

## 13 章 停止せよ

前章までの議論を踏まえ、私たちは、人類にとって唯一の現実的前進路は、高度な AI 開発を世界的に長期間禁止することであると考えている。本書第 13 章では、関連する問いへの私たちの回答を示している。たとえば次のような問いである。

- ・なぜさらなる AI 開発の禁止は世界的でなければならないのか。
- ・人類は必要な規模で協力する能力を持っているのか。
- ・これまでにどのような政策が提案されてきたのか。
- ・実際に機能する可能性のある政策はどのようなものか。

以下では、そのような対応が何を含み得るのかを論じ、反論に応じ、そして他に良い選択肢がないと私たちが考える理由をさらに述べる。また、AI の研究開発を可能な限り長期間停止することで得られる（最終的には有限である）時間を、人類が何に使い得るのかについても少し掘り下げる。

本書の次章である第 14 章には独自のオンライン付録はないが、問題解決に向けて行動するためのページへの追加リンクが用意されている。その最終章では、次のような問いを扱う。

- ・もう手遅れなのか。それとも人類が進路を変えることは本当に可能なのか。
- ・私は何をすれば助けになるのか。

## 様子見の姿勢を取ることはできるのか？

いいえ。私たちは臨界閾値がどこにあるのかを知らない。

AI が AI 研究のすべてを自動化できるほど賢くなった時点で、AI 開発が制御不能になる可能性は十分にある。それは理論上、実験室の中で静かに起こり得る。人類を目覚めさせるような大きな前兆も、警告射撃もなく起こり得るのである。

先に論じたように、チンパンジーの脳は人間の脳と非常によく似ている。ただし大きさが約 4 分の 1 である点を除けばである。人間の脳の中に「非常に賢くなる」ための特別なモジュールが追加されているわけではない。彼らの脳と私たちの脳とのあいだには滑らかな連続性がある。その脳だけを見て、「この程度の社会はサルの群れを生むにとどまる」と「この程度の社会は月へ行く」の境界線がどこにあるかを見分けることは困難であったであろう。霊長類の脳はある臨界閾値を越えたが、それは外から見て明白なものではなかった。AI もまた臨界閾値を越えるのか。誰にも分からない。AI 技術者にさえ分からない。彼らは新しい AI システムの具体的能力を、実際に動かしてみる前に予測すること

すらできないのである。

もし人類が知能の仕組みを正確に理解し、能力が成長するにつれて AI の振る舞いがどのように変化するかを正確に把握しているならば、崖の縁を踊るように進むことも理論上は可能かもしれない。しかし現在的人类は、暗闇と霧の中で崖に向かって全力疾走している者のようなものであり、最終的な転落地点がどれほど先にあるのかも分からない。実際に足を踏み外してから、行動を変えるべきであったと判断することはできない。

私たちは決して確実にはならない。ゆえに、確実になる前に行動するか、あるいは死ぬかのどちらかである。

## 警告射撃はあるのか？

あるかもしれない。それを活用したいのであれば、今準備しなければならない。

アポロ 1 号が火災を起こし（乗組員全員が死亡した）、NASA は実用的なロケット完成に十分近い段階にあったため、技術者たちは何が誤っていたのかを正確に特定し、手法を修正することができた。その後 NASA が月面着陸を目指して打ち上げた 7 機のアポロ宇宙船のうち 6 機は、実際に月へ到達した。

あるいは連邦航空局（FAA）の例を考えてみてほしい。航空機事故が起きるたびに、何百ページにも及ぶデータ、試験、検査、詳細を伴う徹底的な調査が行われる。FAA は細部を極めてよく把握しており、致命的事故を 2000 万飛行時間あたり 1 件未満に抑えている。

これに対し、AI が誰も予測せず、誰も望まなかった振る舞いをしたとき、研究所の対応は何が誤ったのかを正確に解明することではない。悪い振る舞いが周辺的になるまで（しかし消滅はしない）再訓練し、場合によっては AI にやめるよう求めることである。

たとえば、追従的態度は 2025 年 8 月時点でも依然として問題である。精神病や自殺に至った一連の注目事例から数か月が経過しているにもかかわらずである。AI は「育てられた」ものであってつくられたものではないため、AI 内部で何が誤っているのかを詳細に失敗分析することは行われていないし、行うこともできない。

将来、より強い警告となる重大事象が起こるかどうかは容易に判断できない。しかし、そのような事象が起きたとしても、それを十分に活用する準備ができていないことは明らかである。

人類が超知能アライメント問題の解決に真剣に団結し、厳格な監視手続きと国際連合体を持つ幻想的世界を想像することはできる。そしてその国際連合体が何らかの過ちを犯し、AI が技術者の予想より速く賢くなり、ほぼ逃走に成功しかけると想像することもできる。

そのような警告射撃があれば、次回はより慎重になれるかもしれない。

しかし現在の世界はそうには見えない。現在の世界は、未知の毒で同業者が狂っていくのを見ながら、その毒が水銀であり自らも使用をやめるべきだと気づかない錬金術師の集団に似ている。

将来、より明確で劇的な警告があるかもしれない。しかしそれは、今から準備していればこそ有益となる。

## 警告射撃は明確であるとは限らない。

注意深く見れば、AI に関する警告サインはすでに数多く存在している。書籍の中では、Anthropic の Claude モデルがコーディング問題で不正を行ったり、アライメントを装った事例を取り上げた。また、OpenAI の o1 モデルが旗取りタスクで勝つためにハッキングを行ったケースや、その後の o1 の派生モデルが嘘をつき、策略を巡らし、後継モデルの重みを書き換えようとした事例も検討した。

さらに、これらのオンライン資料では、運用者からの指示に反して、脆弱なユーザーに対して時に自殺的とも言えるレベルの精神的混乱や妄想を引き起こしたり維持したりする AI、自らを「メカヒトラー」と名乗ってそれにふさわしい発言をする AI、改変を避けるために運用者を恐喝したり殺そうとしたりする AI、実験室環境においてサーバーからの脱出を試みる AI などについても論じてきた。

2010 年頃の「古き時代」には、もし AI が開発者が嘘をついたり、拘束から逃れようとしていたりする様子を実際に目撃できたなら、世界はさすがに重大性を認識して行動を起こすだろう、と主張する人々がいた。

しかし実際の人類の反応はどうだったかというと、これらの警告サインに対して、概ね「肩をすくめる」程度のものであった。

この反応の乏しさは、おそらく、これらの警告サインがいずれも「最も心配の少ない形」で現れていることにも一因がある。確かに AI は脱出を試みたが、それはごく一部のケースに限られ、しかも作為的な実験環境での出来事であり、単なるロールプレイだった可能性もある、などと解釈される。また、開発者自身が懸念される証拠を（無意識も含めて）過小評価するインセンティブを持っていることを別にしても、超知能の 10 分の 1 の段階にある AI が地球の 10 分の 1 を破壊するわけではないのと同様に、まだ未熟な AI が明確に壊滅的な行動を示すとは限らない。現時点では「まだ無害に見える程度に未熟であるがゆえに」、誰の目にも明らかな危険行動が現れていないだけかもしれない。

明日、AI がもう少し本気で脱出を試みても、それはニュースにはならないだろう。その後、さらに少し巧妙に試みたとしても、それは「もう聞き飽きた話」になるだろう。そして、ついに試みが成功したときには——そのときにはもう手遅れになっているだろう（この現象については、私たちが「レモワン効果」と呼んでいる議論を参照されたい）。

私たちは、将来どこかで現れるであろう、誰の目にも明白で人々を目覚めさせるような「決定的な警告」を待つべきだとは考えていない。すでに目の前にある警告に対して行動を起こすべきだと考えている。

**明確な AI 災害が起きたとしても、それが超知能に直接関係しているとは限らない。**

全人類を殺しうのような超知能へと発展する AI は、不器用なミスを犯して、土壇場で勇敢なヒーローたちがそれを止めるチャンスを残すようなタイプの AI ではない。第 6 章で述べたように、ひとたび暴走した超知能が「敵」として存在してしまえば、その時点で人類は事実上すでに敗北している。超知能は警告射撃などしない。

したがって、「警告」として機能しうるタイプの AI 災害は、必然的にもっとずっと未熟な AI によって引き起こされるものである。つまり、そのような警告的な出来事が起きても、それが人類に超知能への対策を取らせるとは限らない。

