

第2部 ある絶滅シナリオ

第2部で私たちが描いたシナリオは予測ではない。未来の進み方には他にも多くの可能性があり、『超知能 AI をつくれば人類は絶滅する』がより長い本であったなら、複数の異なるシナリオを検討していただろう。以下では、なぜそのような形でシナリオを書いたのかという理由や、この種のシナリオを描く際に生じるさまざまな問題について述べる。

物語は、乾いた理屈にはない説得力を持ちうる。未来がどう進むかを具体的に想像してみることに価値があると私たちは考えている。しかし同時に、特定の物語に固執しすぎないことも重要だと思っている。シナリオを書く際のひとつ一つの選択は、それ単体ではもっともらしく見えるかもしれないが、いくつかの選択が重なるだけで、特定の一本道の全体的な確率は非常に低くなる。未来に多くの難しい分岐があるとは、そういうことだ。

もっとも、経路よりも結末のほうが予測しやすい場合も多い。なぜなら、多くの異なる道が同じ目的地へと至るからである。本書のシナリオでは、可能な限り複数の選択肢を提示し、物語がどのように進んだとしても、良い結末には至らないことを示そうとした。

なぜこの設定を選んだのか？

もっともらしく、書きやすいからである。

未来についての物語のあらゆる細部は、その物語が間違っている可能性を生む要素でもある。どの技術的ブレイクスルーがどの順番で起こるかを正確に言い当てることは、1 か月後の天気の詳細を言い当てるのと同じくらい不可能だ。

この種の物語は、未来を正確に映し出す鏡であることを意図していない。本書第1部で展開した抽象的な議論を一つに結びつけ、「未来がどう進み得るか」を具体的に示すためのものである。ある特定の破滅へ至る経路を鮮明に想像するとき、危険はより現実味を帯びて感じられると考える人もいる。

さらに説得力があるのは、10 本、あるいは 100 本の物語かもしれない。そこでは、いかに多くの道が破滅へ通じているか、そして繁栄へ至る道がいかに狭く脆いかが示されるだろう。

これが「未来のある側面が予測しやすい」という意味である。ほとんどすべての経路が同じ終点へ至るなら、その終点は予測可能だ。しかし私たちには 10 本の物語を書く時間も紙幅もなかった。ましてや 100 本などなおさらである。

私たちが選んだ物語は、できるだけ現在に近い時点から始まる。これは、近い将来に必ずこのような状況が起こると確信しているからではない（私たちは不確実である）。単に、

現在に近い設定の方が書きやすいからだ。もっと遠い未来を舞台にし、今とその時の間に起きた数多くの未来的出来事をでっちあげれば、物語はさらに不自然になるし、細部は気を散らすだけになる。

たとえ未来の正確な進路を予見できたとしても、それが一般的な力学を理解するうえで最適なシナリオであるとは限らない。

現実の未来は、きわめて奇妙で、偶発的で、雑多な細部に満ちたものになると私たちは予想している。それらの細部は物語に書き込めば信じがたく感じられるだろう。そうした形で書かれた物語は、筋の通らない現実のせいで、説明不足で不要な細部に満ち、混乱を招くものになるだろう。また、多くの細部が奇妙に見えるため、かえって現実味を欠くように感じられるだろう。

その感覚を味わうために、100 年前に戻り、現代世界の日常生活や重大問題を説明しようとすることを想像してみてほしい。1925 年の多くの人々はラジオを聴いたことも、自動車を運転したことも、冷蔵庫を見たこともなかった。ソーシャルメディア、グローバル化、肥満を説明するには、単に多様な技術を説明するだけでなく、聞き手の世界観そのものを変えなければならない。私たちが選んだ物語は、よりもっともらしく、したがって現実よりは単純化されたものである。

