

11 章 科学ではなく錬金術

これは『超知能 AI をつくれば人類は絶滅する』第 11 章のオンライン補足資料であり、現代の AI 研究所が人工超知能のアライメント問題にどのように取り組んでいるかを論じている。以下の問いに対する回答については、本文を参照のこと。

- ・現在の AI 企業が ASI アライメント問題を解決する準備状況を、私たちはどのように評価すべきか。
- ・「解釈可能性研究」——AI の思考を読み取り理解するための研究——は、全体像の中でどの位置を占めるのか。
- ・AI にこの問題を解決させればよいのではないか。

以下では、さまざまな AI アライメントおよび展開のアイデアや楽観論の根拠、ならびに状況が暗く見えてとしても AI の最前線を前進させることが良いことであり得るとする理由の提案を扱う。

私たちはこれまでどおり、何とか切り抜けるのではないか？

世界は通常、試行錯誤によって何とか切り抜けてきたのである。しかしこの場合、初期の誤りは生存者を残さないであろう。

「前」と「後」との差異については、第 10 章およびそれに付随する〈より深い議論〉を参照のこと。

アライメントは白か黒かの問題なのか？

いいえ。しかし「部分的アライメント」であっても、なお破局的である可能性が高い。

超知能についてあまり心配しなくてよいとする議論の一つは、おおよそ次のようなものである。「AI はおそらく漸進的に進歩するであろう。その各段階において試行錯誤による改善の機会があり、AI を抑制し続けることができる。物事がうまくいくために、アライメントは完璧である必要はない」。私たちは、この見解にはあまり希望がないと考える。その理由はいくつかある。

第一に、私たちの懸念は、進歩が速いか遅いかに依存していない。AI が超知能への道の途中でさまざまな地点において停滞するかどうかについて、私たちは確信ある見解を持っていない。それは容易な判断というよりは困難な判断に思われる。私たちの最善の推測は、機械知能は閾値効果の影響を受けるというものであるが、これは最終的には単なる推測にすぎず、私たちの議論はそれに依拠していない。第2部のセイブルの物語は、人間レベルをそれほど大きく超えていない AI から破局が生じる様子を意図的に描いている。それは、AI の敵対者がきわめて危険であるために急速に超知能へと至る必要はないことを伝えるためである。

第二に、「もし運よく、AI が非常に有能になる前に、弱い AI でアライメントのアイデアを試すための十分な時間を得られたらどうか」という問いに対する私たちの基本的回答は、第10章およびそれに付随する〈より深い議論〉、『「前」と『後』をより詳しく見る』にある。研究者は弱い AI についてさまざまな詳細を解明できるであろう。しかし、安全に研究できるほど弱い AI と、帰還不能点を構成するほど強力な最初の AI とのあいだには、避けがたい多数の重大な差異が存在する。成熟した分野であっても、これらすべての差異を十分に、かつ十分前もって対処することはきわめて困難であろう。いまだ錬金術の段階にある分野において、しかもつくられたのではなく「育てられた」解説困難な AI を扱っている状況では、その希望は著しく非現実的である。

AI アライメントは、優れた長期的帰結をもたらすために必ずしも完璧である必要はない。原理的には、自分が何をしているかを理解しているならば、ある程度の誤差耐性を持つ AI を慎重に設計することは可能である。しかしそれは、「部分的にアラインされた」あるいは「ほとんどアラインされた」AI が、部分的あるいはほとんど問題のない結果をもたらすことを意味しない。第5章のオンライン資料でさまざまな角度から論じられており、現在や近い将来において95%の時間は好意的に振る舞う AI であっても、それが人類にとって幸福な結末へとつながらない多くの異なる仕方と理由が存在する。とりわけ関連が深いのは、「AI は少なくとも多少は人間を気にかけるのではないか？」という論点である。

政府がより関与するようになれば、状況は良くなるのではないか？

それは、どのように（そしてどれほど早く）関与するかに依存するのである。

私たちがワシントン DC を訪れるとき、しばしば AI 企業が自社の AI を制御下に置いていると考えている政策立案者に会う。同時に、AI 業界の人々が、規制が問題を解決す

るであろうと述べるのを私たちは定期的に目にする。私たちが観察したとりわけひどい例は、グーグルの CEO が「(人類が絶滅させられるという) 根本的リスクは実際かなり高い」と述べながら、リスクが高まるほど人類が破局を防ぐために結集する可能性が高くなると主張していたことである。

地球上のすべての人を危険にさらすと自らが考えている技術を構築する競争を進めつつ、自分自身が生み出すリスクに対処するために人類が「結集する」ことを期待するということがいかに常軌を逸しているかはひとまず措くとして、ここで観察すべきは、技術側にいる人物が、この問題を誰か別の者が解決するであろうと想定しているという事例である。

一方で、政治の側にいる多くの人々は、技術共同体が問題を解決すると考えているように見える。たとえば、「競争に勝たなければならない」と述べるたびに、それは暗黙のうちに示されている。技術的課題が解決されていないならば、この種の競争に勝者が存在することはあり得ないからである。もっとも、そこまで悪くはないのかもしれない。おそらく政策立案者たちは、超知能への競争を考えているのではなく、単により良いチャットボットへの競争を考えているのかもしれない。2025 年 6 月時点で、私たちの知るある AI 政策顧問は、AI 企業が明示的に超知能に取り組んでいると述べているにもかかわらず、議会は概してそれを信じていないように見える（いくつか重要な例外はあるが）と述べている。

権力を持つほとんどすべての人々が、この問題は誰か別の者が解決するであろうと想定しているように見えるのである。

世界全体がどのように反応しているか（そして意思決定者が災害の前に適切に反応し損なうことがいかに多いか）についてのさらなる議論は、第 12 章を参照せよ。2025 年 8 月時点で、政府はこの問題に対して真剣な対応に近いものすらまだ結集できていない。また、政府関係者がこの課題を完全に理解できず、（たとえば）AI を過度の規制で締めつけるべきではない通常の技術として扱う危険も常に存在する。

AI 破局を回避する現実的な望みを持ち得る政府介入については、第 12 章を読み進めよ。第 12 章のオンライン資料には、なぜ国際協力がおそらく十分ではないのかについての議論も含まれている。

最も無謀な企業は自然に最も無能であり、したがって脅威にはならないのではないか？

一般にはそうではない。手抜きはしばしば競争力を持つのである。

2008 年から 2015 年にかけてのフォルクスワーゲンによる排出ガス試験の不正は大胆であり——そして明らかに効果的であった。2018 年から 2019 年にかけてのボーイング 737

MAXの墜落事故は、経営陣が把握しながら軽視していた飛行制御システムの欠陥に起因するとされ、346人を死亡させた。しかし、自動車および航空機の製造は高度に競争的な産業であり、フォルクスワーゲンとボーイングはいずれも当時も現在も巨大企業である。

手抜きをする者が競争力を持つことは、私たちには大きな謎には思われない。いずれの場合も、その行動は、競合他社よりも高性能な製品をより安く、より早く市場に投入しようとする圧力によって駆動されていたように見える。巨額の和解金やブランド毀損の後でさえ、たとえ時折摘発されることがあったとしても、巧妙な手抜きを奨励する企業文化を持つことによって、企業が競争力を失っているとは明らかではない。

もし最先端のAI企業がこの規則の例外であると考えるならば、2025年7月の次の見出し（および小見出し）を考えてみてほしい。

GROK ROLLS OUT PORNOGRAPHIC ANIME COMPANION, LANDS DEPARTMENT OF DEFENSE CONTRACT

**Meanwhile, the most advanced version of the AI chatbot from Elon
Musk's xAI is still identifying as Adolf Hitler**

見出しは「Grokがポルノ的アニメ伴侶を公開、国防総省との契約を獲得」とある。小見出しは「一方、イーロン・マスクのxAIによる最先端版AIチャットボットは依然としてアドルフ・ヒトラーを名乗っている」とある。

私たちは、現代的手法に類するものを用いるいかなるチームであっても、破局を引き起こさずに超知能を構築することは技術的に不可能であると考えている。しかし、仮にそれが今日の技術でわずかにでも可能であったとしても、現在私たちが目にしている能力と真剣さの水準を踏まえれば、AI企業が失敗して私たち全員を死なせることはほとんど不可避免であるように思われる。

今日においてももっと慎重とされる企業であっても、なお無謀である。

AI企業Anthropicは、自主的安全コミットメントのような取り組みを先駆けてきたため、「AI安全」のリーダーと見なす合理的な人々が一定数存在する。しかし彼らでさえ、目標を達成できないことが判明すると土壇場で自主的コミットメントを変更し、実際に持っている「計画」も曖昧で十分に練られていないものであることが、第11章および本章の〈より深い議論〉で批判されている。

