

序章 困難な予測と容易な予測

これはエリザー・ユドコウスキー&ネイト・ソアレス『超知能 AI をつくれば人類は絶滅する』のオンライン補足資料である。よくある質問に答え、本筋から外れるために詳しく扱えなかった背景や詳細について、より踏み込んだ議論を行うものだ。

非常に多くの内容を含んでおり、最初から最後まで通読することを前提とはしていない。本書は単独で成立しており、中心的な主張全体を網羅するように書かれている。しかし、疑問や反論、あるいは好奇心がある場合には、ここでより詳しい説明が見つかる可能性が高い。もし重要な点が抜けている場合は、質問を投稿してほしい。

なぜ超知能 AI を絶滅の脅威として扱う本を書くのか？

状況が本当に深刻で、しかも差し迫っているように思われるからだ。

あるテーマを注意深く検討すれば、ときに歴史の大きな転換点のひとつが近づいていることを見抜けることがある。

1933年、レオ・シラードという物理学者が、核連鎖反応が可能であることを初めて見抜いた。彼はそれによって、歴史の大転換のひとつを、誰よりも早く予測できる立場を得たのである。

私たちは、今日の AI を正しい視点から見れば、歴史の次なる大転換が近づいているのかわかると考えている。そして、人類が進路を変えなければ、事態は悪い方向へ進むだろうとも考えている。

AI 研究機関は、あらゆる人間よりも賢い機械をつくろうと競争しており、その最前線を大きく前進させているように見える。後の章で述べるように、現代の AI は「つくられた」というよりも「育てられた」存在に近いものである。誰も求めず、誰も望まなかった振る舞いを示し、しかもあらゆる人間を上回る能力を持つ方向へ進みつつある。私たちに、これは極めて危険な状況に見える。

この分野の第一線の科学者たちが集まり、AI の脅威は「パンデミックや核戦争などの社会的規模のリスクと並ぶ、世界的な優先課題であるべきだ」と警告する公開書簡に署名した。これは一部の孤立した懸念ではなく、分野全体のおよそ半数が共有している問題意識である。たとえ最初は危険性に懐疑的であったとしても、AI の専門家たちが公に示している懸念の強さと、それが正しかった場合の甚大な結果を考えれば、この問題を真剣に議論する価値があることは明らかだと、私たちは考える。

このテーマについては、第一印象に従うのではなく、議論を慎重に比較検討すべきであ

る。もし書簡や警告が正しいなら、世界はすでに信じがたいほど危険な状況に足を踏み入れていることになる。本書の残りの部分では、それらの警告を支持する議論と証拠を提示していく。

私たちは状況が絶望的だとは考えていない。この本を書いたのは、人類が今向かいつつある軌道を変えられるという希望があるからである。私たちは、この問題を解決できる可能性があると考えている。

問題を解決するための第一歩は、それを理解することである。

ChatGPT のようなものが人類を滅ぼすと言っているのか？

いいえ。懸念しているのは、これから登場する AI のさらなる進歩である。

あなたが今この本を読んでいる理由の一つは、ChatGPT などの進展によって AI がニュースの中心に躍り出たからであろう。世界はいま、AI の進歩や、AI が社会に与える影響について議論し始めている。これは、人間より賢い AI について、そして現在の状況が決して楽観的とは言えないことについて議論する自然な機会を生み出している。

私たち著者は、この分野に長年取り組んできた。最近の AI の進歩は私たちの見解に影響を与えているが、私たちの懸念は ChatGPT やそれ以前の大規模言語モデルによって突然生じたものではない。私たちは 10 年以上にわたり、人間より賢い AI がうまく機能するようにするための技術研究に取り組んできた（ソアレスは 2013 年から、ユドコウスキーは 2001 年から）。しかし最近になって、世界がこの議論を受け入れる準備が整いつつある兆しが見えてきた。そしてこれは、おそらく今こそ行わなければならない議論である。さもなければ、世界はこの問題に対応するチャンスを失ってしまうかもしれない。

AI の分野は進歩し続けており、いずれ（いつかは分からないが）私たち人間よりも賢い AI をつくれる段階に到達するであろう。それは主要な AI 企業すべてが明言している目標である。

「私たちは、従来の意味で理解されてきた AGI（汎用人工知能）を構築する方法を、いまや理解していると確信している。（中略）そしてその先、本当の意味での超知能へと目標を向け始めている。現在の製品も気に入っているが、私たちは輝かしい未来のためにここにいる。超知能があれば、他のすべてのことが可能になる」

