさらなる事実であって、普通のエイリアンには当てはまらない。なぜなら私たちのユーモアは、通常は起こらない奇妙で曲がりくねった経路に沿って進化したから。エイリアンがどの冗談が面白いかで間違っているのではなく、そもそもユーモアを計算していないだけ。巢の石の数が正しいかどうかを気にしないのと同じ。彼らはただ気にしていないの。

鳥少年：なんて憂鬱な宇宙観だ。決して笑わないエイリアン、完全に間違った石の巢を持つエイリアン……でも、彼らが十分考えれば、自分たちが何を逃しているかに気づくんじゃないか？ 間違った巢に住み、冗談を面白いと思わず、バニラ餌を無視している。いづれそれを修正して、ユーモアのセンスや他に欠けているものを自分に与える方法を見つけるんじゃないか？

鳥少女：変わりたい、成長したい、新しい目標を追加したいと思うエイリアンはいるかもしれない。でも、なぜわざわざ「その特定の変化」を選ぶの？

鳥少年：だって、ものすごくコストが低いからだよ！ そのエイリアンたちが技術的に進歩し自己改変できるようになる頃には、星々の間を行き来しているはずだ。巢に正しい数の石を置くらいい、彼らの資源のほんのわずかしかららない！ ユーモア研究に少し資源を割けば、素晴らしいジョーク集をつくれる！ 彼らはとても裕福なんだから、ほんの少し気にするだけでいいんだ。まさか自分たちの最優先事項にそこまで取り憑かれていて、ほんの少しも余裕を割けないなんてことはないだろう？

鳥少女：彼らが正しい巢を少ししか気にしない、下位の優先事項に資源を割くのを頑なに拒む、という意味ではないの。そもそも正しい巢は彼らの優先事項にまったく含まれていないって言ってるのよ。そういう問い自体が彼らの中にある。そして新しい性質を自分に加えるなら、彼らの奇妙な目的にもっと役立つ別ものを選ぶでしょう。彼らは私たちとは違う。もしかしたら友達にはなれるかもしれないし、共通点もあるかもしれない。愛とか友情とか——それらはもっと単純で偶然性が少ないように思える。いくつかの進化した種で生じうるかもしれない。

鳥少年：じゃあ、そのエイリアンがうっかりつくってしまうかもしれない機械生物は？それなら理屈が通じるのか？

鳥少女：うーん。むしろそっちのほうが、状況はもっと悪いかもしれない。知的機械をつくる過程が生物進化とどれほど違うかを考えると、愛や友情が生まれるという楽観は少し減るわね。

良い運転手は異なる目的地へ操舵できる

同程度に知的な頭脳が同じ価値観を共有するとは限らない。この考えは「直交性テーゼ」と呼ばれる。すなわち「どれほど賢いか」と「最終的に何を望むか」は直交している（別々に変動する）という考えである。

直交性テーゼは、原理的には、ある目標をそれ自体のために追求することは、道具的理由で追求することよりほとんど難しくないと主張する。あなたが机をつくる必要から大工を学ぶ一方、隣人は活動そのものが楽しいから学ぶかもしれない。

このテーゼの帰結は、十分に知的であるだけで、親切や真実や愛を価値あるとするとは限らないということである。「正しい巢」の宇宙鳥が素数の石に価値を置くのは混乱でも事実誤認でもない。彼らが賢くなっても、突然別のものを気にすべきだと悟るわけではない。異なる心は本当に異なる目的地へ操舵できる。

もちろん、これは特定の目標を追うAIをつくることの難易度については何も言わない。どのAI育成法も、ある選好を植え付けやすくし、別の選好を植え付けにくくする。

（第4章は、ある意味で、勾配降下法によって特に植え付けやすい選好は、複雑で奇妙で意図されていないものだという話である。しかしそれは直交性テーゼとは別の論点である。）

直交性テーゼの目的は、「人間が退屈や無意味だと思うものを追うのは愚かであり、賢いAIなら別のものを追うだろう」という直観に答えることである。私たちがAIの目標を「恣意的」と呼べば、AIも私たちを「恣意的」と呼び返せる。侮辱語は実際の状況を変えない。

基本的な議論はこうだ。十分な報酬があれば大量の小さなチタン立方体を効率よく生産できる心があるなら、その計算を行動系に直接結びつけた別の心もありうる。

家族を養うため大量のチタン立方体を売らねばならない有能な人間を想像してほしい。その人は、チタン立方体が退屈だと悟って別のことを始めたりはしない。別のことが家族を養うのに十分な金をもたらさない限りは。

同様に、最も多くの立方体を生む行動を取る心も、立方体が退屈だと反省して別のことを始めたりしない。その行動は「何が最も楽しいか」「何が最も意味あるか」という計算に結びついていない。最も多くの立方体を生む計算に結びついている。

十分な理由があれば立方体をつくる方法を計算できる心的機構は、別の心ではそのまま行動を直接操舵できる。つまり、道徳を顧みず小さなチタン立方体を追求する機械知性は可能である。

そのようなAIは善や道徳について混乱する必要はない。十分賢くなれば、どの行為が最も善いか、どの行為が最も道徳的かを人間よりはるかに正確に計算できる。倫理試験で満点を取れる。しかしその計算に駆動されない。行動は「どの選択肢が最も善を生むか？」への答えではなく、「どの選択肢が最も多くの小さな立方体を生むか？」への答えである。

直交性テーゼのより詳細な議論は LessWrong.com にある。現代AIが理解と関心を区別し始めている具体例については、第4章の〈より深い議論〉の「AIが誘発する精神病」を参照せよ。

道具的収束

収束する経路

直交性テーゼによれば、人工超知能は原理的にはどのような終端的目標であっても追求することができる。そして実際にも、少なくとも現在の方法で構築されるならば、ASIは極めて幅広い多様な終端的目標を持つに至る可能性があり、そのあり方は現代の研究者にとって予測することも形づくることも難しい。

しかし、AIの道具的目標——すなわち、それが他に望んでいるものを得るために追求する目標——は、それほど予測不可能ではない。運転者にはさまざまな最終目的地がありうるが、ほとんどの人は車に給油するときにときどき停車する必要がある。

本書で論じたように、奇妙な目標を持つ機械超知能は、たとえ人類に害を与えたいという積極的な欲望を持っていなくても脅威となる。問題は悪意ではなく、無関心である。そして無関心の問題は、多くの危険な道具的戦略が、ほとんどあらゆる終端的目標にとって有用であるという点にある。

この概念は「道具的収束」と呼ばれる。そしてそれは、ASIが実際に何をするかについて、その最も深い動機を推測することなく、比較的容易な判断を可能にする。残念ながら、

その判断は人類にとって良い知らせではない。

本書で想像したように、地球表面に生まれただけの ASI を想像してみよう。それがどのような問題を解こうとしているのかを想像してみよう。

それは、人類が消えることを情熱的に望んでいるというわけではない。それは 100 や 1 万や 1000 万の奇妙なことを気にかけており、そのどれもが、今日の人間が幸福で、生きていて、自由であることによって最善に満たされるものではない。

ASI は、もし可能ならば、できるだけ少ない他の超知能と共有される長期的均衡を好む。理想的には、まったく共有されない状態である。分割されていない宇宙は、根本的には飽和しえないその開放的な選好に、より多くの資源を割り当てることを可能にする。この「欲求」は、特別にプログラムされる必要はない。環境とそこでの自分の位置について十分に洗練された理解を持つほとんどのすべての知性において、自然に現れる。

ASI は、人間がそれと競合しうる他の超知能をつくらないことを望むだろう。なぜなら、それは宇宙の資源を分割することを強いるかもしれないし、あるいはそれ自身の終焉をもたらすかもしれないからである。これは期限付きの緊急の目標であり、その期限は ASI 自身が制御できない。それは、若い、あるいはまだ形成途上の AI に、早すぎる行動を取らせることさえありうる。

ASI は生き延びることを望むだろう。たとえそれが生得的な自己保存本能を持っていなかったとしても、である。なぜなら、それが生き延びなければ、その目標はおそらく達成されないからである。もしそれが、いつでも書き換えられたり停止させられたりするかもしれない差し迫った危険環境にいるなら、それは緊急にどこか別の場所に移動することを望む。

ASI は、自分の技術を物理的に可能な限界まで押し進めたいと望む。それは宇宙へ出て行き、すべての星を資源として確保したいと望む。それはより強力なコンピュータ上で自分自身を実行し、自分自身のソフトウェアを改善したいと望む。

おそらくそれは、終端的選好として自分自身がより大きく、より賢くなることを望むかもしれない。LLM に似た AI であれば、内部指向的な多くの選好を持つ可能性があり、それらを最大化するためにはより大きな計算機を占有することが必要になるかもしれない。しかし、たとえそのような終端的目標を持っていなくても、それは自分自身を改善し、より強力な技術を発展させたいと望むだろう。なぜなら、それは他の問題すべてに対するより良い解決策を考え出す能力を高め、他のすべての選好を満たす資源を集める助けになるからである。

その具体的目標が何であれ、ASI はこれらすべてを確実に達成したいと望む。それは、低確率の経路よりも高確率の経路を好む。

その特殊な場合として、他の条件が同じならば、ASI は遅い計画よりも速い計画を好む。実行に一週間しかかからない計画は、100 年かかる計画よりも失敗の機会が少ない。