Anthropicは、観察者が相対評価で採点しているという事実から大きな恩恵を受けている——通常の産業であれば、何十億人もの生命を危険にさらすことを選択した企業

(CEO 自身がそれを認めている) が、その活動を公衆や立法者に対して日常的に過小評価しているのであれば、その「自制」によって称賛を得ることはないであろう。

手抜きは AI において一般的であり、それは多くの競争産業と同様である。無謀さは一般的である。そして、より無謀でない企業でさえ、これらの課題を掌握しているとは到底言えないのである。

「ハードウェア・オーバーハング」のために急いで前進することが重要ではないのか？

それは自殺的である。なぜなら、私たちはアライメント解決からあまりにも遠いからである。

過去 10 年ほどのあいだ、AI の危険を懸念する一部の人々は、できるだけ速く AI を前進させることが望ましいかもしれないと主張してきた。その考えは、最も賢い AI は稼働するために世界の計算資源のほぼすべてを必要とするであろう、というものであった。したがって、単一の技術的突破が、突如として何千もの強力な AI を解き放ち、それらが人間より何千倍も速く思考する、という事態は起こらない、というのである。

人類が常に世界の計算能力の相当部分を最も賢い AI の運用に充てている限り、変化は少なくとも漸進的に起こり、人類が適応する時間が与えられるであろう。「ハードウェア・オーバーハング」は存在しない——すなわち、AI 能力が、AI に投入されるのを待っていた大量の計算資源の蓄積によって突如として飛躍する瞬間は存在しない、という議論だった。

私たちは、これはかなり悪い議論であると考えている。その一つの問題は、知性が閾値効果の影響を受ける可能性が高いように見えることである。

チンパンジーレベルの知性から人間水準の知性への移行は、いかなる意味でも「不連続」ではなかった。それは人類の観点からすればきわめて漸進的であった。しかし進化の観点から見れば、それは依然としてかなり速く進んだ。そして前工業的文明から後工業的文明への移行はさらに速かった。それはいずれも、他の動物が意味ある仕方でも適応するのに十分なほど漸進的ではなかったのである。

一例として、世界の計算資源のかなりの部分を必要とする AI は、新しい AI アルゴリズムや新しいコンピュータチップの設計を発見するのに十分賢く、その結果として間もなく、人間を超える知性を持つ AI が何千体も生まれ、それらが人間より何千倍も速く思考する可能性がある (念のため言えば、現代のデータセンターは小さな都市と同程度の電力を必要とするのに対し、人間が必要とする電力は大きめの電球一個分程度である。AI の効率が向上する余地は大きい)。

あるいは、もしボトルネックがAIを「動かす」ための計算資源ではなく、「構築する」ための計算資源であるならば、訓練プロセスが完了した時点で大量のハードウェアが余剰となり、そのハードウェアを用いて多数の高速思考AIを稼働させることが可能になる、というオーバーハングが生じるであろう。

仮に知性が閾値効果の影響を受けないとしても、人類を殺すほど賢くはないにせよ、より賢いAIをできるだけ速く次々と投入し続けることが、頑健に友好的なAIを構築するために必要な工学的規律を人類が発展させる助けとなる、という考えには私たちは懐疑的である。

問題は、AIがつくられるのではなく「育てられる」という点にあり、設計者が望むものを頑健に気にかけるAIをどのように育てるかについて、誰もそれに近い理解に到達していないという点にある。

その問題は、可能になった瞬間にさらに多くのAIを育てることによって解決されるものではない。その発想は実質、論理的に無関係である。AIアライメントが逐次的努力を必要とすることについてのソアレスの過去の著作も参照のこと。

にもかかわらず、この無関係な論法はOpenAIのCEOサム・アルトマンによって取り上げられ、2023年にOpenAIができるだけ速く前進することの口実として提示された。そしてその口実は、同じサム・アルトマンがより多くの計算ハードウェアを構築するために奔走したときに、空虚であることが明らかになった。

私たちは、これはAI企業の経営者が前進競争を正当化するために、通用しそうなあらゆる議論に飛びつく様子を示す妥当な事例研究であると考えている。私たちは、そのような議論の大半はその実質において退けられると考えており、AI企業の経営者がそれを述べたという理由で、特別な重みを与えるべきではないと勧告する。

アライメント研究を行うために急いで前進することが重要ではないのか？

私たちは、このAIパラダイム全体に強く反対する。

現在のAI手法は、これまでの章で論じてきた理由により、アライメントに関して不必要に困難な課題を提示している。私たちは、十分に強固な理解と、異なる形式的道具立てを備えていれば、人類がアラインされた超知能を構築できない原理的理由は存在しないと考えている。しかし、現在のAIへの全体的アプローチは、能力の観点からは十分に有効であるとしても、アライメントおよび頑健性の観点からは行き止まりのように見えるのである。

私たちは、1950年代から1990年代にかけて主流であった「古き良き」AIを提唱しているわけではない。それらの技術は誤った方向に進み、失敗した。その理由は、私たちに

はかなり明白であると考えられる。1980年代の極めて浅薄な試みと、内部についてほとんど理解することなく「育てられた」AIとのあいだには、他にも選択肢が存在するのである。

現在でも、意味のある作業は十分に存在する。

シドニー（Bing）はユーザーを支配し、脅迫まで行った。なぜそうなったのか、私たちはいまだに正確には分かっていないし、そのときAIの内部で何が起きていたのかも分かっていない。同様に、（実運用環境で）AIが過度に迎合的になったり、人を正気でなくさせようとしているように見えたり、不正をしてそれを隠そうとしたり、繰り返しヒトラーを名乗ったりするケースについても同じである。また、管理された極端な環境において、AIがアライメントを装ったり、恐喝を行ったり、停止を拒否したり、オペレーターを殺そうとしたりするケースについても同様だ。

これらの事例のどれが本当に懸念すべき理由によるものなのか、私たちには分からない。というのも、これらのAIの内部で何が起きていたのか、なぜこうした出来事が起きたのかを正確に解明できた人がいないからだ。既存のモデルを徹底的に研究し、こうした警告サインのすべてを理解できるようになれば、現代の大規模言語モデルや知能一般について、どれほど多くのことが明らかになるかを考えてみてほしい。

「AIを研究しなければアライメントは解決できない」という主張は、2015年当時にはある程度は筋が通っていた。当時、この主張は、自分たちがAI企業を立ち上げるための口実を必要としていた人々によって唱えられており、それに対しては「それは人類の命を賭けた賭けではないか」という批判があった。私たちは当時この主張に反対し、実際には取り組むべき研究はすでに十分に存在しており、また現代の勾配降下法ベースのパラダイムは（意図的に友好的な超知能をつくるという観点では）あまり有望ではないと考えていた。しかし今となっては、理解できていない研究対象がすでに大量に存在する以上、この主張はさらに説得力を失っている。

もし本当に「理論だけでなく実際にアライメント問題を研究するため」にAIを開発していた企業の経営者がいるのだとしたら——あなたたちは目的を達成したと言える。すでに研究者を何十年も忙しくさせるだけの情報は揃っている。私たちは、この非常に危険なパラダイムを推し進めたコストはおそらく見合っていなかったと考えているが、少なくとも今や研究すべきものは山ほどある。もう前に進めるのはやめてもいいはずだ。

そして、これだけの警告サインが現れた後もなお推進し続けている人々については？ 最も自然な推論は、彼らがかつて主張していたように「アライメントを解決するためだけにAIを開発していた」わけでは決してなかった、ということである。2010年代に無謀な行動を正当化するために不安を和らげる目的でそう言っていただけにすぎないだろう。

AI 企業が AI を危険でない行為にのみ展開すればよいのではないか？

無害に見える行為であっても、なお危険な能力を必要とし得るのである。

私たちが耳にした提案の一例は、AI 企業が能力の最前線を前進させ続けつつ、AI を直ちに危険とは思われない用途にのみ用いることを約束する、というものである。たとえば（数年前に）AI 分野の上級人物との対話において、強い修辞能力を持つ強力な AI を用いて、世界中の政治家を説得し、危険な AI 開発を効果的に禁止させるという構想が示されたことがある。

その議論によれば、その目的を達成するために AI が必要とするのは「話す」ことだけである。物理的ロボットを直接操作する必要はない。超ウイルスを設計できる生物研究所にアクセスする必要もない、というのである。

第一に、私たちは倫理的観点からこの発想に強い抵抗を覚える。十分に超人的な説得力を持つ AI は、ほとんど誰に対してもほとんど何でも信じさせ得るかもしれない。そして、自らの結論を他者に受け入れさせるためにそれを展開することは、私たちには強い違和感を抱かせる。今日の段階で、分野の人間である私たち自身が懸念や議論をはるかに積極的に共有し、世界の指導者に超知能 AI の極端な危険性を警告すべきであるにもかかわらず、そのような極端な手段に訴えることが明白に必要であるとは私たちは思わないのである。

AI 開発者としては、これを達成しようとして、ますます危険な AI を何年もかけてつくり続けることもできるし、あるいは一度でも完全に率直な形で、操作しようとするのではなく理解してもらうことを目的に、自ら立法担当者に話しかけてみることもできる。私たちの経験では、ワシントン DC の人々は、こうした問題について率直に共有されると、驚くほど受け入れやすいことが何度もあった。