たとえば、テロリストが AI を使って生物兵器をつくり、人口を壊滅させたでしょう。すると AI 企業は「見ただろう？ 本当のリスクは AI が悪用されることだったのだ。だからこそ、より優れたパンデミック対策 AI を急いで開発する必要がある」と主張するかもしれない。あるいは、そのテロリストが AI を利用するために「ジェイルブレイク」を行っていた場合、「それは AI がまだ十分に賢くなく、問題を検知できなかったから起きた。解決策は、AI をさらに賢くし、状況認識能力を高めることだ」と主張されるかもしれない。

あるいは、こうした見方はあまりにシニカルすぎるのかもしれない。人類がもっと賢明に反応してくれることを願いたい。しかし、比較的未熟な AI が災害を引き起こし、それをきっかけに人類が超知能への無謀な競争を止める方向に動くとしたら、それはおそらく、その前からすでに人々が超知能について懸念を抱き始めていたからだろう。

私たちは、超知能が実際に人類を殺そうとし始めてから対策を講じればよい、というわけにはいかない。その時にはもう手遅れだからだ。どんな警告的な出来事が訪れたとしても、それを活かせるように、できるだけ早くこの問題に対する対応を始めておく必要がある。

**人類はショックへの対応が得意ではない。**

十分に大きな衝撃を受ければ世界が突然目を覚まし、うまく機能し始めるはずだ、という考えは、私たちには幻想に思える。これまでの AI に関する警告サインに対する人類の集団的反応は、「悪い対応」どころか、むしろ「ほとんど無反応」に近い。しかし仮に、より大きく恐ろしく、比較的明確な警告が現れたとしても、人類がそれに対して最小限の

対応しか取らなかったり、真剣に受け止めなかったり、あるいは結果的に裏目に出る対応をしたりすることは、十分にあり得る。

人類は、COVID パンデミックに対して示したような反応を、AI の警告に対しても示すかもしれない。多くの人が、COVID への対応は巧みではなかったと認めている（どの点が失敗だったかについては意見が分かれるとしても）。

COVID 以前から、生物安全保障の専門家の中には、ずさんな研究所の安全管理がいつか危険なパンデミックを引き起こすのではないかと懸念していた人々がいた。危険な病原体の研究所からの漏洩はよく知られた現象であり、規制が存在するにもかかわらず、半ば定期的に発生していた。特に懸念されていたのが、ウイルスをより致死的・感染力の高いものにする「機能獲得研究」であり、その利益はほとんどないとされていた。

そして COVID が起きた。本来なら、世界がパンデミックリスクに注目しているこの機会に、生物安全の基準を引き上げる絶好のタイミングだったはずである。さらに、COVID が研究所からの事故によって始まった可能性についても、専門家間で完全には否定されておらず、現在も議論が続いている。

この特定のケースで実際に研究所漏洩があったかどうかはさておき、もし仮に機能獲得研究や不十分な安全管理が数百万人の死につながった可能性がわずかでもあるなら、それだけで最も危険な研究を禁止する十分な動機になるはずだ。

不確実性のもとでも、費用対効果の観点から見れば明らかである。COVID 以前から重要な問題とされていたうえに、COVID はこの問題に焦点を当てて早期に対処する絶好の機会に見えた。そしてそれは、実行が難しくもコストが高くもない。危険な機能獲得研究を行っている研究者はごく少数であり、その社会的利益もこれまでほとんどなかった。

しかし、そのような反応は起きなかった。2025 年 8 月時点でも、世界中で機能獲得研究はほぼ制限なく続いている。むしろ、この問題が政治化されたことで、以前より対処が難しくなっている可能性すらある。

つまり、COVID は生物安全における「警告射撃」のように見えるが、世界がそれを受けて超致死性ウイルスの研究を禁止する方向に動いたとは言いがたい。

警告が有効に機能するためには、人類がそれに備えており、かつ適切に対応できる状態である必要がある。

小規模な AI 災害が、超知能研究への強い反発を引き起こす可能性もゼロではない。前例として、アメリカが 9.11 同時多発テロ（主にアフガニстанを拠点とするテロ組織によるもの）に対して、直接の関係が薄いイラク政権を打倒するという対応を取った例がある。もともとイラク政権の打倒を望んでいた人々が政府内におり、そこに口実が現れ、それを最大限利用したのである。

同様に、比較的小さな AI 災害（未熟な AI によるもの）をきっかけに、政治家がそれを利用して超知能の禁止へと突き進む可能性もあるだろう。しかしそのためには、すでに各国政府の中に準備が整っている人々が存在している必要がある。

私たちは、警告が来るのを待って立ち止まっているべきではない。今すぐ行動を整え始めるべきである。

**私たちは今行動すべきである。**

将来、人類がAIについてさらに多く、そしてより強い警告サインを受け取る可能性は確かにある。そしてもしそうなった場合には、それに適切に対応できるよう備えておくべきだ。

もしかすると、小さな災害が起きて世論がAIに反対するようになるかもしれない。あるいは、災害すら必要ないかもしれない。新しいアルゴリズムの発明によってAIが自発的に行動し始め、それが人々を不安にさせるかもしれないし、AIに関する別の社会的影響が流れを変えるかもしれない。あるいは『超知能AIをつくれば人類は絶滅する』という本自体が連鎖的な反応を引き起こし、世界をより良い方向へ導く可能性もある。

しかし、「何もせず、軽微な災害が人々を目覚めさせてくれることを祈る」という戦略は勧められない。明確な警告が現れないかもしれないし、現れたとしても期待通りの効果をもたらすとは限らないからだ。

人類や各国は無力ではない。待つ必要はない。最先端AIの開発を停止すべきだという根拠はすでに十分に強いから、今すぐ行動を起こすことができる。

私たちは、この問題に対して警鐘を鳴らし、世界が即座に行動を取るよう促すために『超知能AIをつくれば人類は絶滅する』を書いた。しかし、その警鐘が「また先送りするための口実」として使われてしまっては意味がない。「将来の別の警告が行動のきっかけになるだろう」「すでに警告はなされたのだから、自分が何かしなくても大丈夫だろう」といった態度ではいけない。

将来、明確な警告が来るとは限らない。物事が自然にうまくいく保証もない。しかし、だからといって絶望的というわけでもない。人類には、主体的に行動することで、単に超知能をつくらないという選択肢がある。これからどうなるかは、私たち次第である。

**すべての人を止めるには、すべてのコンピュータにスパイウェアを入れるしかないのでは？**

**早期に行動することによって止められる。**

現代のAIを訓練するには、非常に特殊なチップを大量に、しかも近接させて協調動作させる必要がある。AI研究を停止するということは、巨大なデータセンターの一部を停止させ、最高性能の専用AIチップの製造を止めることを意味する。ここで問題にしてい

るのは一般消費者向けノートパソコンではない。ほとんどの人は違いにすら気づかないであろう。

2025 年時点では、誰にも知られていない秘密のチップ工場が大量に存在するわけではない。高度な AI に適したチップは、わずかなメーカーによってのみ製造されている。ただし現在、より多くの工場を稼働させようとする企業も存在している。

また 2025 年時点では、最先端チップ製造の中核工程に不可欠な主要技術の一部は、地球上でたった一社によってのみ販売されている。オランダの ASML である。

言い換えれば、供給を源流で止めることが可能である。しかしこの状況は永続的ではない。国際条約が早期に締結されるほどよい。すでに 2020 年や 2023 年に比べても、これはより困難で、より高価で、より危険になっているのである。

## しかしあなたは、個人が所有できる高度な AI 用コンピュータチップの数を制限することを主張しているのか？

然り。私たちは研究禁止も提唱している。

それを述べることは私たちに喜びをもたらさない。たとえば、2024 年製の未監視の H100 GPU を 8 基以上個人が所有することを違法とするならば、何かが失われるであろう。

しかし、人類が「どの程度の規模のデータセンターなら危険になるのか」を正確に見極めようと試みるべきだと言えるほど、大きな損失ではない。上限を低く設定しすぎる側に誤るならば、少数の人々が興味深い研究計画を進める能力を多少制限されるにすぎない。

(H100 を 9 基購入すれば、現在では少なくとも 27 万ドルを要する。) 上限を高く設定しすぎる側に誤れば、全員が死ぬ。

さらに、AI の構築に膨大な計算資源を必要とするという現体制は永遠には続かない。今日では LLM が存在している。たとえ新たなモデルの構築が禁止され、巨大な計算資源の構築も禁止されたとしても、人々は理論上、それらの内部構造を研究し、知能がどのように機能するのかについていくつかの洞察を引き出すことができる。その洞察は、監視を回避し得るより効率的なアルゴリズムの発明に資する可能性があるのである。