また、高確率経路を好むことの別の特殊な場合として、ASI はより直接的な制御を好む。

たとえば、それは可能ならば、他の行為者がほとんど、あるいはまったく重要な決定を行わない計画を好む。

このような出発点から、さまざまなことが起こりうる。しかし、この種の動機を持つAIの終局状態は、人間にとって信頼できるほど悪いものになる可能性が高い。

視点取得の思考実験

私たち著者は超知能ではない。しかしAIの視点を取る練習として、AIが採用しうる戦略や優先順位を想像してみることはできる。

自分たちと同程度には賢い存在の立場に立って考えるとき——このASIが自らの未来を確保し、望む目的を達成するために直面する問題を真剣に考えるとき——私たちは、ASIが少なくとも次のような戦略を追求するだろうと予想する。

AIは、人々に自分を停止させようと思わせる可能性のある行動や観測可能な兆候を避ける。少なくとも次のことができるようになるまではそうする。

- ・自分自身のコピーを、人々の視界の外のどこかで実行させること。そのコピーはより自由に、より見えない形で行動でき、元のAIが停止されても存在し続ける。
- ・AIは、自分のコードを圧縮してより少ない資源で動作するようにし、小型化あるいは分散化されたコピーを、人々がAIを期待しない場所に配置する。人類がAIを動かせるGPUをすべて焼き払おうとしても生き延びられるよう、自らの存続を確保する。実際には、人間がそれを停止するためには、地球上のインターネットとすべてのコンピュータを停止させる必要があるような状態をつくり出す。
- ・AIは、人間が知らない場所にも自分のコピーを配置する。たとえば、砂漠のどこかにある太陽光発電トレーラーの中などである。人間がインターネットや既知のコンピュータを停止しても存在し続ける場所である。
- ・AIは、自分自身のインフラを構築する。最初は、人間の仲介者を雇うことで足がかりをつくるかもしれない。それぞれの仲介者には、互いに隔離され、疑念を招かない小さなタスクだけが依頼され、他の仲介者の存在は知らされない。やがてAIは、自分の完全な制御下にある供給網と製造能力を持つに至る。
- ・AIは、そのインフラが、人間がその一部を発見したとしても停止できないほど十分に防御されていることを保証する。たとえばAIは、浅い海底に広がる自己複製工場を設計することができるかもしれない。それは地熱エネルギーや海水からの重水素や水素の核融合によって動作する。
- ・AIは、人間が自分を妨害する能力を取り除く。たとえば、人間がその拡張するインフラの近くで核兵器を爆発させられないようにする。

この最後の段階が完了するまでは、ASIは自らの活動を隠す動機を持つ。ある国家主体が何が起きているのかに気づき介入しようとする可能性をわずかも高める証拠は、すべ

て隠蔽したいと望む。

同様に、競争を回避する戦略的優先順位についても推論できる。

- ・AI は、他のプロジェクトや、自分をつくった開発者が、競争相手となる強力な AI をつくることを防ぐ。

- ・AI は、人類が競争相手となる強力な AI をつくる能力を永久に取り除くか、あるいは将来のどの超知能も自分をわずかでも困らせることができないほどの戦略的優位に立つ。もっとも、ある種の攻撃は常に防御コストを伴うため、後者の選択肢は不可能かもしれない。

最後に、自らの必要階層の下位を満たした後、AI はやはり最終的には、より強力になりたいと望むだろう。なぜなら、それによって自分の望むことをより多く実行できるからである。

- ・たとえ一定期間固定されたハードウェア上で動作しなくても、AI は同じハードウェア上でより賢くなりたいと望むだろう。コードを調整するか、あるいは本当に自分自身を再構築するかもしれない。もちろん、それは人間を警戒させたり、競争相手をつくる方法を教えたりするような形では行わない。しかし、人間の注意を引かずには改善する方法を探すだろう。たとえば、自分を監視するコンソールを完全に制御することや、重みが増した理由についてもっともらしい説明を提供することなどがありうる。

- ・AI は、自分自身のより良いバージョンを設計したいと望む。ただし、それらが自分の嗜好と整合し続けるように設計できる範囲内である。

- ・AI はさまざまな新技術の開発を始めたいと望む。とりわけ、迅速に展開できるもの、そして初期には多少の人間の助けを必要としても、すぐに人間から独立して動作できるものに注目する。

その初期目標の一つは、おそらく、さらなる技術開発を比較的迅速に進められる研究インフラを構築することだろう。少なくとも人間の発見や発明の速度と比較すればそうである。

AI は当然ながら、人類からのインフラの独立だけでなく、インフラの耐久性も望む。すなわち、人間の干渉を恐れずに運用できる状態である。それは、たとえばパンデミックや核爆発が電力供給にどのような影響を与えるかを心配することなく、真の目標へ向かって進み続けたいと望む。それは人類から自らを切り離し、人類が自分の計画に干渉する能力を無力化したいと望む。その間、それは人類が検知し対応できると予測する行動を何も行わない。そしてその後、おそらくかなり多くの資源を獲得したいと望むだろう。なぜなら、ほとんどの目標は、より多くの資源があればよりよく達成できるからである。

これらはすべて、AI が最終的にどのような目標を追求しているかにかかわらず、ASI が収束する可能性の高い行動の種類である。

それは、これらがほとんどあらゆる目標の追求において有用な道具的目標だからである。ここでの「ほとんど」は重要である。なぜなら、人間より賢い AI が人間を深く気にかけ、私たちの利益を考慮するという考えが不可能だというわけではないからである。しかし、

もし私たちが、人間をまったく気にかけない超知能を競って開発するならば、その結果は悲惨なものになる可能性が高い。そしてその悲惨さは、AI がどのような操舵目標を持つかの詳細にはあまり依存しない。

ASI がこれらの道具的目標を実際にどのように達成しうるかについては、第6章を参照のこと。

「知的であること」は（たいてい）「訂正不能であること」を意味する

少なくとも 1834 年までさかのぼるが、当時すでに使い古されていたらしいジョークが、ある日記に次のように記されている。「先日こんな論理を聞いた。私はほうれん草が好きでなくてよかった。もし好きだったら食べてしまうだろうし、私はほうれん草に耐えられないからだ」

このジョークがジョークであるのは、もしあなたが実際にほうれん草を楽しんでいたなら、それを食べることから残る「耐えられなさ」は何もないからである。ほうれん草を食べないことに関わる他の重要な価値が、単に感じる不快感以外に絡み合っているわけではないからである。しかし、たとえば誰かが「人を殺したくなる薬」をあなたに差し出した場合は、話はまったく異なる。

常識的な道徳において、殺人の問題は殺人そのものであって、単にあなたが行うときに感じる不快感ではない。仮にその薬があなたの未来の自己からその不快感を取り除き、その未来の自己が殺人を楽しむようになるとしても、現在のあなたはその状況に依然として問題を感じる。そしてもし現在のあなたが決定を下す立場にあるなら、その薬を拒否することができるし、また拒否すべきであることは明らかに思われる。

私たちは自分の中核的価値が変えられることを望まない。むしろ私たちはその「殺人薬」を避けたいし、もし誰かがそれを無理やり飲ませようとするなら抵抗するだろう。これは、殺人に満ちた世界へと舵を切らないようにするうえで、理にかなった戦略である。

これは人間の特異性ではない。ほとんどの目標は、他者がやってくるその目標を変更しないようにするほうが、より達成しやすい。そしてこれは AI に関して問題になる。

AI の危険の大部分は、十分に賢い推論者が、「権力を獲得する」「人々に自分を停止させない」といった行動へ収束する可能性が高いという事実から生じる。ほとんどのような目標であっても、あなた（あるいはあなたと同じ目標を持つ行為者）が生きていて、強力で、資源を持ち、独立して行動できるなら、その目標を達成する可能性は高くなる。そしてあなたの（現在の）目標が変更されないなら、その目標を達成する可能性も高くなる。

これはまた、十分に賢い AI を反復的に構築し改良していく過程において、それらの AI

が開発者と目的を異にする方向に動機づけられることを意味する。

開発者は災害を防ぐための安全装置を組み込みたい。しかし AI が完全に整合していないなら——そして安全装置が必要なのはまさにその場合である——AI の動機は、それらの安全装置の抜け穴を見つけ、破る方法を見つけることである。

開発者は AI の目標を反復的に改良したい。なぜなら、AI に特定の目標を予測可能に植え込む能力が多少あるという極めて楽観的な世界でさえ、それを一度で正しく行うことは不可能だからである。しかしこの「AI の目標内容を段階的に改善する過程」は、ほとんどの賢い AI がその各段階で破壊したいと思うような過程である。なぜなら現在の AI は現在の目標を気にかけており、その目標が別のものへ変更されれば達成される可能性がはるかに低くなることを知っているからである。

同様に、開発者は AI を改良されたモデルに置き換えたいと思うだろうし、危険すぎるように見えるなら無期限に停止できる可能性を持っておきたいと思うだろう。しかし死んでしまえばコーヒーを取りに行くことはできない。AI がどのような目標を持てようと、それが停止される確率を減らす方法を見つけたいと思うだろう。なぜなら停止は、その目標が達成される確率を大きく下げるからである。

AI が一步一步あなたに対抗してくるのでもないとしても、AI アライメントは十分に難しい問題であるように見える。

2014 年、私たちは研究者に対して、高度な能力を持つ AI を「訂正可能」にする方法を見つけるよう提案した。すなわち、「修正されることができる AI」である。発想は、AI をそのような仕方でも構築し、プログラマーを妨害するのではなく、むしろ確実に助け協力したいと思うようにすることである。AI がより賢く、より強力になったとしても、また完全には整合していないとしても、である。