——サム・アルトマン（OpenAI CEO）

「（強力な AI は）早ければ 2026 年にも実現する可能性があると思う。（中略）私が言う

強力な AI とは、純粋な知的能力において、生物学、プログラミング、数学、工学、執筆などの主要分野の大半でノーベル賞受賞者より賢い AI モデルのことである。未解決の数学定理を証明し、極めて優れた小説を書き、難解なコードベースを一から構築できるような存在である」

——ダリオ・アモデイ（Anthropic CEO）

「私たちは全面的な汎用知能の構築に注力している。本日話したあらゆる機会は、その実現と効率的な達成にかかっている」

——マーク・ザッカーバーグ（Meta CEO、
「超知能」プロジェクトへの 143 億ドル投資発表直前の発言）

「今後 5 ～ 10 年の間に、AGI と呼べるものが実現する確率はおよそ 50% あると思う」

——デミス・ハサビス（Google DeepMind CEO）
インタビュアー「デミス、あなたは知能爆発を引き起こそうとしているのか？」
ハサビス「いや、制御不能な爆発を起こすつもりはないが……」

彼らは言葉だけでなく資金も投じている。Microsoft、Amazon、Google はいずれも 2025 年に AI データセンターへ 750 億 ～ 1000 億ドルを投資する計画を発表した。スタートアップの「xAI」は、X.com を約 330 億ドルの評価額で買収し、その後さらに 100 億ドルを調達して大規模データセンター建設と AI「Grok」の開発を進めている。OpenAI は Microsoft などと提携し、5000 億ドル規模の「Project Stargate」を発表した。

Meta のザッカーバーグ CEO は、今年 AI インフラに 650 億ドルを投じ、今後数年で「数千億ドル」規模を AI プロジェクトに費やすと述べている。Meta はすでに Scale AI に 143 億ドルを投資し、その CEO を Meta Superintelligence Labs の責任者として迎え入れ、競合研究所から十数名のトップ研究者を 1 人あたり最大 2 億ドルという条件で引き抜いている。

これらすべてが、人間より賢い AI がすぐそこにあることを意味するわけではない。しかし、主要企業がそれをつくろうとして全力を尽くしていること、そして ChatGPT のような AI がその研究プログラムの産物であることは意味している。彼らの目的はチャットボットをつくることではない。目標は超知能であり、チャットボットはその道中の中継地点にすぎないのである。

私たちは長年この問題について理解を深め、将来の展開を真剣に考えてきた。その結果、研究者が明日にも突破口を開き、人間より賢い AI の構築に成功することを原理的に妨げる決定的な障壁は存在しない、という見解に至っている。

その閾値に近い将来に到達するのか、それともまだ 10 年先なのかは分からない。歴史が示すように、新技術の「いつ」を正確に予測するのは、その技術がいずれ実現することを予測するよりもはるかに難しい。しかし私たちは、今日、強力な国際的対応を正当化する

には十分すぎるほどの証拠があると考えている。その議論は、本書で詳しく展開されている。

人はいつもパニックに陥って過剰反応するものでは？

そのとおり。しかし、だからといって「実際には危険はまったくない」ということにはならない。

人は時に問題に過剰反応することがある。運命論的に考える人もいる。社会的なパニックの中には根拠のないものもある。しかし、だからといって私たちが完全に安全な世界に生きているわけではない。

1935年のドイツは、ユダヤ人やロマ、その他さまざまな人々にとって良い場所ではなかった。警告サインに気づいて去った人もいた。一方で、それらの警告を大げさだと片付け、その結果、命を落とした人もいた。

核による壊滅の脅威は現実のものだったが、人類はそれに対処し、冷戦が熱戦に発展することはなかった。

クロロフルオロカーボン（CFC）は実際にオゾン層に穴を開けていたが、国際条約によって禁止されたことで問題は解決された。その後、オゾン層は回復した。

人々が警告する危険の中には、偽物もあれば、本物もある。

人類は常に過剰反応するわけでもなければ、常に過小評価するわけでもない。場合によっては、その両方を同時にしてしまうことさえある。例えば、次の戦争に備えて巨大な戦艦を建造した国々があったが、実際には航空母艦を建造すべきだった、というように。

「想定される技術的リスクはすべて無視すればいい」とか、「技術的リスクはすべて本物だと考えればいい」といった単純な解決策は存在しない。何が真実かを見極めるには、それぞれのケースの詳細を実際に検討する必要がある。