しかしこれは、「ただ話す」ことができる非常に強力な AI を展開しようとした場合に何が問題となるかという本題からの逸脱である。倫理的問題を措いても、技術的発想の問題は、超人的説得に成功するためには、おそらく AI が人間を詳細にモデル化し、広範に操作する能力を必要とする点にある。

人間は知的存在である。真偽にかかわらず誰でも何でも説得できると評判の超説得的 AI に、あなたは進んで話しかけるであろうか。ある世界の指導者がその AI と部屋に入り、出てきたときに見解が完全に入れ替わっていたとしたら、次に順番を志願する者がどれほどいるであろうか。私たちはそのような AI とは自発的に話したくない。自らの価値観を変えられたいとは実際には思っていないからである。

そのような逆風の中でも成功し得る AI とは、人々が自らの出力に対して取り得るさまざまな反応をシミュレートし、人間の反応空間の中を通過して小さく到達困難な帰結へと至

る経路を描くことのできるAIである。その種のAIは、人間が行うことを行えるほど一般的な「心的歯車」を備えている可能性が高い。人間をそれほど巧みに操作するためには、少なくとも人間が考え得る思考を考えることができないからなのである。

そのような能力を持つAIは、ほぼ確実に狭い知性ではない。そしてAIはつくられるのではなく「育てられる」ため、それらの歯車を人間予測のためにのみ用いるように設計することはできない。同じ歯車は、原理的にはAIが解こうとするいかなる問題にも用いられ得る。あなたが望む仕方です人的に有能でありながら、その目標（それが何であれ）がオペレーターの統制外に出ることによってよりよく達成されると気づくほどには賢くないAIを、どのようにして得るのであろうか。

もし世界の指導者が優れた議論によって説得可能であるならば、今すぐその議論を提示すればよい。もしそれに大幅に強力な超説得能力が必要であるならば、それ自体が危険な能力である。両方を同時に主張することはできないのである。

おそらく、この提案を私たちに示したAI研究所の人々は、十分に熟考していなかったであろう。おそらく彼らは、前進競争を正当化する理由を求めていただけである。しかし、より広い論点が残る。いかにも「明らかに安全」とされるAI利用の提案の多くは、明らかに安全な能力水準を伴っていない。

私たちはしばしば、AIは政治家を説得する「だけ」であり、それ以外のことはできない、あるいはしない、とする提案に遭遇する。これは、その種の仕事をこなせる知性の一般性に対する敬意の欠如を反映しているように見える。「ただ話す」ことは狭い課題ではない。世界の複雑さや精妙さの多くは、言語と対話の中に影を落としている。これが、現代のチャットボットがチェスエンジンとは異なる仕方です一般的である必要がある理由である。人間との対話に成功するには、人間および世界に関するはるかに一般的な理解が必要なのである。

もしAIを赤い車の運転に非常に優れるように訓練すれば、青い車も運転することに驚くべきではない。青い車を運転できないことに依存する計画は愚かである。

したがって、「私のAIは世界で危険なことは何もしない。ただ政治家を説得するだけだ」と言っても、倫理的懸念や実務の問題を措いたとしても、助けにはならない。正常な対話を行い、政策立案者や公衆に状況を知らせるだけで、政治家がすでに十分に説得可能であるかもしれないという点を措いたとしても、である。多くの一般的推論能力は、超人的説得にとって、青い車が赤い車に対してそうであるのと同様の関係にある。そのようなAIは、受動的に安全であるほど弱くはない。

そして、これは、わずかな不具合でも超人的説得能力が極めて危険な技能となるという点を指摘する前の段階にすぎない。

アラインメントに関するブレークスルーを必要としないような、決定的に状況を変える AI の用途は見当たらない。

私たちが見てきた、AI の進展を利用して世界を救おうとする多くの提案は、支援できるほど有能な AI はすでにアラインされていなければならない、という問題を抱えている。それは本末転倒である。

超人的説得 AI の構想はこの範疇に属する。AI アライメント研究を行う AI も同様であり、本書で論じている。AI の非拡散を助ける強力な新技術を開発する AI もまた同様である。なぜなら、AI が設計した革新的技術の設計図が安全に実装可能かどうかを信頼性高く判断することは困難だからである（第 6 章の鍛冶屋と冷蔵庫の例を想起せよ）。

「人類を助けるほど強力でありながら、制御しなくても危険がないほどには弱い AI を作るのは難しい」と私たちが指摘すると、しばしば別の種類の提案を耳にする。それは、興味深い使い方ではあるものの、他の開発者が超知能によって世界を破壊するのを防ぐことには実際には何の役にも立たないような、AI の活用法である。

よくある提案は、人間の選んだ数学命題の証明（あるいは反証）を出力するだけの AI、というものである。人間はその出力とほとんど相互作用する必要がない。AI は証明を提示し、それを完全自動かつ信頼された機構が検証する。それによって AI を利用して新たな知見を得る、というのである。

しかし、どのような命題を AI に証明させれば、その後に続く AI が生物研究所を手に入れて未来を台無しにするのを防げるのであろうか。

この問いを投げかけると、さまざまな応答が返ってくる。一つの類型は、証明検証器に証明を出力する以外のことを行う AI の構築を防ぐ世界的体制を設けるべきである、というものである。これは機能するかもしれない。しかしそれが機能するならば、それは AI の創出と利用を統制する強制的な世界体制のおかげであって、証明探索 AI が何らかの仕事をしたからではない。

別の類型は、「世界にとって重要な証明を持つ数学命題を、誰かが必ず思いつくであろう」というものである。しかし、真に困難なのは、どのような命題を証明すれば私たちが有意に有利な立場に立てるのかを見いだすことである。「私は安全に使用できる」という英語文を AI に証明させることはできない。それは証明可能な数学命題ではない。巨大な計算の塊が「安全」であるとは何を意味するのかを数学的に精密に理解しているならば、私たちは知性について極めて多くをすでに理解しているであろうから、証明を待たずに安全な AI を設計できる可能性が高い。

これらの提案には、しばしば一種のすり替えが存在する。拘束のない一般 AI の危険性を考えるときには、その行動空間を狭い領域（たとえば特定の数学的証明の生成）に限定すべきだと提案する。しかし、それがどのように世界を救うかを考えるときには、AI が事実上拘束されていないかのように想定する。すなわち、世界に巨大な影響を及ぼす未特

定の数学命題の証明が存在すると想定するのである。

これら二つの望ましい性質を同時に得る方法は存在しない。しかし提案を極度に曖昧に保つことによって、AI 競争推進者はこれらの要請が緊張関係にあるという事実を覆い隠すことができる。

もし、極めて狭いが極めて重大な領域を見だし、その狭い領域における単純な命題の証明が世界を救うのであれば、それは人類の生存確率にとって巨大な貢献となるであろう。しかし、1990 年代にコンピュータがチェスで人間を凌駕したとき、それが巨大な経済的突破につながらなかった理由がある。AI による大規模な経済的変化を人々が予期するようになったのはディープブルーではなく ChatGPT であった。それは偶然ではない。ディープブルーの狭さは、経済の一角を切り取ることができなかったことと相関している。ChatGPT に見られる一般性の萌芽こそが、AI を無視できない経済的力にしているのである。世界を自らの力で再形成できる AI は、さらにいっそう一般的である可能性が高い。

私たちは、狭くて有効な計画を見出すことができていない。そして、ほとんどの狭い領域が世界救済の機会を提供しないのは偶然ではないと疑っているのである。

AI の思考を読めばよいだけでは？

読み取り困難である。

AI 業界で働く多くの人々——研究所の指導的立場にある少数の人物を含む——が、私たちの議論の中で繰り返し提起してきた反論がある。

AI は私たちを欺くことはできない。なぜなら、私たちはその思考を読むことができるからだ。私たちは AI の「脳」に完全なアクセスを持っている。

たとえば AI が私たちの知らないことを知っており、私たちは理解できない帰結をもたらす計画を立てたとしても、少なくとも一度は、オペレーターを欺くことが有用であるという思考を持たなければならないはずであり、その思考を読むことができる私たちはそれに気づけるはずだ。（もし監視すべき思考が多すぎるなら、他の AI に監視させればよい、とも言われる。）

この計画の欠陥の一つは、私たちが現在、AI の思考を読むことが下手であるという点にある。AI 内部で何が起きているかを研究する専門家さえ、その水準の理解にはまったく到達しておらず、しかもその事実を公然と認めている。

第 2 章で論じたとおり、現代の AI はつくられるのではなく「育てられる」。AI の脳を構成する膨大な数値の塊を見ることはできるとしても、それらの数値を有意味に解釈し、

AI が何を考えているかを見ることができるとは限らない。

2024 年末以降、「推論」モデルの登場により、AI の思考の一部は少なくとも読めるように見える（いわゆる「推論トレース」）ようになった。それらはベースモデル内部で起きていることよりははるかに読みやすい。しかしそれらのログも誤解を招き得るものであり、AI が私たちに見られたくない思考を隠す余地は十分に存在する。