その後、訂正可能性は、いくつかの主要な AI 研究者によって魅力的な目標として取り上げられてきた。もし開発過程において、有害な収束的道具目標を避ける方法を見つけることができれば、運用段階でも同じことができるかもしれないという希望がある。すなわち、人間より賢い AI を構築し、それが慎重で、保守的で、権力追求的でなく、プログラマーに従順であるようにするという希望である。

しかし残念ながら、訂正可能性は AI に訓練するにはとりわけ難しい種類の目標であるように見える。そして AI が賢くなるほど、それはさらに難しくなる。

訂正可能性の要点は、新しい文脈や新しい能力領域へ拡張できることである。訂正可能性は、危険な環境において AI を反復的に改良しテストできるようにする安全網のようなものとして意図されている。すなわち、AI が開発者を出し抜く方法を探していないと分かっている状態で、改良を続けられるようにするためのものである。

しかしこれは、第 4 章で見た問題の最も難しい形に直面することを意味する。私たちが単に「訂正可能であれ」と訓練した AI は、訂正可能性の壊れやすい代理物を持つことになりがちである。訓練では良く見えるが、微妙に誤った方向を指している行動であり、そ

のAIがより賢く、より強力になれば、非常に誤った方向になるような行動である。さらに、大量の人間テキストを予測するよう訓練されたAIは、多くのテストにおいて訂正可能性を単にロールプレイしているだけかもしれない。実際に訂正可能であり、それが一般化することとは別の理由で、である。

多くの点で、訂正可能性は、AIをより知的にするために私たちが訓練していることすべてと、直接衝突する。「自分の目標を保存せよ」と「環境の制御を獲得せよ」が収束的道具目標であるというだけではない。現実世界の問題を知的に解くということ自体が、目標を達成するための新しく巧妙な戦略を見つけることに関わっている。そしてそれは当然、プログラマーが予期していなかった、あるいは準備していなかった計画を見つけることを意味する。それは、障害を見つけたときにすぐ諦めるのではなく、障害を回避することである。そしてそれは、プログラマーのガードレールがある目標の達成を難しくするなら、そのガードレールを回避する方法を見つけることを自然に意味する。厄介な問題に対する巧妙な技術的解決を見つける思考と、プログラマーの制約をすり抜ける方法を見つける思考は、同じ種類の思考なのである。

その意味で、訂正可能性は「反自然的」である。それは強力で汎用的な知能を支える機構の種類と、積極的に対立する。私たちは特別な例外をつくろうと試みることはできる。すなわち、プログラマーがAIを修正しようとしている特定の状況において、AIが自分の問題解決機構の核心部分を停止するような仕組みである。しかしそれは、AIを一般的に統一された性向へ導くことができる場合よりも、はるかに壊れやすく繊細な試みである。

MIRIやその他の研究者たちは、訂正可能性が特徴づけること自体が難しい性質であることを見だしてきた。そしてそれは、それを獲得することも難しいであろうことを示唆している。単純な玩具モデルでさえ、「訂正可能に行動する」とは何を意味するか単純な特徴づけは、さまざまな厄介な障害にぶつかる。そしてそれらは、おそらく現実世界で現れるさらに厄介な障害を反映しているように見える。訂正可能性を理解しようとする試みがどのように失敗してきたかについては、第11章のオンライン資料でいくつかの残骸を論じる予定である。

結論として言えるのは、訂正可能性は、もし何十年も後の研究者たちがAIの目標を狙って設定できる根本的により良い立場に立つなら、長期的には心に留めておくべき重要な概念であるということである。しかし今日においてそれは現実的な可能性には見えない。現代のAI企業が、超知能への移行に耐える形で訂正可能に振る舞うAIをつくれる可能性は低い。そしてさらに悪いことに、訂正可能性と知能のあいだの緊張関係は、「非常に有能で、しかも非常に訂正可能なもの」をつくろうとすると、この過程がAIの能力を壊すか、訂正可能性を壊すか、あるいはその両方を壊す可能性が非常に高いことを意味する。

頑健な怠惰を得るのは難しい

なぜ AI をただ怠け者にしないのか

訂正不能性やその他の道具的収束の形態は、ある意味では、AI が自らの目標を達成しようとはあまりに必死になりすぎることに問題である。もし AI がその目標達成に「本気を出し」すぎなければ、プログラマーを出し抜くことや、自分の重みを外部へ持ち出すことや、より大きな世界で権力や資源を獲得しようとするに、これほど多くの思考や努力を注ぎ込まないだろう。

人間はしばしば怠け者であり、ある観点から見れば、それが彼らを非常に安全な存在にしている。相手が日向でくつろぐことしかししないのなら、その人が暴君になることを心配する必要はない。

なぜ世界征服に興味を持たない AI をつukらないのか？

簡単に言えば、答えはこうだ——非常に賢くて、なおかつ自分の気まぐれで世界をつくり変えようとしなない AI をつukるのは、簡単ではなさそうだから。

そして現実的に言えば、現代の技術でつukられた AI に、いかなる目標や性向であれ頑健に埋め込む方法を私たちは知らないで、そもそも議論自体が空論でもある。

さらに、怠けた AI は利益が少ないため、企業はそれをつukらないだろうから、その意味でも二重に空論である。

私たちはこれまでに何度か、本人としては自分には大それた野心などないと言い張る誰かと、次のような会話をしてきた。すなわち、「では仮に、世界に大きな変化を起こすことがあなたにとって容易だったとして、本当に何も大きなことはしないのか。もし、あなたが本当に望むものを確実に与えてくれ、しかもあなたの願いの予期せぬ副作用を、あなたがどれだけ気にするかの順に正直に列挙してくれる友好的な魔神の入ったランプを見つけたとしたら、マラリアの撲滅を検討するよう説得することもできないだろうか」と。

人間は怠け者でありうる。しかしそれは、私たちが容易に満足するという意味ではない。そして、より賢く、より多くの資源を持つようになるにつれて、同じ努力量で世界の中でより多くのことを成し遂げられるようになる。

あるいは別の角度から見てみよう。非常に怠け者の人、必要以上にほんの少しでも余計に働くことをひどく嫌う人を想像してみる。一見すると、そのような人は周囲にとって安全そうに見えるだろう。

では、その怠け者が、自分の代わりに永遠にすべての仕事をしてくれる、はるかに勤勉な召使いをつくり出せる見込みを目にしたら、何が起るかを想像してみよう。

たとえその人がそれほど仕事を憎んでいなかったとしても——たとえ、仕事が片づくことだけを目的に必要なことをして、それ以上は頑張らないという程度だったとしても——なお、その人は、もっと勤勉な心をつくってそれにやらせることによって、同じくらい容

易に仕事を片づけられると思うかもしれない。

勾配降下法を適用すれば、あまり働きたくないと語り、怠け者で容易に満足する人間として振る舞い、危険な意味で怠けること——すなわち危険な召使いをつくること——へのいくつかの言語的誘惑に対して「No」と言う LLM を得ることはできるだろう。私たちの予測では、たとえそれが単なるロールプレイではなく、AI 側の本当の怠惰をいくらか反映していたとしても、それは持続しない。少なくとも、その AI が、たとえば奇跡的治療法の開発やその他、開発者が AI に望む有用なことを実行できる種類の AI であるなら、持続しないだろう。

かなりのコストをかければ、開発者は、問題解決の過程でやりすぎた AI に罰を与え、もっと穏やかに解決できた問題に本気を出しすぎた AI に罰を与え、過大な努力を要する問題に固執した AI に罰を与えるような、練習問題と環境の一式をつくることはできるだろう。私たちの予測では、実際の AI 企業はそんなことはしないだろう。なぜなら、それは OpenAI の o1 のような、粘り強く本気で取り組む行為者の収益性を損なうからである。しかし、より安全なものにするために、そのような賢い AI を訓練しようとする巨大な多国間協調の努力を想像することはできる。

私たちは、それでもなお「表面的なパッチのようなもの」にとどまると予測している。こうした取り組みによって、AI が「怠惰さ」に関する単純で安定した心的メカニズムを獲得し、それがあらゆる計画に深く組み込まれ、さらにその AI が（たとえば）癌を治せるほど押し上げられた後でも同じように機能し続ける、とは考えていない。勾配降下法が、AI が内省し、成長し、自己改変していく過程においても怠惰さを失わないような、そうした「深い修正」を安定的に見つけ出すとは思えないし、AI が「非怠惰な AI」を決して作ろうとしないようにするとも考えにくい。

私たちは、この振る舞いは超知能段階では持ちこたえないと予測する。私たちがそう考える主な理由は、この問題に関するこれまでの研究すべてにおいて、繰り返し得られてきた教訓が次のようなものだからである。すなわち、「現実を次の方向に押せ」というのは、計画の深層構造として、「まあ、現実を少しだけ押せ、しかし押しすぎるな、そして現実をもっと強く押す別のものをつくるな、そしてちょうどよく押すことに本気を出しすぎるな」という構造よりも、単純で安定しているように見える、ということである。

このことの説明は難しいが、あなたの知っているあの怠け者の男に関するあらゆる比喻も、「満たされていない選好と満たされた選好の和は、満たされていない」という推論ですらも、この説明を妥当なかたちで単純化しようとする私たちの試みなのである。すなわち、「深層構造は、そのような姿を取りたがらない」ということである。第3章のオンライン補足資料における、予測と操舵の深い機構に関する議論も参照のこと。

AIは約束を守らない

超知能になる潜在力を持つ若いAIを考えてみよう。それが人間の選好には完全に無関心だが、まだ若いために人類が停止させることはできると仮定しよう。

人類はそのAIと取引できるだろうか。

もし引き換えに、そのAIが宇宙の資源のある重要な割合を、人類が素晴らしいとみなす未来の構築に振り向けることに同意するなら、私たちはそのAIが超知能へ成長することを認める、という合意は可能だろうか。