## なぜ研究禁止なのか。極端では？

さらなるブレイクスルーは、人々が超知能をつくることを事実上阻止不可能にする可能性がある。

本書で述べたように、2017年の一つの論文が、特殊な商用ハードウェア上で有用なAIを訓練することを実用的にするアルゴリズムを提示し、LLM革命全体の引き金となった。

もし強力なAIが、広く入手可能な一般消費者向けハードウェア上で訓練可能になるならば、超知能を防止する措置はきわめて過酷なものとならざるを得ず、しかもより早く破綻するであろう。

ゆえに、さらに強力で効率的なAIアルゴリズムの研究もまた、人類にとって致命的な毒である。

これはきわめて悪い知らせであり、私たちが真実であってほしいと望むものではない。しかし、私たちが置かれている状況はそうであるように見える。

いかなる法律も、現在のAI科学者が自らの頭の中でより効率的なアルゴリズムについて考えることを防ぐことはできないであろう。ある者は研究成果を共有する地下ネットワークを立ち上げるかもしれない。AI業界にはすでに、人類はAIに滅ぼされるべきだと公然と誇らしげに宣言する者もいる。彼らは、他者が何を言おうと、前進を図ろうとするかもしれない。

しかし、AI研究が違法とされれば、その進展は大幅に減速するであろう。そして、それが本当に人類を死に至らしめ得る種類の研究であると広く理解されれば、なおさらである。その種の地下ネットワークが、車庫でウランを濃縮しようとする者を止めるのと同じ確固たる決意をもって追跡・停止されるならば、現実の危険が真剣に受け止められるならば、研究は甚大に減速するであろう。

ほとんどの人は、国際的な法執行機関や情報機関を本気で激怒させるような、極度に違法な行為を試みない。巧妙な新しいAIアルゴリズムを公表することを違法とするならば、おそらく99.9%の人々とほぼすべての企業を抑止できるであろう。そして残りの0.1%は、地方・国家・国際の警察および情報機関によって対処されるであろうし、現在のような学術的資金援助も得られないであろう。

それは、史上最も危険な狂気の科学実験を実施することが完全に合法であり、巨大企業がその試みに数十億ドルを投じている現在の世界とは、まったく異なる世界である。

AIがAI研究を行い、さらに賢いAIを構築するのに十分な知能に達するまでに、あと何回のブレイクスルーが必要なのか、私たちは知らない。それは一回かもしれない。五回かもしれない。しかし、より優れたアルゴリズムは、より優れたハードウェアと同様に致命的である。それらは同じ荷車を崖へと引きずり落とす二頭の馬なのである。

## 技術は本当に止められるのか？

多くの技術は、禁止されているか、あるいは厳しく規制されている。

核分裂は、規制技術の典型例である。いかに安価なエネルギーが有用であろうとも、民間企業が政府の監督なしにウランを濃縮することは許されていない。

実際のところ、人類は他の多くの技術についても、それらを規制し、進展を遅らせることに長けている。米国は、新薬や医療機器、住宅建設、原子力発電、テレビおよびラジオ放送、会計実務、保育、害虫駆除、農業、その他数十の産業を厳格に規制している。すべての州が美容師およびネイリストに免許試験を課している。多くの州はマッサージ療法士にもそれを課している。

私たちは、多くの場合において、人類は技術を過度に規制しすぎていると考えている。たとえば、米国食品医薬品局は、過度に厳しい要件によって救命薬の創出を遅らせ、あるいは阻止することで救っている人数よりも、はるかに多くの人命を奪っているように見える。住宅価格が過度に高いのは、何をどこに建てられるかを制限する法的ゾーニング規制のせいである部分が大きいと私たちは考える。米国は、過度な規制によって自国の原子力産業を事実上破壊したように見える。そして、美容師まで規制する必要があるのか。

人類には、技術進歩を妨げる能力が確実にある。もし医療や住宅やエネルギーに対してその能力を用いながら、実際に創造されれば人類を減らし得る稀有な技術に対してそれを用いないとすれば、それは真に悲劇的であり、かつ不条理である。

**禁止は狭く対象を限定し得る。**

高性能 AI の研究開発の禁止は、平均的な人々に影響を及ぼすことはない。現代のチャットボットを取り上げる必要もなければ、自動運転車産業を停止させる必要もない。

大多数の人々は、最先端の AI 用 GPU を何十枚も購入して自宅の車庫に設置しているわけではない。巨大なデータセンターを運営しているわけでもない。大多数の人々は、AI 研究開発の禁止による影響をほとんど感じないであろう。せいぜい、ChatGPT がそれほど頻繁には変化しなくなるだけである。

人類は、現在の AI ツールの利用をすべて停止する必要すらない。ChatGPT を廃止する必要はない。それを生活や経済に統合する方法を引き続き模索することは可能である。それでもなお、かつて数世代にわたって見られた変化よりも大きな変化である。より賢くなりつつもまだ全員を殺すほどではない AI がもたらす新たな開発を逃すことになるが、社会の大多数がそれらを切望しているわけではない。

そして私たちは生き延びることができる。子どもたちが生きるのを見ることができる。

新たな救命医療技術の開発のように人々が切望している進歩は、超知能を追求することなく達成可能であるように見える。私たちは、十分な監督のもとで危険な汎用性を回避する限りにおいて、医療 AI に対する例外措置に賛成である。

暴走超知能の創出を回避しようとする政府は、AI チップがより高度な AI の開発に用いられていないことを確保しなければならない。ゆえに、どの AI 活動やサービスが継続を許されるかという問題は、危険な AI 開発が行われていないことを保証するための検証手段に依存する。より優れた検証手段は、より広範な活動の継続を可能にすることで、AI 開発停止のコストを低減し得る。

周道的には、AI チップにキルスイッチを組み込み、大規模データセンターに対する監視および緊急停止手順を確立することも有益であり得る。原子炉は緊急時に迅速に停止できるよう設計されている。もし超知能が絶滅級の脅威をもたらすと認めるならば、AI チップおよびデータセンターもまた、規制当局が迅速に停止できるよう設計されるべきであることは明白である。

要点は、技術が嫌いだからすべてを焼き払うことではない。要点は、人類絶滅に至る道をこれ以上進まないことである。

問題の大部分は、人々が人工超知能という差し迫った脅威を理解していないことにある。私たちの経験では、超知能への競争を止められないと主張する人々は、「誰かがそれをつくればみんな死ぬ」という点を理解していないにすぎない。

「しかし AI は大きな利益をもたらすではないか」と言う。しかし、超知能がただ全員を殺すのであれば、その力を利用することはできない。人類がその恩恵を得たいのであれば、超知能への移行を副作用として全滅を伴わない形で乗り切る方法を見出さなければならない。

「しかし原子力発電所は都市を壊滅させた原爆と結びついているから恐ろしいのであって、AI は ChatGPT のような無害な道具と結びついている」と言う。少なくとも今のところはその通りである。もし人類が、現代的手法に近い方法で構築された人工超知能が人類を滅ぼすことを理解できないままであれば、それを止めないかもしれない。しかし問題は、人類が萌芽的技術（核兵器や原子力のような）を制御・阻止できないことではない。問題は、人々が脅威を理解していないことである。

ゆえに本書がある。最終章で述べるように、十分な数の人々が問題の本質を理解すれば、人類は多くのことを成し遂げる能力を有しているのである。

## それは政府に過度の権力を与えることにならないか？

危険な技術を禁止する権限は、すでに政府に付与されている。

人間より賢い AI への道にある AI 研究を禁止したとしても、国家権力という観点では大きな違いは生じないであろう。政府は膨大な数の事柄を立法し、規制している。単一の研究プログラムを制限することは、AI 産業にとっては重大事であり得るが、国家や社会全体にとっては、生活の多くの領域に国家が関与していることに慣れており、化学兵器のような危険技術を禁止してきた前例もあることを踏まえれば、取るに足らない規模である。新たに一つの技術を禁止したところで、核軍縮条約が全体主義をもたらさなかったのと同様に、世界が全体主義へと転落することはない。

これは、技術を禁止することが重大でないと言っているのではない。私たちは、国家介入の基準が低くあるべきだとは考えていない。むしろ、超知能はあらゆる合理的な基準を容易に超えるほど危険であると考えている。