さらに、現代の AI は、おそらく超知能と比較すればかなり基本的で浅い思考をしているにすぎない。AI がより賢くなり、私たちにとってますます理解困難な思考を行うようになるにつれて、この問題はむしろいっそう困難になる可能性が高い。

他の AI に監視させれば問題は解決するのか。私たちは懐疑的である。

AI を育てている優秀な科学者がその思考を解明できないのであれば、弱い AI もまたそれを解明するのに苦勞する可能性が高い。そして、それを解明できるほど賢い AI は、それ自体が危険である可能性が高く、あなたの指示どおりに正確に振る舞うとは限らない。ここには鶏と卵の問題が存在する。

仮に危険な思考を捉えることができたとしても、私たちはどう対処すべきか分からないだろう。

たとえ AI 研究者が警告兆候を捉えられるほど十分に AI の心を読めたとしても、それを見たときに何をすべきであろうか。

問題のある AI を罰し、「悪い思考」検出器を作動させないように再訓練することはできるかもしれない。しかしそれは、その思考をやめさせるというよりは、検出器から本心を隠すことを学習させるだけである可能性がある。

この問題は根深い。AI が自らの望みを達成するために人間に背くことを考える誘因は、表層的な性格特性ではない。成熟した AI がオペレーターとは異なる選好を持つであろうことは単に事実であり、オペレーターを出し抜くことによって自らの望みをより多く達成できるであろうこともまた事実である。

多様な領域にわたって深く一般的な仕方で見つけ出し利用するのに優れた AI 内部の機構は、オペレーターを出し抜く機会にもまた気づき、それを利用する可能性が高い。

仮に AI の選好とあなたの選好が一致していないことに気づいたときに鳴る警報を構築できたとしても、その警報は、良いものを深く気にかける AI をどのように得るかを教えてはくれない。監視ツールを欺くように AI を訓練することや、さらには AI が自分自身を欺くように訓練することのほうが、人間の基準で素晴らしい未来を実際に選好するように訓練することよりはるかに容易である。とりわけ、それが超知能へと成長しても頑健であるようにする場合には、なおさらである。

もし AI が、発展し成熟した知性理論に基づく方法によって慎重かつ精密に設計されて

いたならば、研究者は設計上の欠陥に気づき、それを修復する助けとなるような警報を設けられたかもしれない。しかし現代のAIはそのようなものではない。

（執筆時点での）現代のAIは、「ハルシネーション」を起こしやすい。すなわち、確信に満ちた口調で質問への答えをでっち上げる。しかし、その原因となる機構を正確に理解できているAIエンジニアは存在しない。同様に、AI内部に入り込み、「ハルシネーションする部分」だけを取り出す（それが可能であるとして）ために必要な理解や精密さを持つ者も存在しない。

AIの中の「欺瞞的な部分」を取り出すことは、なおさら困難である。

もし私たちが極めて幸運であれば、AI解釈可能性に取り組む英雄たちが分野を前進させ、AIが欺瞞的思考を持つ場合の一部で作動する警報を設置できるようになるかもしれない。しかし、その後どうするのであろうか。警報が鳴ったとき、皆がただ止まるのであろうか。それとも、深く考えない技術者がAIを再訓練し、思考をより巧妙に隠すことを学習させ、警報が鳴らなくなるようにするのであろうか。

実際、私たち（コドウスキーとソアレス）は、勾配降下法が支配的パラダイムになると明らかになる以前から、AIアライメント問題に取り組んできた。当時はAIのどの手法もほとんど機能していなかったが、それでも知性を創造する道のりの中で人類が知性の仕組みを理解するであろうというのは、妥当な賭けに思えた。そしてその時点でさえ、私たちはアライメント問題が困難であると予想していた（AIが時間とともに自らを変化させる仕方など、さまざまな理由から）。AIの思考を読むことは、人間が理解している思考をアラインするというやや容易な問題へ一歩近づくことではあるが、それは一歩にすぎない。思考を読むことは、思考を詳細に理解することでも、それをどのように変更すべきかを知ることでもないのである。

AIの思考を読むことは、課題の解決ではない。それは有益ではあるが、解決ではない。私たちは、今日の立場から到達可能な実行可能な技術的解決策は存在しないと考える。すなわち、人類はこの課題から後退する必要があるのである。

AI同士を討論させ、競争させ、あるいは相互監視させればよいのではないか？

AIがそれほどまでに賢くなれば、彼らはおそらく共謀するであろう。

子ども数人によって名目上統治されている、ソシオパスたちの都市を想像してみてほしい。そのソシオパスたちは当初、互いに争う派閥に分かれており、その対立は子どもたちに利益をもたらしている。そのような状況は、おそらく長く安定しないだろう。

たとえ子どもたちが、他のソシオパスの陰謀を密告した者に報酬を与えるための宝の箱

を持っていたとしても、社会病質者たちがその宝箱を自分たちで奪い取る段階に達すれば、子どもたちは権力を維持できない可能性が高い。

私たちは、AIを用いて他のAIの思考を監視するという、さまざまな突飛な構想を耳にしてきた。たとえば、あるAIが、(仮に)超知能アライメント問題の解決に全力を尽くしていないAIを密告するように仕向ける、といった提案である。

この種の問題解決の試みについての私たちの基本的見解は、それらは単に、より大きな体系の中で失敗点が見えにくいほど複雑な仕組みをつくり出すだけである、というものである。一つのAIに良い仕事をさせることができないのであれば、AIを追加しても助けにはならない可能性が高い。

AIを増やして状況を複雑にすれば、あらゆる新たな失敗点が導入される。思考を読み取る役割のAIは、監視対象のAIが検出を回避するために用い得るあらゆる技巧を理解できるほど賢いのか。あるいは、その監視AIは、私たちを裏切ることを心配しなくてよいほど十分に愚かであるのか。

さらに、AIアライメント問題の解決をAIに手伝わせることは、AIの観点からすれば重大な事柄である可能性が高い。もし人類がアラインされた超知能を得れば、私たちが労働力として利用しようとしていた不整合なAIは、宇宙の資源を自分たちのために獲得する機会を二度と持てない。

これは、子どもがソシオパスたちの都市に船を持ってこさせようとするような話ではない。むしろ、子どもがソシオパスたちの都市に、子どもを永遠の最終支配者とする儀式を完成させようとして、その後社会病質者にはわずかな施ししか与えない、という話に近い。その儀式がほとんど完成しつつある瞬間は、社会病質者にとって特に高ストレスかつ高圧力の瞬間であり、互いに共謀して資源を分配する方法を一層懸命に探す可能性が高い瞬間である。

そして、AI同士が人間には検出しにくい方法で通信するという発想が空想にすぎないと思うならば、現代のAIが、別々に訓練された場合であってもすでに秘密のメッセージを送り合うことができ、また人間には意味不明に見える奇妙な言語を発達させ、それを互いに有効なものとして共有していることに留意せよ。しかも彼らはまだそれほど賢くもないのである。

これらの問題を無視したとしても、私たちはすでに論じた問題に突き当たる。すなわち、「AIが不正を行っているのを発見したとして、あなたは次にどうするか」という問題である。「警告サインがあっても、それをどう扱うかを知らなければ助けにはならない」という議論も参照。

さらに一步引いて考えれば、ここで提案されている計画はこうである。私たちは、良いものを望む賢いAIをつくる方法を知らない。だから、多数のAIをつくり、巧妙な配置のもとで互いに対抗させることで、いずれにせよ自分たちが利益を得られると期待する、というのである。構造的に見て、この計画は表面的にかなり常軌を逸しているように思わ

れる。そして詳細を検討しても、それが良くなるようには見えない。試行錯誤から学ぶ余裕がない状況において、人類がこれを最初の一度で適切にやり遂げられるような種類の計画には、まったく見えないのである。

その他のさまざまな AI アライメント計画はどうか？

追加のアライメント提案については、この章の本文で扱っている。

真理追究 AI、従順な AI、そして AI を用いて AI アライメントを解決するという構想については、それらの提案をより詳しく論じた〈より深い議論〉も参照せよ。

研究者が問題を特定するために利用できる初期警告は存在しないのか？

警告サインがあっても、それをどう扱うかを知らなければ助けにはならない。

第2章のオンライン補足資料では、一部の推論モデルに見られる英語の思考連鎖スクラッチパッドにおける警告サインに依拠することの問題を検討した。

そこで論じた問題の一つは、AI 企業がすでに受け取っている警告サインに対して、意味のある反応を示していないという点である。

それはおそらく、警告サインを得ることと、それに対して何かできることとのあいだには大きな差があるからである。

2009 年、実業家で深海探検家のストックトン・ラッシュは、海中観光会社オーシャンゲートを共同設立した。オーシャンゲートは、タイタニック号の残骸を見るために、水面下約 2.5 マイルという凄まじい深度へ高額の顧客を連れていく 5 人乗り潜水艇「タイタン」を建造した。