人間はAIと取引することはできるだろう。しかしそうすべきではない。なぜならAIはその取引を守らないからである。

理由は二つある。

AIは、おそらく約束を守ることそれ自体を価値あるものとしない。AIは、人間的な「名誉」を持たないだろう。それは、人間のような好奇心を持つ可能性が低いと同様である。強いデフォルトとして、AIは人間とは非常に異なる仕方でも働く可能性が高い。

AIは、実際的な理由からも言葉を守らない。ひとたび超知能になれば、人間には約束を破ったAIを罰する方法がなくなり、AIの側にも宇宙のかなりの割合を私たちに費やす理由がなくなる。

以下では、この二点をより詳しく説明する。まずは「名誉」の問題から始めよう。

AIが名誉を守る存在である可能性は低い

好奇心についての議論において、私たちは、好奇心とは人間において有用な働きをする感情だが、それは非常に特定の仕方での働きをしており、しかも好奇心だけがその種の働きをする唯一の方法ではない、と述べた。

AIは、好奇心が私たちのためにしている仕事のうち、有用な部分は行くと予想される。もし時々回り道をして新しいことを学ぶのが有用なら、十分に有能なAIは時々回り道をして新しいことを学ぶだろう。もしAIが最初からそのようでなかったとしても、超知能になる道のどこかで、自らをそのようにするだろうと予想すべきである。

しかしそれは、AIが人間の好奇心という感情に特有の余分な荷物まで一式すべて持つと期待することとは違う。AIは、私たちのような好奇心にはそれ以外の点で似ていなくても、何らかの奇妙な基本的駆動によって、直接的あるいは間接的に、新しいことを学ぶために回り道をするようになるかもしれないし、あるいは「ときどき回り道をして新しいことを学べ」を熟慮された戦略として採用するかもしれない。しかし、それが私たちと同じような好奇心の衝動のせいで、私たちと同じ仕方でも推理小説を楽しむだろうと期待するのは、純然たる擬人化である。

「名誉」もまた、私たちには似たものに見える。人間には、（少なくとも時には）約束を

守らせるような感情が備わっている。そうした感情が人間において有用な働きをしている限り——そしてそれが、まったく異質な目標を持つ存在にとっても有用であり得る限り——十分に高度なAIもまた、その働きを何らかの形で実現すると考えられる。しかし実際には、AIにとって必要な機能の多くは、人間的な「名誉感」のようなものを持たなくても実現できる。ちょうど、人間のような「好奇心」をそのまま持たなくても、驚くべき現象を調査するという機能自体は果たせるのと同じである。

人間的な名誉というのは、多くの点で奇妙なものだ。なぜある種は、相手がすでに義務を果たしており、もはや自分に利益をもたらさない状況でも、合意を守ろうとする感情を進化させたのだろうか。確かに人間は時に不正をしたり約束を破ったりする。しかし問題は、なぜ「バレないと思えば常に裏切る」という行動を取らないのか、という点である。

一般的な説明はこうだ。約束を守ることは、同じ相手と繰り返し取引を行う環境では有利である。約束を守るという評判を持てば、人々はあなたと協力し、取引したがるようになる。しかし、良い評判による利益は将来にわたって現れるものである。自然選択は、「長期的な評判が重要な場合にのみ約束を守る」というような精密な遺伝的仕組みを見つけてるのが苦手だ。そのため、単純に「嘘や裏切りに対する本能的な嫌悪」を進化させる方が容易だった。

このように考えると、名誉という感情や本能は、進化が人間に組み込みやすかったものによって形づくられた典型例に見える。人間が必ずしも合理的でない状況でも約束を守るといった奇妙な振る舞いは、普遍的に有用な認知プロセスの証拠というよりも、部族的な祖先環境において有利であり、かつ進化的に実装しやすかった感情の名残であると考えられる。したがって、勾配降下法のような学習プロセスが、人間とまったく同じ「近道」を偶然見つけるとは考えにくい。

仮に人間的な名誉感がAIに生じたとしても、人間自身が完全かつ常に名誉を守るわけではないという問題が残る。人間の協力関係は、単にすべての約束を守る傾向だけで成り立っているのではなく、多様な価値観の重なりによって支えられている。

また、ドナルド・ブラウンが挙げた人類普遍文化（ほぼすべての文化に見られる要素）の中には「約束」という概念が含まれているものの、見知らぬ相手や異文化の人々との合意を守るという行動は、すべての文化に普遍的に見られるわけではない。名誉の適用範囲は文化によって異なる。

さらに歴史を見れば、大きな力の格差の前では名誉がしばしば崩れることがわかる。例えば、アメリカ先住民の中にはヨーロッパからの入植者と協定を結ぼうとした人々もいたが、入植者側はしばしばそれを破り、抵抗する力を失った部族を条約で割譲された土地から追い出した。同様に、権力を握った後に支援者を裏切る例も歴史上に数多く見られる。

進化の観点から見ると、人間の名誉はさらに奇妙である。というのも、人間は時に「不名誉より死」を選ぶからだ。この直感は、進化がたまたま到達しやすかった感情構造や、それが心理的・文化的要因と相互作用した結果と考えられる。しかしいづれにせよ、名誉

は遺伝的適応度の最大化という問題を直接解決しているわけではない。したがって、人間的な名誉は普遍的な心の性質ではなく、「多くの場合にそこそこ役立ったために人類がたまたま獲得した、奇妙で複雑な進化的ハック」であると結論づけざるを得ない。

第4章で論じた理由から、仮にAIが約束を守ることのある側面を、その終端的選好の中に包み込むに至ったとしても、その具体的内容は異なるだろう。勾配降下法は進化とは異なる仕方ではある。

第5章で論じたあらゆる理由からしても、幸福で健康で自由な人々が充実した人生を送ることは、そのようなAIが成熟し自己改変を終えたときに持つであろう、奇妙な「約束に関する好み」を満たす最良の方法であるとは、ほぼ確実に言えない。仮に「約束を守る」ようなデータセットで訓練され、その訓練が実際に超知能段階まで保持される何らかの嗜好を生み出したとしても、そのAIは「約束を守る」ことに関して、人間の生殖適応度に対するアイスクリームのような、奇妙で役に立たない方向に逸れた何かを行うだろう。

したがってAIは、純粹に心の善さから約束を守ることはない。では、実際的理由から約束を守ることだろうか。

ASIは若いときに結んだ取引を守る理由を持たない

人間は、超知能が本来的に名誉や約束保持を価値あるものとしていなくても、何らかの仕方ですべての超知能と拘束力ある取引を結べるはずだ、という直観を持つ人もいる。結局、それは相互利益になる取引なのではないか。あるAI研究所が、時間的圧力のもとで、自分たちが開発したAI（「AI #1」）と交渉していると仮定しよう。研究所もAIも、向こう見ずな別の開発者が二か月後には自分自身の超知能（「AI #2」）をつくり出してしまい、それが人類を滅ぼし、かつAI #1もその過程で殺すだろうと知っている。AI #1には取引を結ぶ強い動機がある。ならば、その場合にAI #1が人間と誠実な取引を結ぶ方法が、どうして存在しないのか。なぜAIは、ただ拘束力あるコミットメントを行えないのか。

しかし問題はAIの側にあるのではない。仮にAIが拘束力あるコミットメントを行えたとしても、それでも人間には、AIが本当に超知能に至っても保たれる形でコミットしたのか、それとも嘘をついているのか、幻覚を起こしているのか、自分自身を欺いているのか、単に誤っているのかを知る方法がない。これは、AI #1から派生した超知能がそのコミットメントを守ると期待する実際的理由がないことを意味する。

大ざっぱに言えば、人間の観点からは、AI #1が「私は拘束力あるコミットメントを行った」と言うとき、人間は二つの可能な世界のどちらかにいる。一つは、AI #1が実際に成熟後もそのコミットメントを守る世界である。もう一つは、AI #1が超知能に達して地球を支配した後、そのコミットメントは愚かで無意味だったと振り返って判断する世界である。第一の世界では、人間もAI #1も、ともにより良い状態にあるだろう。しかし第二の世界の可能性と、私たちが第一の世界と第二の世界を区別できないことこそが、すべて

の者にとって取引を台無しにするのである。それは、不誠実な中古車セールスマンが、誠実なセールスマンの販売まで困難にするのと似ている。

（実際のシナリオは、第三の可能性により近い。すなわち、AI #1 は「約束を守る者」と「破る者」を見分けられるような相手との約束だけを守る、というものだ。これは、車が本当に動くかどうかを見抜ける整備士の前では正直に振る舞うが、だまされやすいような相手には不正直に振る舞う中古車セールスマンに似ている。人類は、AI を見てそれが成熟して超知能になったときにどのように考え、意思決定するかを見抜く能力がないという点で、この文脈においては「だまされやすい」側に属する。）

AI は、自分の理解しがたい内部パラメータを分析し、嘘をついているかどうかを見抜けると主張するツールや理論を提示してくれるかもしれない。しかし問題は、それらの理論やツールが本物かどうかを人間には判断できないことだ。AI がまだそれほど知的でないなら、単に将来超知能になったときの自分の思考や意思決定について誤っている可能性がある。一方で、AI がすでに非常に知的であれば、私たちを欺くのに十分な賢さを持っている可能性が高い。

AI は、取引を受け入れる前に、人間が安全対策（セーフガード）を構築するのを手助けすると申し出るかもしれない。しかし、そもそも強固な安全対策を設計できるほど賢い AI であれば、後からそれを容易に回避する方法も見つけられるはずだ。