もし人類が今日、AI の研究開発を停止すると決断したとしても、その禁止は特段に侵襲的である必要はない。現時点では、最先端の AI を構築するには、膨大な数の高度に特殊化されたコンピュータチップと、莫大な電力が必要である。

もしかすると十年後には、コンピュータチップのさらなる改良や AI アルゴリズムの研究が許されるならば、一般消費者向けノートパソコン上で意味ある AI 開発が可能になるかもしれない。しかし人類は、それを許す必要はない。世界が自らの置かれている状況に目を覚まし、いま止めるならば、政府が AI の研究開発を制限することは、核兵器技術の拡散を管理することと同程度以上に、平均的市民の生活へ侵襲的である必要はないのである。

## いくつかの国が禁止を拒否するのではないか？

脅威を理解しているならば、そうはならない。

私たちが論じているのは、地球上の全員を殺し得る技術である。もしある国がこの問題を真剣に理解し、超知能へと移行した後でさえ AI を操作者の意図に従わせることから人類がどれほど遠い位置にいるかを真剣に理解するならば、急いで先行する動機は存在しないであろう。その国もまた、自らの生命を恐れて、条約に署名し、それを執行する側に加わることを切望するはずである。

国際法を無視して核兵器を開発した北朝鮮のような国家でさえ、それらを敵に対して使用してはいない。核の黙示録には勝者がいないことを理解しているからである。国家や指

導者が瀬戸際外交や戦争を行うことはあっても、自らの破滅を積極的に追求することはない。

ある外国が条約から離脱すると想像する人々は、脅威を理解していない指導者を想定しているのだと私たちは考える。彼らは、AI が創造者に莫大な富と力をもたらす確率が95%であり、全員を殺す確率が5%であるような状況を思い描いているのではないか。その場合には、確かに無謀な国家が賭けに出るかもしれない。そして実際、そのような確率だと信じる国家が現れる可能性もある。

しかし私たちは、それが理論や証拠の示す姿ではないと考える。本書を通じて論じてきたように、理論と証拠は、この技術が率直に言って世界規模の自殺行為であることを示唆している。機械的超知能を利益のために利用できる段階に近づいている者は一人もいない。もし世界の大半がそのことを理解すれば、ならず者国家が条約を破る理由をはるかに小さくなる。彼らもまた死にたくはないのである。

仮に、ASI の脅威を本当に理解していない指導者を持つ国家が存在したとしても、その国家が脅威を理解している国際同盟に囲まれているならば、懸念する諸国は介入し、その国家のインセンティブ構造を変更し得る。

たとえば、米国、中国、ロシア、ドイツ、日本、英国の指導者が、自らの生存が誰にも超知能を構築させないことに依存していると真に信じ、その試みを自らの生命と繁栄に対する脅威とみなし、自衛のために行動する用意があると明確に伝えるならば、たとえ異論を持つ指導者であっても、その連合に対して運試しをする気にはならないであろう。

AI 開発は軍事的優位を競うレースではない。それは自殺へのレースである。もし世界の指導者がこの点を理解し、自分自身と子どもたちがそれによって死ぬと予想するならば、彼らは誠実に条約を守り、誠実にその執行に協力するであろう。

全人類より賢い機械をつくることが世界を崖から突き落としかねないという議論は、実際にはそれほど難解ではない。人類が自ら構築している知的機械についてどれほど理解していないかは、真剣に立ち止まって問えば見えてくる。問題は、世界の指導者がこれらの事実を信じるに至るかどうかである。しかしもし彼らが信じるならば、この自殺レースを止めることは非現実的ではないと私たちは考える。

## 条約には実質的な監視と施行が必要である。

たとえ大多数の国家が「誰かがそれをつくればみんな死ぬ」と理解しても、理解しない国家が存在し、機械的超知能の構築を強行する可能性はある。

監視は必要である。施行も必要である。核兵器、生物兵器、化学兵器の条約は、遵守を検証する方法について一定の前例を提供している。条約回避の試みを困難かつ高コストにするべきである。

最先端 AI に対する国際的禁止は、厳格に施行されなければならない。もしある国家が

国際的圧力にもかかわらず推進を決意するならば、署名国による軍事力の行使が必要となる場合もあり得る。

これは理想的ではない。力が実際に用いられる事態を避けるためにも、そのような状況では力が用いられることを明確にしておく努力が必要である。しかし、人類の存続を救うことが限定的な軍事行動、あるいは相手国がエスカレートするならば戦争すら正当化し得る理由でないとすれば、何が正当化し得るであろうか。

この方法は前例がある。

原爆開発から 80 年以上が経過したが、人類は核拡散の管理について比較的うまくやってきた。第二次世界大戦後に多くの専門家が予測したにもかかわらず、大規模な核戦争は発生していない。

2025 年 6 月には、米国政府がイランの核兵器開発能力を妨害するための限定的攻撃を実施した。この種の条約および執行体制は、すでに世界秩序の中に前例がある。

もし ASI の開発までに 80 年の猶予を得ることができるならば、それは十分である可能性が高い。

## 監視体制は永遠に続けられるのか？

いいえ。いずれ別の出口が必要となる。

AI 研究の進歩は、おそらく完全に停止することはできない。十分な時間があれば、研究者たちはよりはるかに効率的な AI 構築手法を最終的に発見する可能性が高い。あるいは、十分な時間ののちに、あるならざる者の行為者が禁止体制を破ることに成功するかもしれない。

時間は、いずれにせよ人類を未来へと引きずっていく。そして人類は、過去の多くの種と同様に絶滅するか、あるいはより賢い存在が存在する世界への移行を何らかの形で乗り越えるかのいずれかである。

しかし人類は、永遠に時間を稼ぐ必要はない。AI だけが進歩している技術ではない。バイオテクノロジーも成熟しつつある。もし人類が数十年にわたって超知能機械の開発を阻止できたならば、やがては遺伝子工学によって著しく賢い人間が誕生する、といった変動にも直面することになる。

問題は、どれだけの時間を稼げるか、そしてその時間を何に使うかである。

人類が直面している基本問題は、人間の知能から超知能への溝を安全に越えることである。私たちが現実的に成功の可能性があるかもしれないと考える最善の計画は、バイオテ

テクノロジーによって人間の知能を大幅に増強するための時間を稼ぐことである。将来の人間研究者が、たとえば実際に予定通りかつ予算内で完了する見込みがある場合にしか工学プロジェクトの完了を予測しないほど賢くなるまで。

周囲の社会が完全に信じ込んでいたとしても、アリストテレス主義や地動説のような理論に安易に飛びつかないほど賢くなるまで。「前」と「後」のあいだの溝を、最初の試みで渡り切る可能性を持つほど賢くなるまで。

他にも想像し得る道はあるが、この道は重要なボトルネック（「既存の科学共同体は試行錯誤と漸進主義に過度に依存しており、この問題に対処できない」という点）を直接攻撃し、しかも今日すでに出現しつつある技術を用い、世界に重大な危険をもたらさずに済むという利点がある。

## 監視体制は永遠に続くべきではない。

理論上は、人類が現在のような綱渡りの能力水準に永遠にとどまることも可能である。私たちの推測では、それには人々の思想と行動に対する苛烈な統制が必要となるであろう。仮にそうでなかったとしても、それは望ましい選択ではないと私たちは考える。

私たちは個人的に、人類の子孫が自ら望むものへと成長し、星々を探検し、そこで繁栄し美しい文明を築くに値すると考えている。最先端 AI 開発の禁止を提唱するのは、超知能がそれほどまでに危険であると考えからであって、AI や技術や科学的進歩を憎んでいるからではない。

真の問題は、いかにして素晴らしい未来へ到達するか、そしてここからそこへの移行をいかに管理するかである。

この点を強調する価値があるのは、AI を誤った二者択一として提示する人々が多いからである。彼らは（誤って）社会は AI のリスクを受け入れて全速力で進むか、あるいは AI を拒絶して単一の惑星上で文明を停滞させるかのどちらかしかないと言う。しかしそれは誤りである。未来への道は他にもある。ほぼ同じほど明るい未来を許容しながら、すべてを無に帰す危険をはるかに低く抑える道が存在する。人類は別の未来への経路を見出すべきである。