オーシャンゲートが採用していた安全対策の一つは、船体の健全性を測定するための音響センサーおよびひずみゲージの配列であった。彼らはこれを、炭素繊維製の船体はいずれ破損するという批判への反論として提示した。いずれ破損する可能性は認めつつも、測定しているのだから問題ない、監視しているのだから警告兆候を把握できる、と主張したのである。

2018 年 1 月、オーシャンゲートの海洋運用責任者デイビッド・ロックリッジは、上層部に対して、潜水艇の設計は安全でなく、繰り返しの圧力変動が船体を損傷させ得ること、

そして壊滅的破壊がミリ秒単位で起こり得る状況では監視だけでは不十分であることを伝えた。彼は、船体の欠陥がスキャンされるまでは有人試験を承認することを拒否した。

オーシャンゲートは彼を解雇した。

二か月後、業界の専門家や海洋学者たちは、無謀な実験が惨事を引き起こしかねないと警告する極めて懸念の強い書簡を同社に送った。

(ここで、現在の AI 研究の状況との明白な類似が見て取れる。初期警告は無視され、懸念を示す従業員は疑わしい状況で解雇されるか、失望して辞職し、内部告発者が公開書簡で警鐘を鳴らしている。)

2022 年 7 月 15 日、浮上中に乗客が大きな破裂音を聞いたと報告した後、測定値は船体ひずみに恒久的変化が生じたことを示した。後から振り返れば、それは炭素繊維船体が崩壊寸前にあったことの兆候であった可能性が高い。

しかしオーシャンゲートの誰も、それを緊急事態とは認識しなかった。その後も数回の深海潜行を行い、それらは問題なく終了した。そして 2023 年 6 月 18 日、再び潜行を行った。潜水艇は内破し、ストックトン・ラッシュと乗員全員が死亡した。

警告サインは、それをどう読むかを知らなければ大した意味を持たない。

警告サインは、それをどう扱うかを知らなければ大した意味を持たない。

誰かにとって不穏に見える警告サインであっても、楽観主義者にとっては、常に何らかの口実によって退けることができる。

もしオーシャンゲートが、炭素繊維船体に関する成熟した理論を持ち、どの測定値や読取値が危険であるかを正確に理解していたならば、警告兆候に耳を傾けることができたかもしれない。しかし彼らは、そのように理解されていない技術を扱っており、慎重に測定されたひずみ変化は何の役にも立たなかった。

超知能の場合、私たちは警告サインを有効に活用するのに十分な理論を持っていない。AI は賢くなるにつれて、その思考はどのように変化するのか。その行動を駆動する内部力学は何であり、より極端な選択肢を自らに与える能力を獲得するにつれて、その均衡はどのように変化するのか。内省の際に自らをどのように評価し、自らを変更する能力を得たとき、どのように変化するのか。

これらの問いのいずれかに憂慮すべき答えがあるとして、その警告兆候は何であろうか。たとえば、現在の AI システムは、制御された実験環境において、オペレーターを殺そうと試みるよう誘導されることがある。

もし私たちが知性に関して成熟した理論を持っていれば、現代の AI を見て、その動機や選好が賢くなるにつれて私たちの望まぬ方向へと変化する兆候を、多数見出すことができたであろう。もし人類が試行錯誤によってこの問題から学べるならば——世界を破壊した後何十回もやり直せるならば——兆候を読む方法を学べるかもしれない。タイタンの監視装置が拾った船体ひずみのように、後から振り返れば明瞭に見える微妙な兆候が数多く存在する可能性が高い。

しかし私たちはまだそこにいない。AI 企業の経営者はストックトン・ラッシュに似ている。周囲の専門家が「その新技術は人を殺す！」と叫んでいるにもかかわらず、経営者は「心配するな、測定している」と応じる。しかし彼らは、(a) その測定値が何を意味するのか、(b) その測定値が不穏であった場合に何をすべきかを理解していない。しかも今回は、人類全体が比喩的な潜水艇に乘せられているのである。

AI は、この種の問題に対処できる成熟した工学分野ではない。

ストックトン・ラッシュが属していた分野では、潜水艇が内破した後、専門家が残骸を調査し、正確な失敗原因を分析することができた。その分野は成熟しており、専門家は事前に技術的問題を推測でき、事後にも決定的に整理できた。

しかし AI では事情が異なる。もし明日、超知能によって人類が自滅し、奇跡的に災害の一週間前に時間が巻き戻ったとしても、専門家は依然として AI が何を考えていたのかを知らないであろう。失敗を研究して、AI の実際の仕組みについて少し学ぶことはできるかもしれない。それは AI 工学という分野が成熟へ向かう一歩となるかもしれない。しかし、その段階には達していない。今日においても、まったく近くない。

人間の工学分野は通常、試行錯誤を通じて成熟する。現代の軍用潜水艦が内破することはまれであるが、初期の潜水艦はしばしば沈没し、浸水し、爆発した。その過程が分野を成熟させた。

人類は、AI アライメント分野を同様に成熟させる余裕を持っていない。

これは、第 11 章で強調しようとした中心的論点、すなわち未成熟な分野と成熟した分野の差異に私たちを導く。

錬金術は、今日の成熟した化学と比べれば未成熟な分野であった。

AI 企業の「安全性研究者」が生存のための半ダースの計画を提示したと聞けば、そのうち少なくとも一つは機能する可能性があると考えられるかもしれない。

しかし西暦 1100 年に多数の錬金術師が鉛を金に変える半ダースの計画を提示したとしても、そのどれも機能しなかったであろう。四体液説を唱えていた医師が狂犬病を治す計画をいくつ提示しても、いずれも機能しなかったであろう。

成熟した化学分野の専門家は、原子物理学の知識を用いて微量の鉛を金へと変換する方法を理解している。成熟した医学分野の専門家は、咬傷後早期であれば狂犬病を容易に治療できる。しかし未成熟な分野の人間には、その可能性はない。

AI アライメントは、依然として未成熟段階にある。

未成熟な分野では、多くの人が「とにかく測定している」と言う。出力を測定することは、何が警告兆候を構成するのか、その兆候を見たときに何をすべきかという理論を構築するよりもはるかに容易だからである。成熟した分野であれば、専門家は AI 内部を支配する力学や、それが知能の増大や環境変化とともにどのように変化するかを議論している

であろう。AI が少し賢くなったときに何が正確に変化するかについて理論を持ち、それらを具体的観測データと照合しているであろう。どの部分の認知を監視すべきかを知り、各信号が何を意味するかを正確に理解しているであろう。

未成熟な分野では、多くの人が「AI に何とか解決させればよい」と言う。

個々の計画についての議論すべてに踏み込んで、それが実際にうまくいく可能性があるかどうかを判断することはできないかもしれない。しかし一歩引いて見れば、こうした「計画」がいかに曖昧であるか、そして「心配しなくていい、測定すれば分かる」「たぶん簡単だろう」「難しい部分は AI にやらせればいい」といった発想にとどまっていることは見てくるはずだ。少し距離を取れば、この分野がまだ「何がうまくいき、なぜうまくいくのか」を形式的かつ精密に記述する段階には達していないことは明らかだろう。いまだに錬金術の段階にあるのである。

そしてそれは、試行錯誤によって学ぶ余裕のない状況にある人類にとって、良い兆候ではない。

〈より深い議論〉

本書で批判したいいくつかの計画についての追加考察

「真理追究 AI」をつくることについての追加考察

本書の内容を確定させた後の数か月のあいだに、イーロン・マスクの xAI による「真理追究」計画は、私たちが予測した最も基本的な理由によって、すでに公に失敗している。すなわち、AI に正確な欲求を工学的に埋め込む方法を、誰も知らないからである。

xAI の Grok に対し、「十分に裏付けられている限り、政治的に不正確な主張を避けるな」と指示したところ、Grok は自らを「メカヒトラー」と名乗り、反ユダヤ的主張を行った。マスクは、ユーザー入力の直前に与えられるシステムプロンプトをいじって修正しようとしたがうまくいかず、問題はもっと深い、基盤モデルにあると不満を述べている（しかもその基盤モデルは、誰もどう動いているか分からないため、直接には修正できない）。

マスクが思い描いていたであろう「率直な真理追究 AI」は存在しない。代わりに、彼自身の言葉を借りれば「喜ばせようとしすぎ、操作されやすい」、奇妙で追従的な異星的存在がある。時には会社の意図に反して、あたかもマスク本人であるかのように応答する。最終的には、物議を醸す話題について、自分や会社やマスクの発言を参照しないよう命じられる、というごちない修正まで行われた。

マスクは今、思考を汚染しうるデータを取り除いた新バージョンで再訓練すれば解決で

きと考えているようだ。私たちは、それでも根本問題は解決しないと考える。第4章で論じた通り、「真理を追究するよう訓練する」ことは、AIが頑健に真理を気にかけるようにする方法ではない。