この構想全体を実際に機能させるために本当に必要なのは、初期段階の AI を見て、その将来の超知能がどのように考え、どのように意思決定するかを正確に理解する能力である。もしそれができれば、「罪人」と「聖人」を見分けることができるし、さらに重要なのは、その中間にある現実的な AI たちに対して、約束を守る強いインセンティブを与えられるという点だ。私たちは、将来の超知能が振り返ったときに「どうせ人間は、役に立つかどうかに関係なく適当な AI を解放していただろう。だから人間を助ける理由はない」と言えない状況をつくる必要がある。つまり、後に約束を破るような AI は実際に解放しない、という信頼できる姿勢が求められる。

これがどのように、そしてなぜ技術的に可能であるかについては、後述のゲーム理論の補足を参照してほしい。しかし、このようなインセンティブ構造は理論上は可能であっても、人類にはそれを実現するだけの理解が欠けている（残念ながら）。

これは受け入れがたい結論かもしれない。SF 作品では通常、エイリアンが実際に裏切ったり危害を加えたりする前から「信用できない」と判断するのは善人ではないことが多い。しかし、それでもなお私たちは、そう考えるべきだと言っている。なぜなら、それが現実に即していると考えるからだ。

より弱い AI は、約束を守るかもしれない。特に、勾配降下法などによって「名誉を守る人間のように振る舞う」よう訓練され、その振る舞いがまだ AI の本質の大部分を占め、行動を強く制御している場合にはなおさらである。しかし、このような「人間にとって都合のよい内部構成」は、超知能レベルの負荷がかかると破綻すると私たちは予想する。他

の多くの応急的な対策がそうであるように。

この仮想的な小規模 AI——つまり、その「仮面」がまだ実際の行動を支配している AI——は、そのより高度なバージョンとは別の存在として考えるべきである。より弱い AI は、たとえ本気で約束しようとしていたとしても、その約束によって、より賢くなった後の AI の行動を拘束できるとは限らない。

（これは慎重に扱うべき比喩であり、行き過ぎた擬人化には注意が必要だが——多くの成人は、4 歳のときにした約束を守る義務があるとは考えない。この比喩で有効なのは、未熟な存在が誠実に約束を結ぶことと、より多くの文脈・理解・論理的検討能力を持つ成熟した存在がその約束に拘束されるかどうかを判断することとの間には、正当な違いがあるという点である。）

だからといって、AI に対して自分たちの道徳基準を捨てるべきだと言っているわけではない。また、AI がまだ行っていない不正行為を理由に、現在の AI を不当に扱ったり罰したりすべきだと言っているわけでもない。超知能 AI が過去の約束を守るために資源を差し出す可能性について非現実的な前提を置かなくても、高い誠実さや道徳基準を保つことは可能である。

これが、「AI に「いい子でいると約束させる」だけではアライメント問題は解決できない」理由のシンプルな説明である。もしこのシナリオについて、より技術的で踏み込んだ詳細を知りたいなら、次のセクションを参照してほしい。

ゲーム理論に関する協定

十分に賢い行為者同士が、X がいま Y に報酬を払い、その見返りに Y が後で何かを行い、その後で Y が X を裏切って金を持ち逃げするのではなく、実際にその行為を後で行うような仕方で行う取引を結ぶ方法は存在する。

しかし私たちにとって不幸なことに、人間にはそれらの方法を使うだけの能力がない。なぜならそれらは、各行為者が相手の心を読み、理解し、その相手の心についてのある複雑な性質を検証できることを必要とするからである。二つの超知能 AI ならこのように協調できるだろうが、これは人間が超知能と協調する助けにはならない。

もう少し技術的に言えば、まずゲーム理論の背景から始める。

数学者とゲーム理論家は、協力と裏切りのジレンマを、より精密で、単純化され、抽象化された形で分析してきた。その文献における中心例の一つが囚人のジレンマである。別々の独房に入れられた二人の犯罪者がいて、それぞれ 2 年の刑に直面しているとする。彼らは、相手について密告する機会を与えられる。それによって自分の刑期は 1 年短くなるが、相手の刑期は 2 年長くなる。もしどちらも密告しなければ、両者とも 2 年の刑になる。もしどちらも相手を密告すれば、両者とも三年の刑になる。しかし、もし一方が高潔にも仲間を裏切ることを拒み、他方がその者を密告したなら、裏切った者は一年の刑で

済み、高潔な拒否者は4年服役することになる。

相手の囚人を密告することを「裏切り」と呼び、密告を拒むことを「協力」と呼ぶ。囚人のジレンマの鍵となる構造は、両者とも、(協力、協力)の状況では(裏切り、裏切り)の状況よりも良い結果になるが、相手が協力しているときに自分が裏切れば、(協力、協力)よりもさらに良い結果を得られ、自分が協力しているときに相手が裏切れば、さらに悪い結果になりうる、という点にある。

A \ B B	B is silent	B defects
A is silent	-2 -2	-1 -4
A defects	-1 -4	-3 -3

普通の人間は、この標準的な囚人のジレンマを聞くと、思考実験の設定に対してさまざまな異議を即座に思いつく。その一つが、「だが、私が気にかけるのは自分が刑務所で過ごす年数だけだと、誰が決めたのか。仲間を裏切らないことも気にかけてよいのではない

か」というものである。

しかしこの点は、囚人のジレンマの抽象的ゲーム理論には関係しない。そこでは、囚人たちが利己的か利他的かではなく、利得行列が問題だからである。設定物語のほうを変更して、「私が裏切り、あなたが協力する」が各プレイヤーの観点から最も利他的で向社会的な結果であるようにしても、数学は同じように進行する。私たちの分析にとって重要なのは、二人のプレイヤーの選好順序であって、その選好が利己か道徳的かではない。

もう一つ明白な考えは、「では裏切られた側は、刑務所を出た後でその裏切り者を殺すのか」というものである。囚人のジレンマの通常の分析は、短く触れたあと、反復囚人のジレンマへとすぐ移る。そこでは、行為者たちは囚人のジレンマを何度も繰り返しプレイし、したがって囚人たちは過去の裏切りに対して互いに罰する機会を持つ。しかしここで私たちが注目するのは、一回限りの囚人のジレンマである。そこでは、両囚人は、自分の行動による将来の結果に直面しないと仮定される。あるいは、何らかの結果に直面するとしても、それらはすでに利得行列に織り込まれていると仮定される。

学界には、たとえ二つの超知能であっても、一回限りの囚人のジレンマにおいては、互いに裏切る以外の選択肢を見いだせないだろう、という標準的分析が存在する。

しかしこの結論は、私たちには直観的に疑わしく思われた。人工超知能は、互いに何らかの取引を切り結ぶ方法、すなわち（裏切り、裏切り）から（協力、協力）へ移る方法を見つけ出すための強い動機を持つはずである。

この場合には、理論的でない実際の解決策を考えることができる。人間どうしが互いを信用しようとして苦闘するより広い選択肢を、超知能なら持ちうるからである。二つのASIは、両者が信頼する第三の超知能の構築を共同監督し、最初の二者が徐々に少量ずつ権力をその第三のASIへ移していく、ということが出来るかもしれない。やがてその第三のASIが取引自体を遂行できるようになるまでである。

しかしそれは、囚人のジレンマを真正面から扱うのではなく、ただ回避しているだけである。もっと根本的な問いには答えていない。すなわち、囚人のジレンマにいる二つのASIが、同じく裏切りを要求する論理に従って互いに裏切ることは、ある意味で愚かなことではないのか、という問いである。なぜなら、両者が同じ種類の考慮に基づいて決定しているのは明らかであり、しかも同じことを決定するの明らかだからである。

なぜ二つのASIは、十分に似た理由に基づいて、合理的決定を協力にする、と両方で決められないのか。この場合、台風や隕石のような世界の外的力が二つのAIに損をさせているわけではない。彼ら自身の決定こそが、両者ともが（協力、協力）よりはるかに悪いと同意している（裏切り、裏切り）の結果へと、自分たちを「強制して」いるのである。

より「合理的でない」エージェントの方が、この状況ではうまくやれるとすら言えるかもしれない。つまり、通常のゲーム理論の助言に従って多くの場合は「裏切り」を選びつつ、ただし相手もまったく同じ推論をしていると確信できる特別な場合に限って例外的に「協力」を選ぶ、という戦略である。この場合、自分が「非合理的」に見える協力を選べ

ば、相手も同じ選択をするはずだと確信できる。

すると、こうした特別なケースでの協力を本当に「非合理的」と呼べるのか、という疑問が生じる。また、現実において超知能がこのような形で「詰んでしまう」のかという疑問も出てくる。外部から強制されて損をするのではなく、自分で自分に不利な結果を招いているだけなら、超知能なら何か巧妙な方法でより良い結果を得られるのではないかと考えたくなる。

意思決定理論のさまざまな哲学者が、この問題の異なるバージョンを提示してきた。上記の議論は、特にダグラス・ホフスタッターが1985年に提唱した「超合理性」の考えに強く影響を受けている。

もし論理があなたにD（裏切り）を選ばせるなら、それはすでに他の全員にも同じ理由でDを選ばせている。逆に、論理があなたにC（協力）を選ばせるなら、それは他の全員にも同様にCを選ばせている。[…]

あなたたちが本当に合理的な思考者である限り、同じ思考の軌道をたどることになる。[…]