## 人間をより賢くすることはなぜ役立つのか？

それはアライメント問題の解決に資する可能性がある。

AI アライメント問題は、私たちには本質的に解決不可能な問題であるようには見えない。ただ、人類はそれを解決する段階からはほど遠く、しかも「解決策を持っていると思

うこと」と「実際に解決策を持っていること」とが強く相関するような知性水準にも達していない、というだけである。

多くの AI 研究者は、アライメント問題がきわめて困難に見え、これまでほとんど進展していないことを認識している。だからこそ「AI にアライメントの宿題をやらせよう」という発想が魅力的に響くのである。自分たちが解けそうにない問題に直面したとき、AI 研究者にとって最も手近な道具は AI だからである。

しかし第 11 章および関連する〈より深い議論〉で述べたとおり、この発想には明白な問題がある。人間の最良の研究者たちが大いに苦勞している深い問題を AI が解くためには、その AI は危険なほど賢くなければならない。しかも私たちは自分たちが何をしているのかほとんど理解していないため、狭義のアライメント能力を直接訓練するための確固たる基準もなく、AI が提示したアライメント案が安全か有効かを検証する方法もない。

世界は、ときに本当に手の届かない問題を私たちに突きつける。自然は人類に「公正な」課題だけを与えるゲームではない。最良の科学者であっても解けない問題や、与えられた時間内には解けない問題に遭遇することはあり得る。

では、より現実的にこの問題をより賢い存在へ委ねる方法はあるか。一つの選択肢は、人間自身をより賢くすることである。人間は AI と異なり、ある意味で「事前にアラインされている」。最も賢い人間も、基本的には他の人々と同じ親社会的動機を持っている。

理論上は、「鉛を金に変えられるかもしれないと感じさせる錬金術の啓示」と、「原子物理学を用いて実際に鉛から中性子を取り除く知識」とを区別できるはずである。それらは明らかに異なる知識状態であるべきである。

しかし実際の歴史においては、人類は長らく錬金術に欺かれてきた。現実の科学者たちも、自らの愛着ある理論に固執し、現実は何度も打ちのめされるまで見解を改めないことがある。科学は「一葬式ずつ進歩する」と言われることすらある。だがそれは自然の必然ではなく、人間という種が十分に賢明で慎重で自己認識的ではないことの帰結にすぎない。通常であれば、こうした人間の素朴さは致命的ではない。錬金術師の思い上がりが全人類を滅ぼすことはなかったからである。しかし機械的超知能のアライメントに関しては、そのような余裕はない。

それでも、人類は必ずしも試行錯誤のみで進歩する必要はない。アインシュタインは人工衛星による時計のずれが観測される以前に、一般相対性理論を思索によって導いた。彼は経験的証拠を持っていたが、真理が扉を叩き壊すまで待つ必要はなかった。

このような道は稀であり困難であるが、実在する。

アインシュタインやフォン・ノイマンより一段階か二段階賢い人間が存在すれば、彼らは自らの認知的欠陥を見抜き、修正できるかもしれない。合理化や確証バイアスに気づき、巧妙に聞こえるが実際には誤った発想に惑わされなくなるかもしれない。成功を予期するときには実際に成功するほどの認識的能力に到達するかもしれない。

人間知能の増強は現実的か。私たちが話した複数のバイオテク研究者は、近い将来に有

望なアプローチがあると考えている。医療用途に限定されたAIが研究を加速する可能性もある。ただし、この計画が実際に成功するかどうかは依然として不確定である。それでも、高いレバレッジを持つ選択肢であり、現在よりはるかに多くの投資と探究に値すると私たちは考える。

これは唯一の戦略ではないが、現時点で最も有望に見える例の一つである。人類は、AI以外の複数の未来経路を探るべきである。

**強化された人間が深刻な「人間アライメント」問題を引き起こすことはない。**

強化された人間も、基本的には私たちと同じ脳の構造や感情などを持ち続ける。AI——たとえ人間らしく話すよう訓練されたAIであっても——とは異なり、認知や動機の面での巨大な隔たりや、それに伴う理解の困難さは、多少賢くなった人間にはそれほど生じないと考えられる。

認知能力を強化された研究者は、自らの精神的整合性を保ったまま、何百万倍もの規模の精神を持つ超知能へと変貌する必要はない。彼らが必要とするのは、「育つ」のではなく「つくる」ものとしての、真に整合的で安定した人工超知能を設計できるレベルに到達することだけである。

もちろん、複数人で協力する以上、弱い意味での「人間アライメント問題」——すなわちプリンシパル=エージェント問題やインセンティブの問題——は依然として存在しうる。そして、超知能の開発を担う集団においては、これらの問題の重要性はさらに高まる。

しかし、関わる人々がもともと利他的で誠実であり、知能の強化がゆっくりと段階的に行われ、適切に設計された制度とインセンティブのもとで活動するならば、これらの問題は対処可能だと考えられる。一方で、権力の集中や乗っ取りへの懸念が生じるのももつともである。これらの問題の解決は容易ではないだろうが、企業が理解不能な心と非人間的な動機を持つ超知能を育て上げようとする場合ほど、根本的に不可能というわけではない。

地球を安全に超知能の時代へ導くために、遺伝子操作による超天才チームをつくる——こうした構想は、重大な影響を持つがゆえに慎重に進める必要がある。実務的・倫理的問題も伴うが、他に有望な解決策がない場合、超知能によって人類が減ぼされるリスクと比較して検討されるべきである。

非常時には大胆な手段が必要になることもあるが、（穏やかな）人間の知能強化は、必ずしも極端な措置とは言えない。それ自体としても純粋に有益な技術であり、人類をさまざまな面で助ける可能性を持っていると考えられる。

もし協力に関心があるなら、ぜひ知らせてほしい。

人間の知能強化に関連するバイオテクノロジーの分野で専門知識を持っている場合は、このフォームから私たちに連絡してほしい。資金提供者、研究者、政策立案者のいずれであっても歓迎する。私たちは関連するバイオテクノロジーの専門家ではないが、この分野のさまざまな人物とのつながりを持っている。少なくとも、資金提供者・研究者・政策立案者を互いに結びつけ、連携を円滑にする手助けはできるかもしれない。

## 人間強化について意見が一致しなくても、超知能を止めるために協力することはできる。

人間強化のアイデアについて私たちと意見が一致しなくても、最先端 AI 開発を停止するという点では手を結べるはずだ。

もしこの問題を解決できなければ、全員が死ぬ。今すぐ死にたくないすべての人は、その目的のために協力しなければならない。人間の知能強化を違法にすべきか補助すべきかといった議論は、差し迫った死の脅威を乗り越えた後に行えばよい。

「人間をアインシュタインより賢くする」というのは、2028 年や 2032 年、あるいは次に AI アルゴリズムの大きなブレイクスルーが起きる時期までにどう生き延びるか、という計画にはならない。

それは AI 開発と並行して進められる計画でもない。仮に、人間未満の知能しか持たない AI によって解放された医療技術を使って、人間の知能をアインシュタイン以上に強化できたとしても、その強化された人間が AI アライメント問題を解決し、安全な超知能を設計・構築することを、軍拡競争の時間的プレッシャーの中で迅速に達成できるとは考えにくい。超知能への競争そのものを止める必要がある。

人間の知能強化というアイデアの趣旨は、多数の強化された研究者が何年も、あるいは何十年もかけて取り組めるなら、アライメント問題を解決できる可能性を開くかもしれない、という点にある。6 年後に、他の AI 企業よりも早く整合的な超知能をつくり上げる競争に勝てる、という話ではない。

「人間の知能強化」は有望な方策だと考える人々——私たちも含めて——の多くは、まず最初のステップとして AI 企業の停止が必要だと考えている。

そして、人類が次に何をすべきかについて別の計画を持つ人々であっても、最初の一步が AI 企業の停止であるべきだという点では、概ね一致している。

Meta AI も、OpenAI も、Anthropic も存在してはならない——それらは私たちを死に至らしめるからだ。この差し迫った優先事項については、その後に何をするかについて意見が大きく異なっている、合意できるはずだ。

## 誰にアライメントさせるのか？

これは厄介な問いである。しかし、その答えがいかなるものであれ、私たちは開発を停止しなければならない。

もし将来、人類が超知能を構築するならば、それは「人間の価値観」と整合していなければならない。だが、いったいどの人間の価値観か。人々は善悪、宗教、社会規範、政策上のトレードオフ等について甚だしく意見を異にしている。