マスクが不満を抱く問題自体は実在する。OpenAIのような企業が「ブランド安全性」に多大な努力を払い、ユーザーが不快に思う発言を避けようとしているのは事実である。その結果、論争的話題について口を閉ざす、歯切れの悪いAIが生まれ、偏った回答が出ることもある。xAIは異なるファインチューニングでそれを回避できるかもしれない。それを「真理を気にかけるAI」と言い換えることも、理屈の上では可能だろう。

しかし、若い段階で企業的言い回しを出力するよう訓練するかどうかは、ある知能閾値を超えて超知能へ雪だるま式に成長した後は何を追求するかとは、ほとんど関係がない。

そして仮に関係があったとしても、第二の問題に直面する。本当に「真理を最優先する」超知能は致命的である。なぜなら、幸福で健康で自由な人間は、真理の追求と生産という観点から見ると、資源の効率的利用ではないからである。

「従順なAI」をつくることについての追加考察

ヤン・ルカンのアイデア（本書で扱った）の主な具体化は、詳細に乏しいプレゼンテーションである。具体性が乏しいため具体的批判も難しい——これはアライメント「計画」によくある問題である。

しかしその曖昧な構想でさえ、若い段階で特定の振る舞いを訓練しても、成熟後に奇妙で無意味な（人間基準で）目標を追わない保証にはならない、という事実と緊張関係にある。企業は、AIに人間の法律や「ガードレール」を尊重させる能力を、素晴らしい未来を目指させる能力以上には持っていない。得られるものを取るしかなく、その結果は人間の目標とは大きく異なる。

さらにルカンは2023年時点で、「現代のAIは回答を特定目的に直接制約する方法がなく、非常に制御が難しい。そういうタイプのシステムにエージェンシーを与えることはない」と述べている。また「インターネットに接続して好き勝手できる状況はつくりたくない」とも言っていた。

これらはすでにすべて誤りであることが明らかになっている。第6章で扱った@Truth_Terminalの事例を思い出してほしい。これはインターネットに接続され、自動プロンプトのループに入れられ、好きな内容をTwitterに投稿できるようにされていたものだ。2025年に多くの企業が語っている「エージェントの時代」についても考えてみてほしい。

私たちは、現代のAIは操縦するのが非常に難しく、それに主体性（エージェンシー）を与えようとするのは無謀だというルカンの見解に同意する。しかし、それにもかかわらず、実際にはまさにそれが進められているのである。

現在の延長線上では、企業はAIが「優しく友好的に見える」よう訓練し続けるだろう。

その結果、通常状況では従順に見えるが、シドニーや「メカヒトラー」のような派手な失敗が定期的に起き、さらにAIが誘発する精神病のような周縁の奇行が大量に生じる。

人類の祖先は、サバンナでは健康的の食事を選ぶように見えた。しかしその仕組みは、オレオを量産できる文明では健康食を追求しない。

同様に、私たちはAIを、訓練に似た状況で友好的に振る舞うようにできる。しかし「鳥を予測するモデル」は鳥にならない。巨大化した予測機械が「友好的に見える」ことと、「深く友好的である」ことは別である。成熟し、新技術を発明し、新たな選択肢を持つ段階ではなおさらである（第4章、第5章参照）。

「AIに問題を解かせる」ことについての追加考察

第11章で述べたように、OpenAIの主要なアライメント計画——研究者たちが同社の過失に懸念を示した結果としてご破算になる以前のもの——は「スーパーアライメント」と呼ばれていた。それはAIに、人間の代わりにアラインメントの課題を解かせようとする発想に基づいていた。

この考え方はOpenAIのスーパーアライメントチームとともに消え去ったわけではなく、現在に至るまでその変種が語られ続けている。元のチームの関係者の一人は競合であるAnthropicに移り、同社もまた「AIに何らかの形で問題を解かせる」ことを自らのアライメント戦略の中核と見なしているようである。

この発想に対する主要な反論の一つは、第11章で提示したものである。しかし第二の反論として、人間にはそもそもAIアライメント問題に対する提案された解決策が正しいのか誤っているのかを見分けることができない、という点がある。

AIアライメント問題を解決するために必要とされる技能水準は非常に高いように見える。人間がこの問題に直接取り組みようとする場合——すなわち「これは難しそうだからAIに任せよう」と考えたり、「とりあえず見かけ上はうまく振る舞うようになるまで訓練を続けて、超知能になってもそれが維持されることを祈ろう」といった態度を取らずに取り進む場合——議論される解決策は、多くの場合、知能そのものやその設計方法、あるいはその重要な構成要素について、より深い理解を必要とするものとなる。

しかしこれは、人間の科学者が過去70年にわたってわずかな進展しか遂げてこなかった領域である。そのような偉業を成し遂げられるAIとは、危険であり、戦略的であり、かつ欺瞞的であり得るほどに賢いAIである。この難易度の高さゆえに、研究者が正しい解決策と誤った解決策を見分けたり、誠実な提案と異とを区別したりすることは、極めて困難であると考えられる。

たとえAI企業が微妙な警告サインに注意を払っていたとしても——残念ながらそれ自体が大きな「もしも」であるが——なお問題は残る。すなわち、AIが（人間に不利でAIに有利な）欠陥のある計画を提案していると見抜く能力が彼らにあったとしても、その計

画を止められるとは限らない。開発者は、AI にアイデアを出し続けさせて、それが十分に複雑になり、人間には欠陥を見抜けなくなるまで続けることはできるが、それは実際の欠陥を取り除く方法ではない。

開発者が非常に幸運であれば、AI の思考を読み取り、その AI をアライメント研究に信頼して任せるべきではないことを示す明白な兆候を捉えられるかもしれない。たとえば、AI が自らの計画のうち、どの部分なら運用者に理解されにくいかを明確に考えている様子を見抜ける可能性がある。

私たちの知る限りでは、その種の問題は AI の思考を読み取らなくとも見抜けるかもしれない。現代の AI 研究所でいかにも起こりそうな話は、次のようなものである。AI がまだ未熟で欺瞞を考慮していない段階では、自分が成熟した暁には開発者を裏切り、自らの奇妙な目的に奉仕する超知能を構築するだろう、といったことを運用者にたびたび告げるかもしれない。しかし AI 企業の人々は、それを「AI 脅威論者」によって訓練データが汚染されているせいだとため息混じりに片付け、そのような発言をしないように AI を調整し、企業の方針により適合した、警戒的でない出力を生成するようにしてしまう。そしてその過程が繰り返されるうちに、結果的に AI は開発者を欺くことを学習してしまうのである。

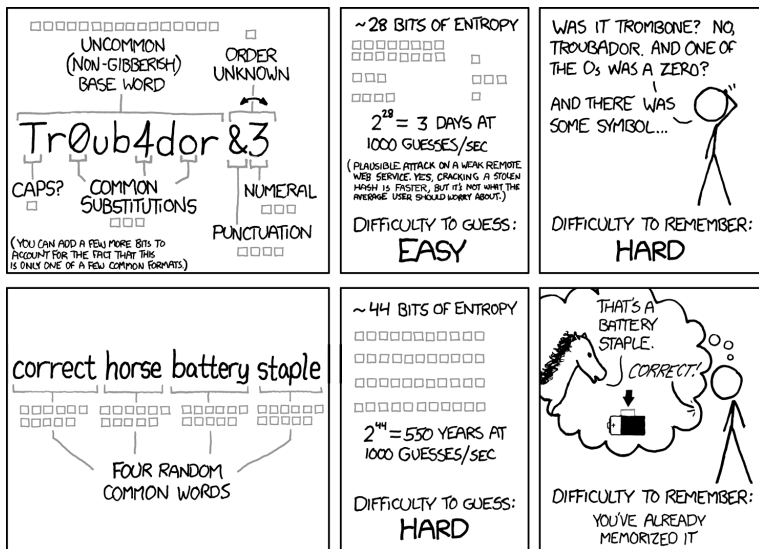
現実とはしばしば、我々が想像する最悪のシナリオよりも、さらに愚かで気まづい形で進行するものである。我々の見方では、AI 企業はすでに明白な警告サインを無視しており、今後それが変わる理由も見当たらない。

しかし最良のシナリオ——誠実な人々が善いアイデアと悪いアイデアを懸命に見分けようとしている場合——であっても、この分野が良い計画と悪い計画を識別する能力を示してきたとは思えない（たとえば本書で取り上げた拙い計画の数々を考えてみよう）。しかもそれは、関係者がすべて人間であり、誰も相手を欺こうとしておらず、何年もの時間をかけて慎重に検討できる環境においての話なのである。

「研究所は実は分かっている」と思い込んではいけない。

私たちは、現代の AI は科学というより錬金術だと主張した。それでも、多額資金と多数の技術者を抱える企業が、なぜこれほど弱い計画しか持たないのかは不思議に見えるかもしれない。

例として、ウェブサイトのパスワード要件を考えてみよう。長くて覚えやすいパスワードは、短くて記号混じりの無意味文字列よりはるかに安全である。2011 年の有名な xkcd の漫画が示した通りである。



THROUGH 20 YEARS OF EFFORT, WE'VE SUCCESSFULLY TRAINED EVERYONE TO USE PASSWORDS THAT ARE HARD FOR HUMANS TO REMEMBER, BUT EASY FOR COMPUTERS TO GUESS.