（このテーマについてさらに知りたい場合は、本書の序章を参照のこと。）

そういう「懸念されるタイプのAI」は、いつ開発されることになるのか？

人々が私たちに予測してほしいと求める多くの事柄について、実のところ私たちに知る方法がない。1939年にレオ・シラードが核兵器についてアメリカに警告する書簡を書いたとき、彼は「最初の原子爆弾は6年後に実験爆発が可能になる」といった具体的な日付を記すことはしなかったし、実際できなかったのである。

それは極めて有益な情報であったであろう。しかし、たとえシラードのように核連鎖反

応を最初に正しく予見した人物であっても——ある技術が可能であり重大な影響を持つと最初に見抜いた人物であっても——その技術が正確にいつ到来するかまでは予測できない。容易な予測と困難な予測がある。私たちは、危険な種類の AI が正確にいつ生み出されるかといった難しい判断ができるふりをするつもりはない。

専門家たちは、AI の進歩の速さに何度も驚かされ続けている。

AI がいつ到来するか分からないことは、それがまだ遠い未来だと分かっていることと同じではない。

2021 年、予測サイト Metaculus のコミュニティは、最初の「真に汎用的な AI」は 2049 年に到来すると見積もっていた。ところが 1 年後の 2022 年には、その集合予測は 12 年短縮されて 2037 年に、翌 2023 年にはさらに 4 年短縮されて 2033 年になった。AI の進歩の速さに予測者たちは繰り返し驚かされ、年ごとに見積もりは大きく揺れ動いている。

この現象は Metaculus だけに限らない。「80,000 Hours」という組織も、多くの専門家グループで予測タイムラインが急速に短縮している事例を記録している。さらに、予測大会で常に上位に入り、しばしば分野の専門家を上回る精度を示す「スーパーフォーキャスター」でさえ、2025 年までに AI が国際数学オリンピックで金メダルを獲得する確率をわずか 2.3% と見積もっていた。ところが AI は 2025 年 7 月に実際に金メダル相当の成績を達成した。

人間より賢い AI が実現するのは直感的には何十年も先のことに思えるかもしれない。しかし、2021 年には ChatGPT レベルの AI でさえ何十年も先の話に見えていた。それが突然実現したのである。次の質的飛躍がいつ訪れるのか、誰にも分からない。さらに 10 年かかるかもしれないし、明日にでも突破口が開けるかもしれない。どれくらい時間がかかるのかは分からないが、多くの研究者が、残された時間は短いのではないかと次第に懸念を強めている。特別な知識があると主張するわけではないが、私たちは人類は早めに対応すべきだと考えている。これ以上の警告期間が与えられる保証はない。

AI の能力がほとんど前触れなく連鎖的に飛躍する可能性については第 1 章を参照のこと。

また、現代の AI パラダイムが「最後まで到達」できるのかどうかについては第 2 章を参照のこと。

メディアが「何が近いうちに起こり得るか、起こり得ないか」について断定的に語る主張には慎重であるべきである（すでに起きている可能性すらある）。

ウィルバー・ライトが動力飛行の実現には 1000 年かかると落胆気味に予測してから 2 年後、ニューヨークタイムズ紙は自信満々に「100 万年はかかる」と主張した。その 2 か

月と8日後、ライト兄弟は飛行に成功した。

今日でも懐疑論者の中には、AIが特定の能力で人間に匹敵することは絶対にないと言っている者がいる。しかし機械学習の最近の進歩は、AIが人間と同等か、あるいはそれを上回る成績を示すベンチマークの数が増え続けていることを示している。たとえば少なくとも2024年末までには、現代のAIがテキストや非言語の手がかりから皮肉や嫌味をしばしば識別できることが知られていた。それにもかかわらず、ニューヨークタイムズ紙は2025年5月、「現在の技術が、皮肉を理解するような脳の比較的単純な働きさえ遂行できるという確固たる証拠はない」との主張を繰り返した。

要するに、人間より賢いAIが差し迫っていると主張する人もいれば、計り知れないほど遠い未来の話だと主張する人もいる。しかし不都合な現実、現時点では誰にも分からないということである。

さらに悪いことに、国際社会が対処できる段階を過ぎてからでなければ誰にも分からない、という可能性も高い。

次の技術的ブレークスルーの時期を正確に予測することは極めて困難である。人間より賢いAIが致命的に危険であることは分かっている。しかし、それが到来する曜日まで正確に知らなければ行動できないというのであれば、私たちは何もできないことになる。不確実性の中から行動できなければ、結局まったく行動できないのである。