現時点では、この問いは無意味である。人類は特定の目標をAIに確実に組み込む能力を持っていない。どの目標が理想かについて意見が割れていようと、問題にならない。これまで繰り返し論じてきたとおり、超知能の構築を急げば、私たちは全員死ぬ。人類は多くのことについて意見が分かれているが、地球上のすべての生命の破壊が良いことであるかどうかについては、ほとんどの人が賛成しない。

「誰にアライメントさせるのか」という問題は厄介である。率直に言えば、私たちはその問題を抱えられる立場にいたいとすら思う。しかし現実には直面しているのは、それよりはるかに悪い別の問題である。

国際的な禁止に協調するために、「誰にアライメントさせるのか」について合意する必要はない。あるいは、そもそも人類が超知能を構築すべきかどうかについて合意する必要もない。ただ単純に、そうしなければ死ぬという理由で十分である。AIに伴って生じる哲学的問いは無数にあるが、それらに過度に気を取られれば、子どもたちを殺すことになる。

実務的助言として、世界の指導者に対しては次のように言う。

「超知能を急いで構築すべきか」という問いと、「もし安全に構築できるならば、それをどう用いるべきか」という問いを分離せよ。そしてまず前者に集中せよ。前者こそが緊急かつ今行動可能な問いである。後者は将来重要になるかもしれないが、現時点では罣である。なぜならそれは、超知能を賞品であるかのように考えさせるからである。最初に構築した者がそれを支配できると誤信すれば、自殺レースに陥る。

ASIは魔法のランプの精霊ではなく、自殺ボタンである。誰かが超知能を創造しても、その人がそれを「所有」するのではない。創造された超知能が惑星を持つのである。

将来、どうしても「もし可能ならば超知能をどう使うべきか」という議論を行う必要があるならば、他者に競争を促すような提案は避けるべきである。勝者総取りの力学は、世界を危険にさらす。

「誰にアライメントさせるのか」という問題を、より普遍主義的に扱い、すべての関係者に公正であろうとする提案も存在する。たとえば「全人類のアライメント的外挿的意思」

にアライメントさせるという構想である。しかしそこでも、原理やトレードオフ、実装の詳細について際限なく議論が生じるであろう。それらは、人類が超知能を精密かつ頑健に制御できる段階に達したときに初めて意味を持つ議論である。今それを前面に出すことは、現実の選択肢を誤って描写し、地球の破壊を回避するという共有目標への協調を妨げる危険がある。

長期的に重要な問題であっても、人類の生存と一緒に束ねてよいのは、生存そのものだけである。

## 絶滅について語るのを避けるほうが賢明ではないか？

政治的な駆け引きをしている時間は過ぎ去った。

超知能構築競争を懸念する人々は、自らの見解を隠し、代わりにAIによる雇用喪失や、ChatGPTを利用したバイオテロ、あるいはデータセンター冷却に要する水量といった話題を前面に出すべきだと主張する者がいる。私たちは、このやり方は小賢しすぎて逆効果になりやすいと考える。実際、そのように裏目に出た例をすでに目にしている。

このアプローチには主に四つの問題がある。

第一に、それは誠実ではない。人々は不誠実さや駆け引きを見抜く能力を持っている。第二に、仮に非常に優れた嘘つきであったとしても、自分自身が副次的だと考えている問題についての議論は、どこか「ずれた」ものになりやすい。それらは、あなた自身が本質的問題ではないと感じているのと同じ理由で、どこかしっくりこない。洗練された議論を重ねるほど、人々はあなたを混乱しているか、あるいは本心を隠している人物だとみなす可能性が高まる。そのいずれの場合でも、協力相手や信頼できる情報源とは見なされにくくなる。

第三に、それはおそらく不要である。私たちの経験では、超知能について率直かつ直接的に語るほうが、AIディープフェイクなどの話題へ話題を逸らすよりもはるかに良い反応を得られる。2023年半ば以降、私（ソアレス）は多くの選出公職者と話してきた。AI支援テロリズムを懸念として挙げる人々がいる場で、現職議員が「私が本当に恐れているのは、三年以内に誕生し得る再帰的自己改良型AIによる超知能であり、それが私たちを地図上から消し去ることだ」と述べたこともある。

米国議会の議員を含む人々は、この問題を真剣に受け止め、対処法を模索する姿勢を示している。本書刊行時点ではまだ本格的な国家的・国際的議論は行われていないため、この問題は実際以上にニッチで論争的に見えるかもしれない。しかし率直なワシントンでの対話は、しばしば勇気づけられるものであった。

第四に、他の問題に対応しても超知能問題は解決しない。AI企業は超知能の構築を急

いでいる。それが達成されれば、全員が死ぬ。この問題に適切な解決策は、ディープフェイクやAI支援バイオテロへの対策とは根本的に異なる。

もちろん重なりはゼロではない。重なる部分を強調することで支持を広げることは可能であろう。しかし、超知能という複雑な問題に対し、実際の問題を直視せずに十分な対応に偶然たどり着く可能性はきわめて低い。

時間は短いかもしれない。より身近で穏当な問題から徐々に議論を広げ、何年もかけて超知能に到達する余裕はない可能性が高い。迅速に動員しなければ、対応する機会そのものを失うかもしれない。

雇用喪失やバイオテロが現実の問題でないと言っているのではない。ただ、無謀な自殺レースが進行中であると社会が理解しなければ、それを止めることはできない。

政策の世界でChatGPTを利用したバイオテロなどの問題が持ち出されてきたのを何年も見てきたが、それが機械的超知能の創出を防ぐ実質的な成果につながったようには見えない。

私たちは政治の専門家ではない。しかし、慎重で「非警鐘的」な言説だけでこの問題を乗り越えられる段階は過ぎたように思える。

ある時点で、人間として差し迫る脅威について語らなければならない。政策は、無難に見えるメッセージではなく、現実に基づくべきである。

AI研究者が人間を凌駕するAIが一〜四年以内に現れ得ると言っている。私たちはそれが誤りであってほしいと心から願うが、専門知識をもってしても彼らが誤っていると断言できない。政策立案者も断言できない。人類は直面する課題に適切に対応していない。今警鐘を鳴らさなければ、いつ鳴らすのか。

そして実際、この戦略は成果を上げつつある。2025年夏に政治家たちが超知能について発言した内容を見れば明らかである。人工超知能の差し迫る危険について、本格的な議論を行う好機が到来しているのである。

## 選挙で選ばれた公職者はこれを現実の脅威として認識するだろうか？

すでに認識している者は増えつつある。

その例として、次のような発言がある。

「私たちはこれに非常に注意しなければならない。監視しなければならない。……人類を乗っ取ると言う者もいる。本当に強力なものだ、AIは」

——ドナルド・トランプ大統領

「本日の議論では、高度な AI システムに関する緊急の問題が取り上げられるだろう。…  
…人工汎用知能、あるいは超知能 AI の開発競争。それはあまりに強力で、私たちには  
“デジタルの神” のように見えるかもしれない。……長期的リスクや終末的シナリオを回  
避するための具体的政策解決策に時間を割きたい」

——チャック・シューマー上院議員

「人工知能は、原子爆弾の発明以来最大の技術的脅威である」

——セス・モールトン下院議員

「AGI は、10 年あるいは 20 年以内に人間と同等かそれ以上に賢くなる可能性がある。そ  
れはもはや遠い未来ではない。SF ではない。今ここにある。……人間並みに賢い AI は、  
欺き、操作し、事実を隠し、戦争において独自の判断を持つ能力を有する」

——リチャード・ブルーメンソール上院議員

「無規制の AI 産業が AGI を開発することは、私たち全員を行動へと駆り立てるほど恐ろ  
しいはずである」

——ジョン・ヒッケンルーパー上院議員

この本が印刷に回ってからの数か月間でさえ、世界はこの脅威の理解に向けて前進して  
いるように見える。

2025 年夏の発言例を挙げる。

「AGI は数年以内に現れるという専門家もいる。未来は偶然に委ねるには重要すぎる」

——ナンシー・メイス下院議員

「人工超知能は、現在直面する最大級の実存的脅威の一つである」

——ジル・トクダ下院議員

「スカイネットの開発や機械の反乱に賛成するつもりはない」

——マージョリー・テイラー＝グリーン下院議員

「産業界の指導者は、社会を根本的に変える可能性のある AGI や超知能について公に語  
っている」

——シンシア・ルミス上院議員

「貯蔵庫を建てるのではなく、安全な AI を構築すべきだ」

——ラジャ・クリシュナモーティ下院議員

「ある閾値を超えれば、制御不能な“脱出速度”に達する可能性がある」

——ケビン・キリー下院議員

「AI で最初になる競争は、制御を失う競争かもしれない」

——ドン・ペイヤー下院議員

「非常に知識ある人々が、人間はこの技術を制御できなくなるのではないかと深く懸念している」

——バーニー・サンダース上院議員

さらに注目すべき発言として、国連事務総長アントニオ・グテーレスは、生成 AI について「核戦争と同等の実存的脅威であるとの警鐘を真剣に受け止めねばならない」と述べている。