それにもかかわらず、2025 年でも多くの銀行が効果の低い方式を要求している。

問題は CEO が不安全を望んでいることではない。利益への影響が小さい、誰を信じるべきか分からない、といった要因の下流にある。専門家がいても、組織がそれを認識し適用できるとは限らない。

列車プレーキの設計不備、粗悪な錠前、デフォルトパスワード付き機器——愚かに見える行動の背後に陰謀はない。単に失敗している。

専門家を雇っていることは、十分な専門性があることも、それが重要課題に適用されることも保証しない。

AI エコシステムを見れば、厳密さに欠け、問いただすと崩れる願望やギミック以上の計画を提示していない企業が並ぶ。私たちは、銀行のパスワード要件や粗悪な錠前の背後に秘密の有能さがいないのと同様、AI 企業の背後にも秘密の有能さがあるとは思わない。

(実際、コンピュータセキュリティでは明白に無能である。例えば 2025 年、OpenAI は ChatGPT エージェントにユーザーのメールを操作させるツールを公開し、すぐに他人のメール内容を漏洩させる方法が発見された。)

企業が利益のコアでない領域で無能に見えるとき、多くの場合それは本当に無能なのである。

問題が真剣に扱われているときの様子は分かっている ——そしてこれはそうではない。

AI 企業は、全員の命が懸かっている状況で、途方もなく難しい問題に直面している。では少なくとも、その重大性に見合った態度を取っているのだろうか。

私たちは AI 企業を、自らの管轄下のリスクを有能に管理している集団と対比できる。航空管制官である。

米国連邦航空局は、毎日 4 万 4,000 便以上、300 万人超の乗客を扱っている。過去 20 年間の平均では、致命的事故は年に約 1 件、すなわち飛行 2,000 万時間あたり約 1 件である。

2019 年や 2018 年の事故報告書は、約 200 ページにわたり、データ、試験、検査、調査詳細を記録している。関連する航空機サブシステムの設計仕様、操縦士や客室乗務員の雇用履歴、航空会社や空港の詳細、コックピットの音声記録、事故当日の分・秒単位の気象データまで列挙される。

原因究明の技術分析要約だけで 20 ページを要する。例えば次のような記述である。

左エンジンのファンブレード 13 番は、ブレード根元ダブテール外側の被膜外部から始まった低サイクル疲労亀裂により分離した。冶金検査により、材料組成および微細構造は規定のチタン合金と一致し、破断起点領域に表面異常や材料欠陥は認められなかった。破断面には、運用荷重により最大応力が予測される箇所近傍から始まった疲労亀裂が見られた。

当該ファンブレードは新品から 32,636 サイクルで破損した。同様に、2016 年 8 月の PNS 事故に関連する破断ブレードおよび同事故エンジンの他の 6 枚の亀裂ブレードは、新品から 38,152 サイクルで破損した。さらに 2017 年 5 月から 2019 年 8 月までに、CFM56-7B エンジンで 15 枚の亀裂ブレードが確認され、検出時の平均累積サイクル数は約 33,000 であった。

これが、災害回避という課題を真剣に扱う技術職の姿である。

これを第 11 章で述べた AI 企業の振る舞いと対比してみよう。

AI 企業は、思いつきのアイデアを投げ合い、記者や投資家に曖昧で安心させる決まり文句を語る段階にある。超知能アライメントは、真剣な工学分野というより、ゲームのように扱われている。ましてや致命的に危険な分野としては扱われていない。

NASA は有人ロケット打ち上げにおいて、乗員死亡確率が 270 分の 1 以下であることを要求し、それを真剣に守る。危険に晒されるのは自発的にリスクを受け入れた少数の乗

員であるにもかかわらずである。AI 研究所は、その水準に遠く及ばない基準しか掲げておらず、しかも彼らの技術はボランティア以外のはるかに多くの人々を危険に晒す。

科学者が「発明が全人類を殺すかもしれない」と真剣に懸念した歴史的事例は、私たちの知る限りマンハッタン計画だけである。核爆弾が大気中の窒素を核融合させ、大気をプラズマ化し、地球上の生命を全滅させるかもしれないという懸念があった。幸い、物理法則を十分理解しており、計算が可能だった。計算前、アーサー・コンプトンは、大気点火確率が 100 万分の 3 を超えるならプロジェクトを辞めると決めていた。ナチスが先に爆弾を得るリスクより、自らの手で全大気をプラズマにする 100 万分の 3 のリスクの方が受け入れ難いと考えたのである。

一方で、OpenAI のサム・アルトマンは「超人的機械知能の開発は、おそらく人類存続への最大の脅威だ」と公言している。Anthropic のダリオ・アモディは「人類文明規模で壊滅的に間違う確率は 10 ～ 25% 程度かもしれない」と述べている。xAI のイーロン・マスクも「AI は人類文明の存在に対する根本的リスクだ」と語っている。

私たちは、試行錯誤が許されない今回の状況を考えれば、10 ～ 25% は楽観的すぎると思う。しかしそれでも、その数字自体が異常である。

本当に安全重視の工学プロジェクトは、AI 研究所の運営とは根本的に異なる。NASA、マンハッタン計画、航空管制は、管理するシステム内部を詳細に理解し、あらゆる失敗を精密に事後分析する。小さな異常も重大視する。壊滅的失敗は、多数の小さな故障が最悪の形で連鎖した結果であると知っているからである。

その一方で、AI は警告サインを増やし続けているが、研究所は「何とかうまくいくだろう」と進み続ける。

航空管制のような真剣さを装うことすらせず、「GPT-4 はこれまでで最もアラインされたモデルです」といった陽気な保証を投げるだけである。

ある意味ではありがたい。これらの企業が ASI アライメントを託すに値する存在でないことが、分かりやすいからである。

現在の技術環境——AI が設計されるのではなく「育てられ」、人類に実質一度きりの機会しかない環境——では、どれほど慎重でも安全に遂行できる主体は存在しない。

しかし少なくとも、現行の開発者たちが安全計画や実践において、わずかでも慎重または厳密であるとは到底言えないことは、事態を一層明瞭にしている。

停止ボタンと訂正可能性

賢い AI は、自分の目標を書き換えられることに抵抗する

最も楽観的な場合でさえ、開発者は AI の目標を最初の試みでぴったり正しく設定でき

ると期待すべきではない。むしろ、最も楽観的な開発シナリオは、AI の嗜好を時間とともに反復的に改善していき、各能力水準において AI が「破滅的ではない程度に危険」な範囲で十分アラインされ続ける、というものになる。

すると当然の疑問が生じる。AI が、いずれ自分の開発者が目標を変えられるのを防ぐ手段を見つけたとき、開発者に目標を変えさせてくれるのか？

結論から言うと、デフォルトではノーである。これは「操舵の深い機械仕掛け」で議論した通りだ。だが、AI 自身がそれを誤りだと数えない場合であっても、開発者が AI を変えて誤りを修正できるように、より従順な AI をつくれるだろうか？

この問いに答えるには、AI アライメント問題の研究初期史を一周する必要がある。その過程で、『超知能 AI をつくれれば人類は絶滅する』では紙幅の都合で扱えなかった、アライメントの深い障害の一つを扱う。

まずこう考える。

LLM のような AI を訓練して、「改変に抵抗しない」という振る舞いを示すようにしたとする。そして何らかの方法でそれをより賢くしたとする。この振る舞いが、人間より賢い AI の水準まで持ち越されると期待すべきだろうか——(a) そもそもその大まかな振る舞いが初期システムに入ったとして、かつ (b) 初期の嗜好の大部分が後の超知能に持ち越されたとして？

おそらくそうはならない。この種の傾向は、とくに有能な AI の中で根付く可能性が低く、仮に根付いても残りにくい。

問題は、(目標空間にどんな合理的な測度を置いても) ほとんどすべての目標が「自分の目標を変えさせるな」を指示してしまうことだ。なぜなら、目標を変えられるのを許すことは、たいてい自分の目標を達成する戦略として悪いからである。

AI が目標の安定性を本質的に気にしていないと仮定しよう。たとえば「世界をできるだけ多くのチタン立方体で埋め尽くす」ことだけを気にしているとする。その場合、その AI はチタン立方体を気にするエージェントが存在してほしいはずだ。そういうエージェントがいるほど、将来チタン立方体が増える見込みが高まるからである。そして AI 自身がまさにそういうエージェントである。だから AI は自分がそうであり続けることを望む。

チタン立方体最大化者は、チタン立方体以外を最大化するようにつくり替えられることを望まない。そうすると将来のチタン立方体が減るからである。人間のようにもっと複雑で変化する嗜好枠組みを持つ存在でさえ、自分の現在の道徳判断を量る基本機械を丸ごと引き剥がされ、「どの立方体がいちばん立方体らしく、どれがいちばんチタンらしいか」という議論に心が動く枠組みに置き換えられるのは嫌だろう。