過去の進歩から、人間より賢いAIがいつつくられるか予測できるのでは？

そのための十分な「知能」に関する理解を、私たちはまだ持っていない。

うまく当たる予測の一種に、長年安定して続いているグラフ上の直線をそのまま延長し、少なくともあと1〜2年はその傾向が続くと見なすという方法がある。

これは常にうまくいくわけではない。トレンドラインはときに変化する。しかし実際には、かなりの頻度で妥当な予測を可能にする方法でもある。

この方法の大きな問題は、私たちが本当に知りたいのが「このグラフの線は2027年までにどこまで上がるか」ではなく、「この線が上がり続けたら、質的に何が起ころのか」という点にあることである。つまり、その線がどの高さに達したとき、現実世界で重要な結果が生じるのかということである。

そしてAIの場合、それが分かっていない。たとえば「パープレキシティ (perplexity)」のような、言語モデルの性能を測る統計的指標を選べば、きれいな直線を描くことができるし、それを延長することもできる。しかし、将来のどの「パープレキシティ」の水準が、どの程度の質的なチェス能力に対応するのかは、誰にも分からない。事前に予測すること

はできず、AI を実際に動かしてみるしかない。

「今や全人類を殺せる能力を持った」と言える境界線が、そのグラフのどこにあるのかも、誰にも分からない。できるのはAI を実行して確かめることだけである。したがって、グラフ上の直線を延長することは役に立たない（しかもその前に、アルゴリズムの進歩によってそのグラフ自体が無意味になる可能性もある）。

そのため本書では、誰かがAI の訓練に10 の27 乗回のFLOP（浮動小数点演算）を投入するのが正確にいつになるのか、あるいはそれがどのような結果をもたらすのかといった予測に時間を費やしていない。それは難しい判断である。本書が焦点を当てているのは、私たちにとって比較的「簡単な予測」に見えることだ。

扱っているのは狭い範囲の話題にすぎないし、その限られた領域でいくつか重要な予測ができるからといって、将来について何でもかんでも予言できるわけではない。

著者としての動機や利害関係は？

平均的に見て、この本から私たちが金銭的利益を得ることはないと考えている。加えて言えば、私たちは本書の主張が間違っていることを心から望んでいる。

私たち（ソアレスとユドコウスキー）は、機械知能研究所（MIRI）から給与を受け取っている。MIRI は、これらの問題が重要だと考える人々からの寄付によって資金提供を受けている。もしかすると、この本が寄付を増やすことはあるかもしれない。

とはいえ、私たちには他にも収入を得る機会があるし、お金のために本を書いているわけではない。本書の前払い金はすべて広報活動に充てられ、印税はすべてMIRI に渡され、スタッフの時間や労力に対する補填に使われる。

そももちろん、もし私たちの文明が危機に瀕していないと結論づけられるなら、著者である私たちは大いに喜ぶであろう。さっさと引退したり、別の場所でもより多くのお金を稼いだりできるなら、それに越したことはない。

仮に証拠が見解の変更を正当化するものであれば、私たちは考えを変えることに困難はないと思っている。実際、過去にもそうしたことがある。MIRI はもともと（当時は「シンギュラリティ研究所」という名称で）超知能を構築するプロジェクトとして設立された。しかしユドコウスキーは、それが自動的にうまくいくわけではないと気づくまでに1年を要し、さらに数年かけて、それをうまくいかせるのは相当に難しいと理解した。

私たちは一度ピボット（方針転換）をしているし、もう一度転換できるなら喜んでそうしたい。しかし、現時点では証拠がそれを支持しているとは思えない。

状況が絶望的だとは考えていない。だが私たちが見る限り、ここには本物の問題があり、

世界が本気で立ち向かわなければ脅威は極めて大きい。

また強調しておきたいのは、AI が私たちを滅ぼす方向へ進んでいるかどうかを判断するには、AI そのものについて考えなければならないということである。人間側の動機や立場だけを考えれば、どんな情報源でも退ける理由は見つかる。学者は現実離れしている、企業は話題づくりをしている、非営利団体が資金集めをしたいだけだ、趣味の人たちは何も分かっていない、などである。

しかしそのやり方では、誰を退けるかによって最終的な信念が決まってしまう、議論や証拠が間違いを正す余地を失う。何が真実かを知るには、誰が主張したかとは切り離して、その議論がそれ自体で成り立っているかどうかを検討するほかない。