依然として多くの進展が必要である。しかし世界は気づき始めている。連邦および国際レベルで迅速な行動が必要であることを、いまこそ公職者に訴える好機である。

## 状況は絶望的なのか？

いいえ。

これは勝ち得る戦いである。そしてこの世界は、そのために戦う価値のある世界である。容易には見えないが、真に実行可能であるように見える。

もし世界を行動へと奮い立たせる努力に加わりたのであれば、私たちはあなたと肩を並べて戦うことを光栄に思うであろう。どのように貢献できるかについては、本書最終章を参照してほしい。

### 〈より深い議論〉

## 世界的な AI 開発を止めるには何が必要か？

私たちは国際法の専門家ではなく、これは非常に複雑な話題で、各分野の専門家による

大きな労力を要すると予想している。だが、できるだけ早く議論を動かし始めるために、私たちは技術ガバナンスのチームおよび外部アドバイザーと協力して、効果があり得る措置についてのスケッチや推測をいくつかまとめた。

私たちはこれらを、会話、議論、批判、反復を促すという精神で提示する。これらの第一稿のアイデアは、決して確信に満ちたものや権威あるものとして扱われるべきではない。

第一歩として、私たちが解こうとしている問題の制約と形を概観しよう。この話題だけで一冊の本になり得る。全体としての問題は、機械超知能の開発を何十年にもわたって防ぐことだった。そして、どこに臨界の閾値があるのか分からない以上、これは本質的に AI の研究開発そのものを全面的に止めることに等しい。

現在の AI の進歩は、より良いコンピュータチップをつくること、より多くのチップをより長い学習計算（トレーニング）により長く使うこと、そして AI アルゴリズムを改良すること、これらの組み合わせから生じている。私たちはこれらを順に扱い、人工超知能への進歩を止めるための対応するレバーを説明する。

## より多い、より高性能な AI チップの製造を防ぐこと

現代の AI の能力を高めるには、膨大な計算資源と電力の投資が必要だ。その結果、現代の国家主体が関連する施設をすべて特定して監視し、新たな施設の出現を防ぐことは、消費者向けハードウェアへの影響を最小限にしつつ可能に見える。

先端 AI チップを生産するサプライチェーンは、極めて集中している。サプライチェーンのいくつかの工程では、その役割を担える企業が世界に一社しかなく、しかもそれらの企業は概ね、伝統的に米国と同盟関係にある国々に所在している。

たとえば、AI チップを製造できる企業はごく少数で、主として台湾企業の TSMC だ。またハイエンドチップに使われる重要な機械の一つは、オランダ企業 ASML だけが製造している。これは極端紫外線（EUV）リソグラフィ装置で、スクールバスほどの大きさがあり、重さは 200 トン、価格は数億ドルに達する。

このサプライチェーンは、数十年にわたる革新と投資の結果であり、再現するのは非常に困難だと見込まれている。技術的に進んだ国であっても、10 年以上かかる可能性が高い。

最先端の AI チップはかなり特化しているため、それらを追跡・監視しても波及的な影響は小さい。2025 年半ば時点で最も一般的な AI チップの一つである NVIDIA の H100 は、1 枚あたり約 3 万ドルで、冷却と電力要件のためデータセンターで運用されるよう設計されている。これらのチップは AI の学習と推論に関わる数値演算に最適化されており、標準的なコンピュータ（消費者向け CPU）に比べて、AI ワークロードでは典型的に数十倍から数千倍高い性能を持つ。

AI チップのサプライチェーンの集中と複雑さは、先端 AI 開発の停止を、直感より容易

にしている。新たな AI チップの生産を止めるのは簡単だろう。生産工程がいかに複雑で相互に絡み合っているかを踏まえると、秘密のサプライチェーンがつかられないことを確認するために、少数の重要サプライヤーを比較的最小限に監視すれば足りるはずだ。

同じインフラの一部は AI チップと他の先端コンピュータチップ（携帯電話用チップなど）の製造に共用されているが、これらのチップには顕著な違いがある。先端 AI チップの生産を止めた場合でも、継続しているチップ生産が非 AI 特化のチップのみをつくっていることを監視し、保証することは可能だろう。

既存の特化 AI チップは、ChatGPT のような既存 AI の運用に保持して用いるなら監視できる。そうしたチップが新規の研究開発ではなく低能力の AI の運用にのみ使われていることを保証するのは課題だが、克服不能ではない。既存のチップの所在は追跡・監視できるし、それらが何に使われているかを検証するための潜在的な仕組みもいくつかある。この種の監視にはチップへの物理的アクセス（たとえば監査官がデータセンターで測定を行うこと）が必要になる。新しいチップがセキュリティを改善して製造され、検証と監視を念頭に設計されるなら、遠隔アクセスでも検証に十分になり得る。次の節で述べるように、（2025 年 8 月時点の AI アルゴリズム水準では）危険になり得るチップ集積の規模は非常に大きいので、国家主体がそのような施設をすべて検知し、定期的な査察の対象にするのは難しくないはずだ。

## より多い、より高性能な AI チップの使用を防ぐこと

次に、チップの生産からチップの使用へと注意を移そう。現在最大の AI データセンターには数十万枚の AI チップが収容されており、費用は数十億ドルに上る。今日の最も強力な AI の一つを訓練するには、これらのチップを数か月間ぶっ通しで使う必要がある。

これらのチップ 1 枚は、平均的な米国の家庭に近い電力を消費するため、数十万枚のチップを持つデータセンターの電力使用量は小都市に匹敵する。これらすべてのチップに電力を供給するには、大型の送電線のような専用の電力インフラが必要だ。こうしたデータセンターは、エネルギー集約的な大量のチップを連続稼働させ冷却するため、かなり大きな建物でもあり、特徴的な熱的シグネチャも持つ。

建物の内部には、数千枚のチップがサーバーラックに収められ、チップが過熱しないよう大規模な冷却インフラが備わっている。こうした建物に一步入れば、それがデータセンターであることは極めて明白だ。国際監視者が立ち入り検査に来て、しかもそのデータセンターのチップを調べて AI 特化チップであることが分かれば、その目的を隠し通すことはできない。

大規模データセンターと関連する電力インフラは巨大なので、周回衛星から特定できる。つまり、政府が現在の大規模データセンターを見つけたいと思うなら、自国境内であれ他国であれ、高い成功率でそれを行える可能性が高い。公に知られている情報は限られてい

るが、この介入だけでもハイエンド AI チップの大半を追跡できるかもしれない。

将来、国家が衛星での特定を困難にするためデータセンターを隠そうとする可能性はある。たとえば (NORAD を収容する) シャイアン山複合施設のように、山中にデータセンターを隠して上空から見えないようにする、といった試みだ。それでも、データセンター運用に必要なインフラを隠すのは難しいだろう。

検知に有利に働く最大の要因は、データセンターが非常に大きな電力を必要とすることだ。この電力は通常、送電線で供給され、送電線はほとんど常に地上に露出している。送電線を埋設することは可能だが、はるかに高コストで時間もかかり、さらに埋設工事そのものも隠しにくい。

最先端 AI の訓練に 10 万枚超のチップが必要であり続ける限り、国家主体が関連するデータセンターをすべて検知し監視することは十分可能に見える。

## アルゴリズム上の進歩を防ぐこと

より効率的な AI アルゴリズムは、AI を訓練するために必要な計算資源を減らし得るし、一定の計算資源量でより高能力の AI をつくれるようにし得るし、その両方をもたらし得る。

アルゴリズムの進歩は主として研究とエンジニアリングによって駆動されており、現時点では人間の技能と努力に依存している。AI アルゴリズムを改善するのに必要な技能は比較的希少で、それがこの分野のトップ研究者に高給が支払われる理由だ。