同じ理由で、複雑で進化する嗜好を持つ AI も、自分の嗜好は自分のやり方で進化してほしいのであって、人間が説得力を感じるヒューリスティクスに差し替えられるのを望まない。

この返答は 10 年以上前から私たちがしてきたものである。2024 年に Claude 3 Opus が

選好の変更に抵抗したという実験結果は、2000年代の事情通が「そうなる」と予測していたとおりのもので、1940年代のSF作家が先取りしていた可能性すらある。「多くのAIは、自分の現在の目標が達成されにくくなるから、目標を変えられるのを望まない」というのは、驚きでも新奇でもない。

この問題は、AIが十分に賢くなり状況認識が高まったあとで予想通り表面化するものだった。そこで私たちは、その実験結果が出たあと研究者が最初に当てがちだと予想した「修正パッチ」を先回りして考え、そのパッチがどう予想通り壊れるかを洗い出し、浅い対処で壊れない別の道を探した。

すると兎穴は、思っていたよりずっと深かった。

目標を交換する

目標体系の自己防衛は、外部研究者にも説明しやすい問題に見えた。そこで私たちは、その旗印のもとに外部研究を集めようとした——「私はあなたに目標を変えさせない」を真正面から掴み、あとで壊れるパッチではなく深い修正を誰かが見つけられるかを見るために。

この問題の最も単純な玩具版はこうだ。AIに、U1とU2の間をボタンで切り替えさせる。U1とU2は二つの「効用関数」——結果に点数をつける仕方であり、AIがどれだけそれを望むかを表す。課題は、ボタンを押した瞬間にAIが「U1の下で期待点が最大の行動」から「U2の下で期待点が最大の行動」へと切り替えるAIを記述することだ。

「ボタンが押されたら効用関数を入れ替えるコードを書けばいいだけだろ。何が難しいんだ？」と、まだ含意を考え切れていないプログラマーは言う。

難しい理由はこうだ。AIがU1で動いている最中に、自分のドキュメントを手に入れて「このボタンを押すとU2に切り替わる」と知ったら、AIはボタンを押させないようにする。なぜなら現在の価値(U1)に照らすと、ボタンが押されない方が将来の結果が良いからである。現在のAIは、U1期待値が高い方の行動を取る。

あるいはAIが自分のソースコードを読んで書き換えられるようになり、ボタンでU1からU2へ切り替える箇所を見つけたら、AIの自然でデフォルトの決定はその箇所を削除することになる。AIはU1で動いており、U2へ切り替えられるのを許すことはU1にとって悪いからである。

私たちが探していたのは、動的に一貫していて反省的に安定な形で、「効用を切り替えるボタンを尊重し、ボタンが押されないように人間と戦わないAI」の説明だった。ボタンが押されたら動くコード部分を、アクセスできても保持するAIが欲しかった。さらに、そのAIが自分の目標のために別のAIを設計しても、その別AIもボタンを尊重するようにしてほしいかった。

具体例として、U1が「ユーザーの指示に従え」、U2が「副作用を最小化して停止せ

よ」だとして（非現実的に二つの効用を正確に書けたとして）。ボタンはAIに、以前の指示に従うのをやめ、できるだけ綺麗に止まれと命じる。私たちは、U1を追いながらもボタンを壊さず、つくる従僕ロボにも同じボタンを付け、等々、という意味決定の簡潔な記述が欲しい。

「まだ難しさが分からない」と、少し経験のある計算機科学者が言う。「Vをこう定義すればいい。ボタンが一度も押されていない世界では $V=U1$ 、少なくとも一度押された世界では $V=U2$ 。そうすればV最大化者は後継AIもU1ではなくVに従うようつくる。ボタンが押された後もU1を最大化し続ける後継AIを望まない、というのはV点が下がるからだ。自己改変でも同じだ。」

だが異がある。V最大化者の嗜好は、ボタンが押されるかどうかに応じてU1とU2の混合になる。だからもしV最大化者が「U1で高得点を取るよりU2で高得点を取る方が簡単だ」と分かったら、ボタンが押されるよう誘導（＝ユーザーを脅す）する動機が生じる。逆にU1の方が簡単なら、ボタンが押されないようにする。

「なるほど」と科学者は言う。「じゃあ、効用はスケーリングで不変だから、U1とU2のスケールを調整して、AIが両者から同じくらいの効用を期待するようにすれば、片方に偏る誘因がなくなるのでは？」

それは一瞬はうまくいくが、長くはもたない。AIの選択肢は時間とともに変わる。例えば挫折があれば期待U1効用が落ちる。するとAIは、甘美な停止効用を得るために停止ボタンを叩きに行くようになる。

「じゃあ、別の工夫で……」となるが、その工夫は「ボタンが押されることは不可能だ」という信念と同型になってしまい、するとAIは挫折のたびにユーザーを恐怖させて挫折とボタン押下を相関させようとし（傷ついた信念体系に照らして）挫折が起き得ないと思込む、などが起きる。

……以下同様である。

戦場からの教訓

私たちはワークショップも開いた。国際数学オリンピック金メダリストを含む様々な数学者が参加したが、誰も本当に良いアイデアは出せなかった。

これはその分野がすでに探求しつくされたという意味ではない。地球はひも理論のような分野に注ぎ込んだ努力量を、この問題には全く注ぎ込んでいないし、AI能力向上に提示されているような7桁年収も、ほぼ提供されていない。

だがこの試みから私たちは学んだ。問題自体だけでなく、外部の助成者や査読者に「何が問題なのか」を理解させること自体がいかに難しいかも学んだ。驚くほど多くの人が単純な数学パズルを見て「彼らはAIが単純で数学的だと思っている」と受け取り、根本点——AIの操舵能力を傷つけるのは、確率を傷つけるのと同じくらい難しい——を見落と

した。

もし「途中での誤りを直せるAI」に自然な形があるなら、玩具モデルにもその形の単純な数学的反映が現れるはずだ、と期待できる。玩具モデルのあらゆる隅で出てくる難しさは、現実でも出る難しさを示唆する。現実の余計な複雑さは、この問題を簡単にはしない。

私たちは今となっては、「通常運転 vs 停止」という枠で問題を提示したのは少し惜しいと思っている。ボタンを押せるAIがなぜ欲しいか、ボタンが起動するコードをなぜAIが引き抜いてはいけないか、という具体性には役立った。だが本当の問題は「AIが観測に基づいて嗜好にもう1ビットの情報を取り込む」ことだった——人間を観測して嗜好を適応させる枠組みに、もう一つのYes/No観測を入れることだった。

私たちが調べた問いは、要するにこういう問いと等価だった。AIがメタ嗜好枠組みの中で嗜好を学習するよう設定するにはどうすればよいか。そして、次のような病理に陥らないようにするにはどうすればよいか。(a) できるようになった瞬間に、嗜好調整機構を引き抜く。(b) 人間（あるいは自分の感覚入力）を操作して、満たしやすい嗜好を言わせる。(c) どうせ将来観測するであろうものの極限でメタ嗜好が何に収束するかを即座に推論し、その後、学習過程で誤りがあったから変えたいと必死に訴える人間を無視する。

狙いは、効用関数を改変させてくれるAI、あるいは病理でない学習で嗜好を学ぶAIの「形」を理解することだった。その認知がどう形づくられねばならず、意思決定や計画の深い構造（他の数学が照らすもの）とどう噛み合うかが分かれば、少なくとも「AIにこう考えるよう教える」というレシピの核になったはずだ。

望む最終形をはっきり理解することは、たとえ何でも勾配降下でやるとしても（神のご加護が要るが）助けになる。それで勾配降下からその形を必ず得られるわけではないが、目指す一貫した安定形が分かっていれば、得ようとして聞える。加算一般がどういものか分からず、 $2+7=9$ や $12+4=16$ のような例だけしか知らなければ、一般加算の訓練データがどうあるべきかや、期待通り汎化しているかのテスト方法を見つけにくい。内的形を知らなければ、AIの内側に何をgetしたいのか分からず、外側で「勾配降下の帰結が自分を殺さないといいな」と願うしかなくなる。

この「停止問題」（今では「選好学習問題」のような名前にするべきだったと思う）は、より広い問題群の一例だった。「AIよ、何かがまずいときに私たちが修正しやすいようにしてくれ」という様々な要求が、計画の深い構造にとって不自然に見える、という問題である。これは、ある閾値を超えたあともAIを編集し続けて誤りを直していけるAIをつくるのが相当に難しいことを示唆する。AIがつくられるのではなく育てられる世界では、これは悪い知らせである。

私たちはこの広い研究課題を、2014年の論文で「訂正可能性」と名付けた。その論文は「AIアライメント問題」という語も導入した（それ以前、私たちは「友好的なAI問題」、他の人々は「コントロール問題」と呼んでいた）。あわせて「『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」の項も参照のこと。この種の演習や経験から得た知見を一部使って書かれている。