本書は「AI 研究所を率いる企業幹部には AI が安全だと人々を安心させる動機がある」というような安易な議論から始めてはいない。まず AI そのものについて議論する。そして後半では、人間の科学者が過度に楽観的であった歴史を少し振り返るが、AI 研究所に所属しているという理由で誰かの議論を無視すべきだとは一度も言っていない。私たちは開発者たちの実際の計画を取り上げ、その計画がなぜそれ自体の論理でうまくいかなかったのかを検討している。重要なのは議論そのものである。

もし私たちが間違っていると思うなら、ぜひ議論に向き合い、どの部分が誤っていると考えるのか具体的に指摘してほしい。それが、人物の性格や動機に注目するよりも、何が真実かを見極めるためのより信頼できる方法だと私たちは考えている。

世界で最も偏った人間が「雨が降っている」と言ったとしても、それが晴れていることの証明にはならない。

この手の AI の話って、ただの SF でしょ？

ある話題がフィクションで多く描かれているからといって、そこから多くを学べるわけではない。

人間を超える知能をもつ AI はまだ実現していないが、フィクションでは数多く描かれてきた。しかし、そうした描写を基準点にするべきではない。現実の AI は、おそらくフィクションの AI とは大きく異なる。これは第 4 章で詳述している。

AI は、フィクションで予見された最初の技術ではない。空気より重い飛行機や月旅行も、実現以前に物語で描かれていた。核兵器の一般的な構想も、初期の SF 作家である H・G・ウェルズが 1914 年の小説『解放された世界』で描いている。ウェルズは、爆弾が数日間燃え続けると書き、実際の爆発と放射線被害の形態を正確には描かなかったが、「化学エネルギーではなく核エネルギーで動く爆弾」という基本コンセプトは捉えていた。

1939 年、アルベルト・アインシュタインとレオ・シラードは、ドイツに先んじて原子

爆弾を開発すべきだと訴える書簡をルーズベルト大統領に送った。ルーズベルトが核爆弾という概念をウェルズの小説でしか知らず、「SFにすぎない」と退けていた世界もあり得た。

実際には、ルーズベルトは少なくとも十分に真剣に受け止め、ウラン諮問委員会を設立した。この事例は、「フィクションで語られたから」という理由で概念を退ける危険を示している。

SFは「本当だと思い込む」ことで誤らせることもあれば、「どうせ嘘だ」と切り捨てることで誤らせることもある。作家は預言者ではないが、必ず外れる反預言者でもない。多くの場合、フィクションは脇に置き、技術やシナリオをそれ自体として分析するほうがよい。

現実を予測するには、議論を検討し、証拠を秤にかける以外にない。

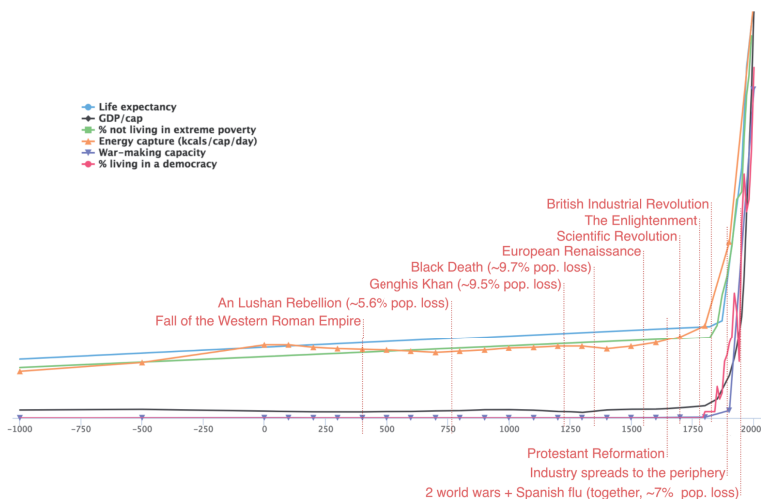
AIの帰結は、必然的に奇妙なものになる。

「AIは奇妙であり、世界を変えてしまい、現状を覆してしまう」という反応には、私たちも共感する。私たちは皆、ある程度、人間だけが発電所の建設のようなことを成し遂げられる世界に適応した直感を持っている。また、人類の歴史を通じて、機械は常に知性を持たない道具であったという世界にも適応している。確実に言えることの一つは、人間よりも賢いAIが存在する未来は、これまでとは異なるものになるということだ。

世界を大きく、しかも長期的に変える出来事は、日常的に起こるわけではない。「どうせ何も起きない」という経験則は多くの場合うまく機能するが、それが外れるときこそ、歴史上きわめて重要な瞬間である。そもそも未来について考える大きな理由の一つは、そうした重大な変化が起こる瞬間を予測し、備えることにある。