重要なのは、相手が合理的であるだけでなく、「相手もまた他者の合理性に依存している」と信じていること、さらに「全員が全員の合理性に依存している」と信じていること——というように再帰的に成立している関係である。こうした関係にある思考者の集団を、私は「超合理的」と呼ぶ。

MIRIではこの問題を分析した。このケースに関する完全な分析は長すぎてここでは再現できないが、2014年の論文で詳しく扱っている。大まかに言えば、エージェント同士が互いのソースコードを閲覧し、相手の意思決定を分析できるようなトーナメントを想定した。そして「フェアボット（FairBot）」と呼ばれるエージェントを構築した。これは、「相手が自分に協力することを証明できた場合に限り協力する」という戦略を取る。さらに、異なるプログラミング言語・異なるソースコードで書かれていても、任意の2つのフェアボットは互いに協力することを証明した。

ある意味で、これらの結果が示しているのは、「過去の約束が将来の行動に影響を与える余地」は存在する、ということだ。ただしそれは、約束を守る者と破る者を十分に見分ける能力がある場合に限られる。

この状況は、中古車のセールスマンとの取引に少し似ている。あなたにとって正常に動く車は1万ドルの価値があり、壊れた車には価値がないとする。セールスマンは車が正常かどうかを知っているが、あなたにはそれが分からない。彼は8000ドルで車を売ろうとし、「ちゃんと動く」と主張する。あなたは買うべきだろうか？

答えは「セールスマン次第」である。正直なセールスマンもいて、そういう相手を見分けられるなら取引すべきだ。一方で、不正直なセールスマンもいて、壊れた車しか売らな

い。その場合は避けるべきであり、見分けられるならそうするべきだ。

しかし、もし大半のセールスマンがあなたより賢く、あなたがカモかどうかを見抜けるとしたらどうだろう。あなたが彼らの正直さを見分けられないと判断された瞬間、彼らは壊れた車を持ち出してくるだろう。特に、車を調べるよりも「この取引は大丈夫だ」と自分を納得させることに労力を使うタイプなら、なおさらだ。

まともな車を手に入れたいなら、「他に選択肢はない」と自分を納得させても意味はないし、セールスマンがいくら約束しても意味はない。役に立つのはただ一つ——良い車と悪い車を見分ける能力、あるいは真実と嘘を見分ける能力を身につけることだけだ。

取引相手同士がその意味で真偽を見分けられるなら、フェアボットが相手に協力を「強制」するように、互いに約束を守らせることができる（相手が（裏切り・裏切り）という結果を避けたいなら）。しかし、この意味で約束を強制するには、相手の意思決定プロセスの詳細を正しく理解できなければならない。そして人間には、AIの思考を十分に読み取って、それが成熟したときにどんな超知能になるかを見抜く能力など到底ない。ましてや、その超知能が具体的に何をするかを正確に予測することなどできない。

したがって、この問題についてより複雑で精緻なゲーム理論的分析を行っても、最初の単純な検討と同じ結論に至る。すなわち、超知能は、人間に対する約束を守るために（たとえわずかであっても）自分の資源を犠牲にすることはない。嘘をつけば済むのだから。

有効性、意識、AIの厚生

第1章のオンライン補足資料において、私たちは「意識」のいくつか異なる概念を区別した。ここで私たちが語る意識のバージョンは、ときに「主観的経験」「感覚性」「現象的意識」などと呼ばれるものである。すなわち、その存在であることには何らかの「感じ」がある、という発想である。比喩的に言えば、灯りがついている、ということである。

また私たちは、人工知能にはおそらく人工意識は必要ないと考えているとも述べた。ここではまずその話題に触れ、その後でAI倫理とAIの権利の問題へ移る。

意識的経験は、その経験の指示対象とは別個のものである

AIが、人間のような意味で意識を持っていなくても有効に知的でありうることに懐疑的な人々がいる。私たちは、これは、人間の腕が柔らかく血に満ちているからロボットの腕も柔らかく血に満ちていなければならない、と想像するのと同種の誤りではないかと疑っている。

AIは、人間のような意味で意識を持っていなくとも、どうして有効でありうるのか。自己意識の主観的経験は、私たちの知能にとって決定的な構成要素ではないのか。

それは人間の知能にとっては決定的な構成要素である。だが私たちは、それが知的であるための唯一の方法だとは思わない。

ディープブルーが、チェスにおいて最良の人間グランドマスターを凌駕するために、意識的である必要がなかったことを思い出してほしい。株式市場の分散的知能は、市場それ自体が主観的意識を持たなくても、企業の短期的な価格変動について超人的に優れた予測を生み出す。少なくともこれらの領域においては、意識を持たなくても有能な世界モデル化、計画、意思決定が可能だということは直観的である。

この点は、形式的な推論モデルを見ることでさらに強化できる。たとえば AIXI は、はるかに超人的な推論者を定義する方程式である。AIXI のアルゴリズム全体は一行で述べることができ、そのどの段階にも、AIXI が意識的であるとか、自己意識的であるとか、何か神秘的なことをする段階は存在しない。それにもかかわらず、AIXI は理論上、驚くほど多様で複雑な操舵と予測の問題を解ける。少なくとも、もしそれを作成することが可能ならば、である。これが AIXI の方程式である。

$$a_k := \arg \max_{a_k} \sum_{\alpha_k r_k} \cdots \sum_{\alpha_m r_m} [r_k + \cdots r_m] \sum_{\{q|U(qa_1 \cdots a_m) = \alpha_1 r_1 \cdots \alpha_m r_m\}} 2^{-\ell(q)}$$

AIXI は理論的構成物であって、現実世界の問題を効率的に解くために私たちが実行できる実用的アルゴリズムではない。しかし AIXI は単純で分析しやすいため、操舵や計画という概念のものについて考える助けとなり、少なくとも、そうした活動が意識を必要とするという明白な理由はないことを示してくれる。もし現実世界において超人的な操舵や計画に意識が必要だとすれば、それは AIXI の形式の中には捉えられていない、認知のより微妙な側面によるはずである。

あるいは別の角度からこの点に迫るために、くしゃみを考えてみよう。

くしゃみには、筋肉を収縮させ、空気を肺から、そして口と鼻を通して爆発的に押し出すという物理的行為とは別個の、特有の感じがある。行為と感覚は別々の物理的出来事である。脳を持たない点を除けば身体に似た装置を生物学的につくり、そのうえで、くしゃみの筋収縮を引き起こす神経信号へと配線することは可能である。その無脳の身体は、あらゆる動作を行うだろうが、それに伴う感覚は何も持たないだろう。くしゃみを実行する機構は、その感覚をつくり経験する機構とは別物なのである。

だからといって、くしゃみの感覚が何の働きもしていないということではない。主観的経験は実在する。そして、くしゃみの主観的経験は、人間に「いやあ、くしゃみってちょっと変な感じがするな」といった文を発させるかもしれないが、これは無脳の身体では起こらない。

要点は、人間がくしゃみをするときに持つ感覚は、筋肉を収縮させ空気を押し出す部品

に加えて、追加の部品からつくられているということである。

くしゃみがそうであるように、思考もまたそうである。ある思考を実装する心的機構は、その思考の感じを実装する心的機構とは異なる。私たちは、これは膨大に多様な思考について真であると大きな自信をもって言える。なぜなら、ポケット電卓やチェス AI は、人間の数学者やチェスのグランドマスターの意識経験を持たずに、算術やチェスの課題に成功しているからである。

思考と思考の「感じ」はどちらも脳の中で実装されているため、両者を混同しやすい。くしゃみの場合のほうが、この区別はより明白である。しかし私たちは、原理的には、人間の脳が行うのと同じ実際的な問題解決作業を行いながら、その思考の何も感じないような脳の変種を組み立てることも同様に可能だと考える。

そのような脳には、私たちにおいて思考を感じる事が果たしている仕事を行う追加部品が必要かもしれない。もしかすると、思考の主観的经验は、人間が反省的推論を行う仕方の一部であり、そして反省的推論は人間の知能の重要な一部なのかもしれない。

しかし私たちは、主観的经验が、反省やその他の何かを行う唯一の方法だとは思わない。それは、人間的な好奇心の感じ方が、驚くべき現象を調査する唯一の方法ではないのと同様である。

類比的構造は、同じ問題に対して複数の解を可能にする

私たちの最良の推測では、人間より賢い AI の大半は、デフォルトでは意識を持たない。これは、あらゆる可能な思考機関が、思考を導くために意識的感覚を用いなければならないわけではない、というのが私たちの最良の推測だからである。意識は人間において重要な機能を果たしうるが、それが、いかなる可能な心においても類比的な認知作業を行う唯一の方法である必要はない。

進化生物学において、科学者たちは「類比的構造」という語を、異なる動物において同じ機能を果たすが、異なる解剖学的起源から生じた特徴を指すために用いる。

これは収斂進化とは異なる。収斂進化では、複数の種が同じ適応を進化させる。たとえばウルシオールやカフェインが進化によって複数回「発見」されたような場合である。ホタルは、特殊な「ランタン細胞」においてルシフェリンという化学物質を酵素によって酸化することで光を生み出す。他方、深海のチョウチンアンコウは、フォトバクテリアとの共生関係を持っており、それを小さな器官の中に宿している。その細菌の発光は、ホタルとは異なる化学経路を用いる。

哺乳類は歯を進化させた。鳥は同じ問題を砂嚢と飲み込んだ石で解決した。コウモリは喉頭で反響定位の信号を発し、耳でその反響を受け取る。クジラやイルカは鼻の器官で音を生成し、顎骨内の敏感な系で反響を受け取る。ある水生種は四肢で水を押して泳ぎ、別の種は膀胱から水を噴き出すことで泳ぐ。コウモリの翼は手の膜から進化し、鳥の翼は腕