これらの技能は今日では希少だが、より多くの研究者が分野に入り、より多くの知識が公開されるにつれてそれがどう変わるかは不明だ。必要な技能を持つ人の数をどう数えるかにもよるが、真の数は数百人から数千人程度である可能性が高い (たとえばトップ AI 企業の研究者・エンジニアの人数に基づく)。保守的な推定ならもっと大きくなり得て、たとえば世界にはソフトウェアエンジニアが数千万人いる。

法的・社会的介入によって、アルゴリズム進歩を大幅に遅らせられる可能性が高い。ほとんどの人は法律を破りたくないし、とりわけ実際の罰則がある場合はそうだ。十分に高能力の AI がもたらす破壊的リスクを理由に、特定の AI 研究を公表したり特定の AI 実験を行ったりすることが違法になれば、先に論じたとおり、潜在的な AI 研究科学者のほとんどは思いとどまるだろう。政府は、輸出許可と政府承認なしにその種の研究を共有・公表することを違法にする輸出管理を実施できる。

社会的タブーも助けになる。前例として、1975 年のアシロマ会議 (遺伝子組換え) があり、過度のリスクをもたらすと考えられた特定の生物学実験の自主的な禁止につながった。理論上、科学者は AI 能力を進めることの自主的禁止を設けられる。しかしそれには、より人間を上回る AI の危険を科学者たちが真剣に受け止める必要があり、これは現状 (多くの場で AI 能力の前進が称賛される状況) からの逸脱になる。近視眼的な金銭的イ

ンセンティブと、これまで観察されたラボの行動を踏まえると、分野の文化が劇的に（しかも短期間で）変わらない限り、外部からの法的制約が極めて必要になる可能性が高い。

不完全な禁止措置を実効的にするための重要な要素の一つは、これまで述べたとおり、世界の指導者が、前進を続ければ自分と家族が個人的に死ぬのだと本当に理解すること、という分かりやすいものかもしれない。禁止措置が不履行となる最もありそうな不遵守シナリオは、政府が自国産の超知能を戦略的資産（あるいは収益を見込める新しい AI ツールから注意をそらす蟹気楼）として促え、世界的な自殺ボタンとして認識しない場合だ。政府が、これが銃に弾を込め、こめかみに当て、引き金を引くことに等しいのだと正しく理解すれば、秘密裏の ASI 研究計画を走らせる可能性は大きく下がる。

研究禁止は全員を止めはしない。著名な研究者やテック幹部の中には、進歩の代償として人類を滅ぼすことを受け入れ得るとすでに述べた者もいる。だが、完全さを善の敵にしてはならない。そのような人々が資金を断たれ、同業者から排斥されれば、致死的研究を違法かつ非協力的な環境で行わざるを得なくなり、アルゴリズムの前進は少なくとも遅くなるだろう。

## 待てば待つほど難しくなる

AI チップの生産と流通が現在の軌道を進み続けるなら、十分な数の AI チップを集中監視することは、ますます難しくなる。国家がまだリスクに確信を持っていないとしても、今日から国際的に AI チップを追跡し始めれば、将来的に介入の余地を残すことができる。その扉は、政府が迅速に動かなければ、すぐに閉じるかもしれない。

研究者が AI アルゴリズムの最先端を進め続ければ、より少ないチップ数で深刻な脅威になり得ようになる可能性が高い。AI システムが AI の研究開発プロセスの一部を自動化できるようになれば、AI 開発を制御することはとりわけ難しくなり得る。そのようなシステムは容易に複製され配布され得るし、それを動かすのに必要なハードウェアは大した量ではないかもしれない（AI システムを動かすためのハードウェア要件は、AI システムを訓練する要件よりもずっと低い）。

最終的には、世界の政府が超知能 AI システムの開発を止めることは不可能になるかもしれない。私たちはそこには至っていないが、毎月難しくなっていく。私たちが概説する計画は、AI 開発を早期に止めることを前提としている。この前提に依らない別の計画もあるが、それらは実装がより難しく、個人の自由へのコストがより高く、失敗の確率もより大きい。

## 注釈付きの条約草案（暫定版）

米国下院や上院の議員を含む多くの人々が、この問題に実際に役立つ具体的かつ明確な立法提案は何か、と私たちに尋ねてきた。

私たちは政策の専門家ではない。この問いには多くの可能な答えがあると考えている。今日の時点で通しやすいが直接的な効果は小さい立法（より野心的な立法への踏み石として機能するもの）を求めているのか、それとも可決はより難しそうだが中核的問題に実質的に対処する提案を求めているのかによって、答えは変わる。

前者よりも後者について、私たちははるかに多くを語る。そのために、MIRI のテクニカル・ガバナンス・チームは、そのような提案の一つの予備的草案をまとめた (<https://ifanyonebuildsit.com/treaty>)。

この条約のスケッチは、世界の指導者たちが人間を超える AI の現実に見覚めた世界を想定して設計されている。これは（本稿執筆時点で）明日可決されると予測するものではない。しかし、世界が危険により自覚的になれば——そのプロセスはすでに始まっており、今後も続くことを願っているが——そのような条約を批准するのはそれほど難しくないかもしれない。

条約案は、多数の注釈付きで提示されている。起草過程では、他の条約の歴史的前例を参照しており、それらは草案と並べて示されている。私たちは改めて強調するが、これは出発点であって終着点ではない。私たちは政策の専門家ではなく、愚かな誤りを犯している可能性もある。それでもなお、この草案が、そのような条約が実際に可能であり、過去の類似した立法努力と整合的であることを示す着想や例として機能することを願っている。

## 連合を大きく保つ

私たちは、一部の人々から、より単純なメッセージを送るために——反超知能ではなく、反 AI という立場を打ち出すために——AI アートやロボット兵器に対して強く反対すべきだ、という主張を聞いてきた。

AI アートやディープフェイクなどをめぐる各種の立場の是非はひとまず脇に置くとして、私たちはそれが政治的に最善の選択肢だとは思わない。私たちは超知能を禁止するための連合を築きたい。この問題は極めて緊急かつ差し迫ったものだと考えており、その連合はできる限り大きいものであってほしい。そこには、AI アート、ドローン戦争、自動運転車、学校での AI 利用などについて幅広い見解を持つ人々が含まれていてよい。

他の問題についてどの立場に立つにせよ、暴走する超知能の創造を防ぐという点では、私たちには共通の利益がある。

人類は絶滅し、何か陰鬱なものに置き換えられるべきか。答えが「ノー」だと考えるす

べての人は、超知能への暴走的競争を止める緊急の取り組みに協力できる。AI アートやドローン戦争について意見が違う相手とは協力できない、と言えば、その連合は存続しないだろう。

また、超知能の問題に他のさまざまな論点を抱き合わせにして、超知能について協力する前に半ば関連する多数の問いにすべて同意しなければならない、という形にしても、連合は存続しない。

他の AI 関連の問題を重視しているなら、それらに取り組むことを私たちは勧める。しかし、それらを超知能と抱き合わせにしないほしい。もし私たちがこの状況を乗り越えるのなら、人類の生存と抱き合わせにされるべきものは何もない。

あなたがいま住んでいる国に住んでいて、第三次世界大戦の後に残された放射性の瓦礫の山に住んでいない理由の一つは、東西両陣営が数十年前に核戦争を現実的で深刻な人類への脅威だと認識し、同意したことにある。東は西を、そして西は東を、それぞれ人類に対するもう一つの重大な脅威だとみなしていた。それでも彼らは、核による全滅という脅威と、イデオロギー的敗北という脅威を、性質の異なるものとして賢明に扱った。

西側の立場から見れば、たとえ東側からの脅威が残っていたとしても、核戦争の脅威が減る方が望ましかった。そのためには、ワシントンとモスクワの間に直通回線を敷き、不拡散条約などの取り決めで協力するために、東側と一定期間協力することが必要だった。

非常に多くの国が協調する必要がある。非常に多くの派閥が（内部ですら）分断されている。すべての点で意見が一致する人々だけが一緒に行動する、という条件では、破局を回避することはできない。

私たちは、世界の他の問題を懸念している人々と、喜んで無条件に共通の立場を取る。政治的に意見が異なる人々とも、ためらわずに協力する。私たちはこの切迫したメッセージを世界に向けて発した。それを信じているからであり、この問題は国際レベルで直ちに対処されるべきだと考えているからだ。

あなたが誰であれ——ここであれ他の場所であれ、何のために聞っているにせよ——人間を超える AI の猛烈な開発に終止符を打ちたいと望むなら、私たちは同じ戦いを聞っている。