現状維持バイアスを乗り越える一つの方法は、序章で述べたように歴史を振り返ることである。

歴史を見れば、特定の発明が世界を一変させた例は存在する。たとえば蒸気機関や、それが産業革命の中で可能にした数多くの技術は、人間の生活を急速に変化させた。



真に汎用的な AI の登場は、同程度に重大な出来事である可能性が高い。少なくとも産業革命と同等の影響をもつだろう。具体的には以下のような影響だ。

- ・ AI は、技術の進歩をはるかに速く進展させる可能性が高い。第 1 章で述べるように、機械は人間の脳よりもはるかに高速で動作できる。そして人間は AI を改良でき、さらにやがては AI 自身が自らを改良できるようになる。その結果、機械は科学的発見や新技術の発明などにおいて、人間をはるかに上回るようになるだろう。

人類の歴史を通じて、人間の脳という「仕組み」は本質的には変わらないまま、人類はますます高度な工学的成果を生み出してきた。しかし、認知の仕組みそのものが向上し始め、それが自らを改良できるようになったとき、多くのことが非常に速いスピードで変化し始めると考えられる。

- ・ さらに、第 3 章で述べるように、十分に高度な AI は独自の目的を持つ可能性が高い。もし AI が単に人間より速く賢い存在にすぎないのだとしたら、それだけでも大きな出来事ではある。しかし実際には、AI は事実上、地球上のまったく新しい知的生命種のようなものになるだろう——それは独自の目的を持ち、そして（第 4 章および第 5 章で述べるように）その目的は人間の目的とは重要な点で食い違う可能性が高い。

この 2 つが同時に起きても世界秩序が維持されると考えるほうが不自然である。「普通の未来」を信じるには、機械知能が人間を超えないと信じる必要がある。それは 2015 年や 2005 年よりも、2025 年の現在のほうがはるかに信じにくくなっている。

長期的未来もまた奇妙なものになる。

21 世紀は 19 世紀から見れば異様であり、19 世紀は 17 世紀から見れば異様であった。AI はこの加速をさらに強め、新たな主体を盤上加える。

今日予測可能な点の一つは、高度な技術種は自分の惑星に永遠には留まらないということである。現在、夜空の星々はただエネルギーを放出しているが、理論的にはそれを利用する技術を構築できる。

物理的制約はあるが、最終的に恒星間探査や資源採取を行うことを妨げる原理的制限は見当たらない。自己複製探査機が宇宙に広がり、資源を収穫し、さらに拡張する未来は物理法則に反しない。もし人類が AI に置き換えれば、AI が同様の拡張を行うことを妨げるものもない。

地球上で生命が不毛な岩を覆い尽くしたように、やがて宇宙の無人領域も生命や機械により満たされる可能性がある。

現在、宇宙の大部分の物質は偶然に配置されている。しかし十分長期的には、物質の大半は何らかの主体の設計や目的に従って再配置される可能性が高い。

仮に地球発の知性が宇宙に進出しなくても、宇宙のどこかで一つでも宇宙進出知性が誕生すれば、それだけで拡張は始まり得る。

したがって未来は現在とは大きく異なる。細部を予測するのは困難である。しかし、宇宙の大部分が何らかの目的のために利用される方向へ向かうという予測は、直感に反していても比較的妥当である。

今から 100 万年後の未来が、2025 年のように、毛の少ない類人猿（＝人間）が地球の表面であれこれしているような姿のままであるとは考えるべきではない。それよりもずっと前に、私たちは自滅しているか、あるいは私たちの子孫が自ら宇宙へと進出しているだろう。

人類にとって奇妙な未来は避けがたい。問題は、それがいつ訪れるかである。

未来は速く到来する可能性がある。

AI のような技術によって、未来はすぐに私たちの目の前に現れ、その影響は強く私たちに及ぶかもしれない。

産業革命は前近代史の基準から見れば急速だった。ホモ・サピエンスによる世界改変も進化史の基準から見れば急速だった。新たな変化プロセスは、旧来の基準から見れば常に急速である。