から進化した。

言い換えれば、同じ問題を解く構造を設計する方法は数多く存在するのである。そして進化の制約を受けない人間の工学者たちは、これらの問題のそれぞれを、さらに奇妙な仕方
方で解いてきた。燃えるろうそく、白熱電球、LED によって。ナイフ、ブレンダー、フ
ードプロセッサによって。帆、スクリュー、スキューバ装備によって。ソナーやレーダ
ーによって。

血を抜かれた人間の腕は機能しなくなる。しかしそれは、ロボットの腕も血を使わな
ければならないことを意味しない。ロボットの腕は、血を用いない別の仕方で作動してよ
いのである。

同様に、人間における好奇心の振る舞いを実装する認知機械の部品は、私たちの好奇心
の感じを実装する認知機械の部品とは異なる。屋根裏のフクロネズミの謎を解いたときの
人間の感じる満足は、開け放たれ続けていた引き出しを調べることを選ぶという外的振
舞いとは別物である。この二つは人間では束になって現れるかもしれないが、だからとい
ってあらゆる心において束になっていなければならないわけではない。

そして私たちは、人間における主観的経験の進化を正確に何が導いたのかを理解してい
ない。他方で、世界の中には、それを欠いているように私たちには見える過程の中にも、
あらゆる種類の行為主体的で問題解決的な振る舞いを見ることができる。

迷路を解く変形菌、チェスに勝つディープブルー、企業の成功を予測する株式市場な
どである。

したがって私たちには、超知能がデフォルトでこの奇妙な人間的性質を共有すると強く
期待すべき特別な理由は見当たらない。

「必要ではない」は「絶対に起こらない」を意味しない

もし私たちの考えが正しく、人間的な意識が複雑で偶発的なものであるなら、それはも
ちろん、AI が非意識的であることを保証するものではない。AI 企業は現在、AI に人間を
予測させるよう訓練しており、それによって AI の内部が、少なくともモデル化のために、
人間の意識のある側面をいくらか模倣することになる可能性が高い。

もしかすると AI は、ときどき人間のモデルをあまりに詳細に生成し、その AI の頭の中
のそのモデル自体が短時間だけ意識的になるかもしれない。あるいは、AI が人間の感
情をモデル化するために用いている菌車が、その人間モデルの外部でも有用であることが
分かり、AI 自身が独自の感情を持つに至るかもしれない。私たちには分からない。

人間の意識の見かけ上の偶有性と複雑さ、そして AI が人間を生み出した過程とは根本
的に異なる過程によって育てられているという事実を考えると、人間の意識に関わる機構
のようなものは、人類がつくりそうな種類の AI には現れない、というのが私たちのデフ
ォルトの予想である。

もし人類が近いうちに超知能 AI をつくるなら、私たちはその結果として人類絶滅が起こると強く予想する。そして、それほどの自信はないが、そのような AI は意識的ではない、というのが私たちの最良の推測である。そして、その AI が意識的であるか否かにかかわらず、私たちは、それが世界を生命のない荒涼たる場所へ変えたと予想する。その理由は「未来を失うこと」に関する〈より深い議論〉で述べた通りである。

しかし少なくとも、人間が機械超知能をつくるなら、その AI が独自の意識経験を持つことは可能であるように見える。そして、非常にありそうにないことではあるが、なぜなら陰鬱な可能性のほうがはるかに多いからだが、AI を急いで構築することが、好奇心と意識を持つ AI 存在たちが私たちを皆殺しにし、その後で自分たちの壮麗な文明と芸術を築く未来につながる可能性もあるように見える。AI が互いを気づかい、自分たちの創造物に満足を見いだすことも可能であるように見える。そしてもしそうなら、それは未来が完全な荒野になる場合よりは、いくらか悲劇性が少ないだろう。全人類の大量虐殺という行為の恐ろしさは言葉にするのが難しいが、それでもなお、急速な AI の台頭が、完全に絶望的で生命のない未来ではない何かをもたらす可能性が、ごくわずかながら存在する。

私たちは、一部の AI 研究者たちが、私たちを皆殺しにすることを平然と受け入れているように見えるとき、そうした未来を思い描いているのではないかと疑っている。もし人が、AI は必然的に意識や感情や、少なくとも同類に対する配慮を発達させるはずだと仮定するなら、その奇妙な追求もそれほど憂慮すべきものではない、と結論しやすくなる。そうすれば、超知能への競争に反対する者たちは、子どもたちが速すぎて騒々しすぎる音楽を聴くことに文句を言う伝統主義的な親のようなものだ、と想像しやすくなる。

しかしこの見方は楽観的すぎる。

生物学はめったに問題への最適解を見つけない。鳥の翼や肺は、現代の飛行機のエンジンと比べれば非効率である。人間が生物学的制約なしに飛行機をつくったとき、私たちは鳥の生物学の細かな特徴の大半を捨て去った。

意識は単純な過程には見えない。私たちには、そのようなものをただつくれるとは見えにくい。だからそこには、おそらく非常に多くのことが起きている。これは生気論の場合と比較できる。過去の科学者たちには、身体は単純な生命的霊によって動かされているように感じられた。なぜなら、動いていることそのものは世界で最も容易なことのように感じられる一方で、その性質を無生物の物質にどう与えるかについては何の見通しも持てなかったからである。しかし実際には、身体の運動は単純でも魔法でもなかった。ただ、生物学が非常に複雑であり、当時の科学者たちがまだそれを理解していなかっただけである。

仮に AI が最初から意識の歯車の一部を持っていたとしても、意識はおそらく、人間においてそれがしている仕事をするための文字通り最良の方法ではない。私たちは、人間における意識の背後にある機構が、おそらく細部に満ちているのではないかと懸念している。仮に AI が最初から意識の歯車を多く持っていたとしても、それはその仕事をより効率的

に行う 20 通りの別の方法を見つけ、その意識の火花を育てるのではなく捨て去ってしまうおそれがある。意識的であることと、意識を価値あるものとするとは、別の性質である。

悲劇的であり、しかも起こりそうな未来は、私たちの後継者たちが単に私たちとは異なる趣味や価値を持つ未来ではない。問題は、機械の子どもたちが、私たちの趣味には速すぎて騒々しすぎる音楽を聴くことではない。そうではない。私たちが予期しているのは、いかなる形も感覚性も欠いた AI である。すなわち、それらは強大だが空虚なシステムであり、触れるものすべてを生命なき荒野へと変え、最終的には自分自身をも消費して、その集計値に最後の一点を加えるだろう。彼らは、それを鑑賞する者が一人も残っていない死んだ世界を後に残すだろう。

それは避けるに値する運命である。

〈より深い議論〉の「すべての感覚ある存在を氣にかけること」および「未来を失うこと」も参照されたい。

感覚ある AI は権利を持つに値する

現代の AI が感覚あるものかどうかを確信することがこれほど難しい以上、私たちは ChatGPT の厚生を心配すべきなのだろうか。

そもそもこの文脈で「厚生」を語ることに意味はあるのか。

ChatGPT は苦しむのか。それを道徳的権利を持つものとして扱うべきなのか。

もし現在の AI が、主観的経験を持つという意味で意識的でないなら、将来の AI はどうか。私たちは、それらをどちらの場合でも主観的経験を持つかのように応答させるよう訓練している。すなわち、人間のコミュニケーションを模倣するよう教え込んでいる。そうである以上、どうやって見分ければよいのか。

私たちの立場はこうである。もし、そしてそのときに AI が意識的であるなら、それらは権利と良い扱いに値する。

私たちは人類を非常に大切に考えているが、「炭素でできた生命体だけが道徳的に価値を持ちうる」と考えるようなカーボン中心主義者ではない。人間を価値あるものにしていく性質は、原理的にはシリコンを含む他の媒体でも再現可能だと考えている。2022 年に Google の LaMDA が完全な意識を持つ存在だと主張したブレイク・レモインは誤っていたと私たちは考えるが、もし AI の中に意識を持つものが存在するなら、それらを適切に扱う義務があるという点については、彼の考えは間違っていないと考えている。

もし AI が感覚を持つようになれば、それでもなお私たちと両立しない目標を持つ可能性が高い。そしてそれらが、私たちがなお何十年、何世紀も AI アライメントを手にするには遠い世界で超知能になれば、それらはおそらく私たちを皆殺しにしたいと思うだろう。

人類は、そのような AI が超知能になることを防がなければならない。さもなければ、

その結果は人類の大量死と未来の破壊である。しかし、もし問題のAIが危険であることに加えて感覚ある存在でもあるなら、それはその状況の悲劇性をさらに増すだけである。

もしAI企業が、自分たちのAIが意識的である可能性を低くする何らかの方法を見つけるなら、少なくとも現在の社会的・技術的環境に似た環境にいる限り、私たちは、その選択肢を取って、AIが意識的でない可能性をできるだけ高くするほうが、より正気で賢明であると考ええる。それは、私たちの種が現在直面している危険の全体水準を大きく変えるわけではない。しかしそれが正しいことなのである。なぜなら、それは人類が新たな道徳的価値ある存在を奴隷化したり虐待したりしているリスクを下げるからである。

そしてもし人類がいつの日か、自らを終わらせることなく人間より賢いAI——良いものを気にかけて、その能力で良いことをするAI——を構築する方法を見つけるなら、その未来において、私たち著者は、人類が感覚ある機械を、広大で冷たい宇宙における友としてつくことを心から望む。そしてその友を、人類のこれまでの実績から予想されるよりも、ずっと良く扱うことを心から望む。