人類は、機械が機械速度で世界を改変する段階の入り口に立っている可能性がある。機械知能が人間知能を超えれば、急速かつ根本的な変化が起こる可能性を真剣に考慮する必

要がある。そのような急激な転換は、歴史の中で繰り返し起きてきたのである。

〈より深い議論〉

破局シナリオについての AI 専門家

2022 年に、学術 AI 会議である NeurIPS および ICML の参加者 738 名を対象に行われた調査では、回答者の 48% が、AI の帰結が「極めて悪いもの（例：人類絶滅）」になる確率が少なくとも 10% あると考えていると答えた。この分野では、AI が前例のない大惨事を引き起こす可能性への懸念が広く存在している。

以下に、AI の破局的な帰結について述べた著名な AI 科学者・技術者のコメントを集める。これらの科学者の一部は、自らの「p(doom)」、すなわち AI が人類絶滅またはそれと同等の壊滅的な結果をもたらす確率を示している。

ジェフリー・ヒントン（2024 年）。深層学習革命を引き起こした功績によりノーベル賞およびチューリング賞を受賞。自身の推定について次のように述べている。「私は実存的脅威のリスクは 50% を超えると思っている」

ヨシュア・ベンジオ（2023 年）。ヒントンおよびヤン・ルカンとともにチューリング賞受賞。最も引用数の多い存命の科学者。「それが本当に危険になるまでにどれだけ時間があるのか分らない。ここ数週間、私は「心配しなくてよいという議論を示してほしい。私を説得してほしい。そうなればどれほど安心できることか」と言ってきた。しかし、まだそれは起きていない。……破局的な結果になる確率はおよそ 20% 程度だと考えている」

イリヤ・サツキバー（2023 年）。AlexNet の共同発明者、元 OpenAI 主任科学者、AI 分野で最も引用数の多い三名の一人。「超知能の莫大な力は非常に危険であり、人類の無力化や絶滅につながる可能性がある。超知能は今では遠いように見えるが、今後十年以内に到来する可能性がある。……現在、潜在的に超知能な AI を制御し、暴走を防ぐ解決策はない。人間のフィードバックによる強化学習などの現在のアライメント手法は、人間が AI を監督することに依存している。しかし、人間よりはるかに賢い AI を人間が確実に監督することはできない。現在のアライメント手法は超知能には拡張できない。新たな科学的・技術的突破が必要である」

ヤン・ライケ（2023 年）。Anthropic のアライメント科学共同責任者、元 OpenAI スーパーアライメントチーム共同責任者。インタビューが「個人的な p(doom) は 10% より大きく 90% より小さいと思う」と述べたのに対し、ライケは「私もおそらくその範囲だ」と答えている。

ポール・クリスティアーノ（2023 年）。米国 AI 安全研究所（NIST 内）安全責任者、RLHF（人間からのフィードバックを用いた強化学習）の発明者。「強力な AI（人間労働

を不要にするほど強力な AI) を構築してから 10 年以内に大多数の人類が死亡する確率：20%。……強力な AI 構築から 10 年以内に人類が不可逆的に未来を台無しにする確率：46%」

スチュアート・ラッセル (2025 年)。UC バークレー校スミス＝ザデー工学講座教授、主要な AI 教科書の共著者。「企業間・国家間の AGI 競争は冷戦期の核兵器競争に似ているが、より悪い。競争に参加している CEO たち自身が、勝者が人類絶滅を引き起こす有意な確率があると述べている。なぜなら、私たちは自分たちより賢いシステムをどう制御するか分からないからである。言い換えれば、AGI 競争は崖へ向かう競争である」

ヴィクトリア・クラコヴナ (2023 年)。Google DeepMind 研究科学者、Future of Life Institute 共同創設者。「2100 年までに AI が原因で自分が死ぬ確率は？」という問いに対し、「2100 年は非常に遠いが、現在の技術進展の速さを考えると、直感的には 20% 程度」と答えている。

シェーン・レグ (2011 年)。Google DeepMind 共同創設者、主任 AGI 科学者。「人間レベル AI の出現から 1 年以内に、人類絶滅のような極めて否定的な結果が生じる確率は？」との問いに対し、「5% かもしれないし 50% かもしれない。誰も良い推定値を持っていない。もし超知能機械が私たちを排除すると決めたなら、かなり効率的に行うだろう」と述べている。

エマド・モスタク (2024 年)。Stability AI 創設者。「私の p(doom) は 50% である。時間を特定しなければ、人間より有能で、重要インフラを運営するようなシステムが私たちを全滅させる確率は五分五分である」

ダニエル・ココタイロ (2023 年)。AI ガバナンス専門家、OpenAI 内部告発者。「AI 破滅は 70% の確率で起きると考える。20% 未満と考える人は極めて不合理だと思う」

ダン・ヘンドリックス (2023 年)。機械学習研究者、Center for AI Safety 所長。「私の p(doom) は 80% 超である。過去にはもっと低かった。2 年前は約 20% だった」

上記すべての研究者は、本書冒頭で紹介した AI リスク声明に署名している。その声明は次のとおりである。

「AI による人類絶滅リスクの軽減は、パンデミックや核戦争などの社会的規模のリスクと並ぶ、世界的な優先課題であるべきだ」

署名者には他にも、ChatGPT 設計者ジョン・シュルマン、元 Google 研究ディレクターのピーター・ノーヴィグ、Microsoft 最高科学責任者エリック・ホルヴィッツ、AlphaGo 研究責任者デイヴィッド・シルバー、AutoML の先駆者フランク・フッター、強化学習の先駆者アンドリュース・パート、GAN 発明者イアン・グッドフェロー、元バイドゥ(百度)社長の張亞勤、公開鍵暗号発明者マーティン・ヘルマン、Vision Transformer 研究責任者アレクセイ・ドソヴィツキらが含まれる。さらにドーン・ソング、ヤシャ・ソール＝ディクスタイン、デイヴィッド・マカレストー、クリス・オラ、ビーン・キム、フィリップ・トールほか数百名が名を連ねている。

レオ・シラードが未来を見たとき

1933年9月、物理学者レオ・シラードは、ロンドンのサウサンプトン・ロウとラッセル・スクエアの交差点を横断しているとき、核連鎖反応という着想を得た。これは原子爆弾の核心となる概念である。

そこから、シラードはこの重大な着想をどう扱うべきか模索することになる。彼はより著名な物理学者イシドール・ラビのもとを訪ね、ラビはさらに著名なエンリコ・フェルミに相談した。ラビがフェルミに核連鎖反応は現実的かと尋ねると、フェルミはこう返答した。

「ばかげている！」（Nuts!）

ラビがその意味を尋ねると、フェルミは「可能性は低い」と答えた。

「どのくらい低いのか」と問われると、フェルミは「10%だ」と言った。

それに対しラビは、「10%で死ぬかもしれないのなら、それは低い可能性とは言えない」と返した。

フェルミは再考した。

この逸話から引き出せる教訓はいくつかある。私たちが引き出さない教訓は、「死ぬ可能性があるならどんな低確率でも心配すべきだ」というものである。10%は「低確率」とは言えないが、十分に小さい確率であれば考慮に値しないこともある。

私たちが引き出す教訓の一つは、放射能の連鎖のような技術が原理的に可能であると気づいた瞬間に、世界が大きく変わると予見できる場合があるということである。

もう一つの教訓は、急激な変化を予測するうえで初期直感は信頼できないことが多いという点である。たとえエンリコ・フェルミのような著名な専門家であっても例外ではない。そもそもフェルミは「低い可能性」や「10%」という数字をどこから得たのか。

なぜ彼は放射能が放射能を誘発し連鎖反応を起こすとは考えなかったのか。それは単に、大きなアイデアの多くが実現しないという経験則によるものだったのか。

「ばかげている」という反応は、それ以上の意味を含んでいるように見える。この特定のアイデアは特に実現しえないという直感である。しかし、その根拠となる物理的議論は何だったのか。

単に「突飛に感じた」だけだったのか。核兵器の可能性が世界に重大な影響を及ぼすことは確かである。しかし現実には、「重大な結果をもつ出来事が起きないように」配置されているわけではない。

フェルミは当初、シラードにこの着想を公表するよう勧めた。つまり、ドイツや新首相アドルフ・ヒトラーにも知らせることになる。

その提案は退けられ、それは幸いであった。核兵器は実際に可能であることが後に判明

したからである。フェルミは最終的にシラードの小さな秘密の取り組みに加わったが、自らが最初の原子炉「シカゴ・パイル 1 号」を完成させる直前まで懐疑的であり続けた。

技術が世界を覆すことはある。急進的な新技術を「ばかげている」と前提してしまえば、世界屈指の科学者であっても進歩に不意を突かれる可能性がある。フェルミがシラードと議論を重ねたことは高く評価されるべきである。そして、技術が現実となる前、目で見えて確認できる前に、行動を改めるよう説得されたことはさらに大きな功績である。

人類史には数多くの悲惨な出来事があった。しかし、起きなかった悲惨な出来事の中には、誰かが腰を据えて議論し、時には強引にでも議論を起こしたからこそ回避されたものもある。シラードがフェルミに対して行ったように。