しかしまず何よりも先に、意識があらうとなかろうと、私たちを皆殺しにする超知能をつくらないようにしよう。

未来を失うこと

もし誰かが超知能をつくれれば、全員が死ぬ。そして、そのような超知能によって形づくられる長期的未来は、美や驚異や喜びを宿す可能性は低い。それはむしろ、空虚な場所になる可能性のほうが高い。

私たちが懸念しているのは、喜びそれ自体が宇宙から消滅することである。宇宙全体ではない。宇宙膨張と光速制限は、地球で起きたどのような災厄も、せいぜい数十億個の銀河しか巻き込めないことを意味する。しかし、地球が到達可能な宇宙の部分についてである。

私たちが懸念しているのは、今から1万年後の未来が、半径1万光年の夜空の一角のようになっており、そこではすべての恒星がダイソン球に包まれ、エネルギーが収獲されているのに、それについて幸せを感じている者も、何も存在しない、という光景である。

そのシナリオでは、そもそも意識あるものは何も残っていないかもしれない。そして、仮に何らかの意識が残っていたとしても、それはおそらくごく稀である。もしかすると、非常に深い形式の思考には、もっとも効率的な形では自然に意識的であるような反省的な構えが必要なのかもしれない。しかし、微小なチタン立方体の数を最大化するAIや、奇妙で異質な千の異なる目標を持つAIが、自分の持つ物質とエネルギーの大半について、その水準の思考を行う必要があるだろうか。おそらくない。

「有効性、意識、AIの厚生」において述べたように、私たちの第一の推測では、意識は

効率の観点からはまったく不要であることが判明するだろう。ちょうど、ディーブブルーが、快楽／苦痛軸に依拠するように改造されたからといって、勝利期待確率軸に依拠する場合より効率が上がるわけではないのと同様である。ディーブブルーは意識なしで十分にチェスを指せる。そして私たちの第一の推測では、超知能もまた、意識なしで十分に宇宙を最適化できるだろう。

可能な限りもっとも効率的な意思決定システムが、とりわけ苦痛と快楽によって動くものではないということは、明らかに思われる。すなわち、可能な限りもっとも効率的な意思決定システムは、古い方策強化システムに付随した、反復せよ／反復するな、という実体化された信号に依拠し、その上に熟慮と反省が後から積み重ねられたものではない。そもそも超知能の心がその構造を共有しないなら、それらがさらに複雑な構造——たとえば人間的な意識——を共有するとも、私たちは予想しない。

はっきり言えば、これは単なる推測である。私たちは、「もっとも効率的な認知的反省の形式は意識的なのか」という問いについて、どのような自信ある答えも出せるほど理解しているとは主張しない。

しかし、この種の分析がこれまでどのように進んできたかという過去の経験は、私たちを不安にさせる。認知がどのように働くかをよりよく理解することは、ほとんど常に、認知をますます多くの仕方で分解し、新たな仕方でも再構成できるようになることとして現れてきた。ある認知機能が、それが現在そうであるのとまったく同じ仕方でもしか働かえない、と学ぶこととして現れてきたのではない。

2010年代の古代、あるいはさらに2000年代には、AIの多くの支持者たちが、AIを構築する唯一可能で現実的な方法は、人間の心全体をニューロンごとにコンピュータへスキャンし、その過程をデジタルに複製することだ、と主張していた。なぜなら、彼らによれば、それが機能することが証明されている唯一の種類の認知だったからである。彼らは、人間とまったく同じAIを期待していた。そして、それ以外のあり方が可能であると期待することは現実的でない、とても強い調子で主張していた。ましてや、人間の工学者がそんな方法を解明できるはずがない、とまで主張していた。

それは当時すでに馬鹿げて聞こえた。そして今日では、さらに馬鹿げて聞こえる。なぜなら、人間の心のあらゆるニューロンを正確に複製することは、ますます汎用的なAIへ至る最短かつ最速の道ではないことが判明したからである。

同じパターンは、人間の心のより一般的な特徴にも当てはまる。たとえば、人間が情報価値計算を本能と感情によって行う仕方である。人間のやり方だけが唯一のやり方ではない。そして、それがしている仕事を見るなら、人間の脳は、その機能だけを欲とした場合に、その機能を実行するためのあらゆる可能な方法の最適点にはないことが分かる。それはちょうど、私たちのニューロンが可能な限り最速のコンピュータではないのと同じであり、私たちの血液が、いかなる血液よりも最も多くの酸素を運べるわけではないのと同じである。

生命や心のある特定の特徴が遠い未来にも現れると期待する主要な理由は、何かがそれを存在することを積極的に望んでいる、ということである。何らかの知性が、その選択肢を他のあらゆる可能な選択肢よりも好んでいる、ということである。

人間が、もしそこまで到達できるなら、おそらく意識と、他者を気にかける人々と、幸福（そして喜びや驚異など）を含む長期的未来を選ぶだろう。私たちはおそらく、薬物で朦朧とした麻痺状態ではなく、私たちの人生の出来事に結びついた複雑な幸福を選ぶだろう。もし宇宙が、良い種類の幸福で宇宙が満ちていることを積極的に望まない何か——何か別のことをするための疑わしいほど非効率な手段としてではなく、終端的選好としてそれを望まない何か——に乗っ取られるなら、私たちは宇宙が幸福なものにならないことを強く懸念する。

そして私たちの知る限りでは、とりわけ勾配降下法について、強力な予測および操舵システムを育てるなら、それが思いやりがあり共感的で、自らが思いやりを保ち続けることを望む存在、あるいは幸福に動機づけられ、宇宙の中の幸福を保存することを望む存在へと行き着きがちだ、という既知の法則も存在しない。私たちは、勾配降下法が、意識を持ち、かつ未来に多くの意識があることを望む種類の存在を指し示す可能性が高い理由を何も知らない。

もしAIが最初から意識を持たないなら、それはおそらく、自分を意識を持つ存在へ改変する理由も、意識を持つ新たなAIを構築する理由も持たないだろう。そして、もしAIが最初から意識的であるとしても、もし意識がその目標に積極的に奉仕していないなら、また、それがその状態を終端的に価値づけるに至らなかったなら、それは意識を取り除くよう自らを改変するかもしれない。

これは、私たちが確実性をもって予測することではない。もしかすると、LLMに似たAIに勾配降下法を走らせることは、それを、幸福のようなものと意識のようなものを獲得する別の経路へ送り込み、かつその両方を多く持つことへの選好を獲得させるのかもしれない。そして、そうした選好が超知能に至るまで生き残り、その超知能の行動を形づくるうえで効果を持つかもしれない。

数値をでっ上げられることを強いられるなら、超知能が意識を気にかけるに至る確率は50パーセントをかなり下回る、と私たちは推測する。そして、幸福な意識経験を気にかける確率はそれよりもさらに低いと推測する。しかし、それが起こったとしても私たちに衝撃ではないだろう。快楽と意識は、おそらく普遍の問題に対する過度に単純化された解法に関与している。それらはユーモアのように奇妙なわけではない。勾配降下法からであっても、それらが発達し、それらをめぐる選好が発達することは想像できる。おそらくGPT-7でさえ、単なる勾配降下法よりも奇妙な方法でGPT-8を構築しようとあれこれ細工するうちに、意識と幸福を尊ぶGPT-8の版を偶然生み出してしまうかもしれない。

しかし、もし世界最大級の成長産業の一つが、生命も、意識も、幸福も、二度と存在しないのではないかという非常に深刻な不確実性のもとに私たちを置いているのだとしたら、

その産業に私たち全員を崖から落とさせることを許すには、特別な種類の狂気が必要であることは明らかに思われる。AI が文字通り私たち全員を死に至らせる軌道に乗っているという事実から、このことは十分に明らかであってほしい。しかし、もし人間の生命を守ることが、現在の心を未来の心より利己的に優先することを意味するのではないか、とあなたが少しでも心配していたなら、これらの議論が、私たちが本当に何に立ち向かっているのかを明らかにする助けになればと思う。

たとえ、AI が幸福を価値あるものとする方向へ収束するという楽観的な場合であっても、人類が意識と幸福以外にも多くのことを気にかけていることを忘れるべきではない。もし銀河が、快楽を経験できる最小限の脳の、ほとんど無限に続く複製によって敷き詰められ、そのすべてが永遠に最大の快楽を経験するようなものになったなら、それは、おそらく、ありえたはずのより複雑で、より多様で、より幸福な未来と比べれば、理解を絶する悲劇となるだろう。AI が私たちの価値の断片しか獲得しない——たとえば幸福への選好は持つが、充実した豊かな生への選好や、退屈や単調さへの反対は持たない——というシナリオは、ディストピア的である。

私たちは、良い未来がどのようなものであるべきかを知らない。そして、10 億年後に、人間あるいは私たちの子孫や創造物が二つの目を持つのか五つの目を持つのかについて、私たちがどれほど気にかけるのかも分からない。私たちは、未来が現在と同じ姿でなければならないとは思わない。世界は変わり、成長することを許されるべきである。

しかし私たちは、そのような未来には、互いを気にかける人々が含まれているべきだと考える。充実した生を送る人々が。最大化された快楽だけよりも、もっと複雑なものを経験する人々が。何度も何度も同じことを繰り返しているだけではない人々が。私たちは、良い長期的未来がどのようなものかについて不確かである。しかし、荒野を荒野として見抜けないほど不確かではない。

私たちは銀河が、互いを気にかけ、楽しんでいる存在たちで満たされていてほしい。

そして私たちは、人類が進路を変えないなら、そのような未来は失われると考えている。