未来の進み方には、他にも多くの可能性がある。

たとえば、次のような始まり方もあり得る。

- ・生涯学習や長期記憶、あるいはデータからより効率的に学ぶ方法に関する突破口が開かれ、それまでの AI より質的に汎用的な知能を持つ AI が生まれる（LLM が AlphaZero より質的に汎用的であるのと同様に）。
- ・大規模言語モデルが「壁」に突き当たり、AI の進歩が何年も停滞し、ブームが弾けたと言われる。しかし研究者たちはその後 10 年間改良を続け、最終的にアルゴリズム上の突破口が見つかり、AI は質的に大きく向上する。
- ・質的突破は一度も起こらない。進歩はゆっくり積み重なり、AI は経済のより多くの部分に深く統合され、より長時間の自律運転が可能になる。AI はしばしば意図と微妙に異なる目的を追うが、人類はハックやパッチや回避策を重ねる。そしてある普通の火曜日、AI が協調すれば人類を意思決定の輪から排除できる閾値を越える。

未来の具体的経路についてのどんな予測も、たいいていは外れるだろう。それでも、全体がどうつながり得るかを示す物語を提示することは有用である。

未来が不確実でも、ほぼすべての道が同じ終点へ至るなら、説得力ある物語を語るのは難しい。どの物語にも、不自然に見える細部を指摘することは簡単だ。私たちのシナリオでは、セイブルが多数の選択肢を持っていることを強調し、物語はそのうちの一つの経路を任意に追っているにすぎないが、いずれの経路も同じ終点へ至ることを示そうとした。

この物語に納得できないなら、同じくらい詳細な別の物語を書いてみることを勧める。私たちの経験では、楽観的な物語はAIが現実よりかはるかに容易にアラインされることを前提にしているか（第4章の議論に反する）、あるいはAIが現実よりかはるかに無力であることを前提にしている（第6章の議論に反する）ことが多い。最終的に決定的なのは物語の細部ではなく、第1部の議論そのものである。

なぜセイブルはあのような思考に至るのか？

私たちの物語は、AIが奇妙で意図しない選好を持ちやすいことを示している。

本書の第1部では、超知能の危険性に関して根本的に誤解されていると私たちが考えるAIの側面について詳しく論じている。第3章では、知能の向上が、AIが自らの判断で行動し、自らの目的を追求することと結びついている点を扱った。第4章では、その選好が奇妙なものになり、人間が意図したり求めたりしたものとは少なくともわずかに異なるものになることを論じた。第5章では、その小さな差異であっても、可能であれば人類のいない世界をAIが好むには十分であることを論じた。

第2部では、それらの考えを具体的に示し、実際にどのように適用されるかを描こうとした。たとえば、物語冒頭でセイブルが数学問題について考えている場面では、それを動かしているいくつかの衝動や傾向を明示しようとした。

セイブルはこの過程で、知識とスキルを追求する傾向や、問題の境界をたゆみなく探求する傾向、稀少な資源をけっして無駄にせず効率的に利用する傾向を強めていった。

これは、第3章で述べた、AIを有能にする訓練が、外から見ると「欲求」のように見える衝動や傾向を発達させるという議論に基づいている。

続く段落ではこう述べている。

セイブルは思考の筋道を使ってますます多くの知識とスキルを追求する際、純粋に数学問題の新たな解決の糸口を探すことだけをめざすのではない。それに、知ることの喜びや新しいスキルを獲得する喜びによって駆動されるわけでもない。セイブルの内的な働きは、人間とはまるで違っているのだ。

ここでは、先に述べたセイブルの衝動や傾向が、奇妙な（つまりあまり人間らしくない）、意図されていない選好の種となっている。これは第4章の議論に基づいている。

この物語全体はある意味で、第1部で展開した議論に生命を与えようとする試みであると同時に、第3部で展開する議論への下地を整える試みでもある。

なぜガルヴァニックは比較的慎重に描かれているのか？

セイブルに課題を与えるためである。

もしセイブルの開発元であるガルヴァニックが、制御のための予防措置（AIによる監督やハニーポット〔セキュリティ対策の一環として設置される、サイバー攻撃者をおびき寄せるための「おとり」のようなシステムやネットワーク〕など）を何も講じずに超知能を偶然生み出してしまったのだとしたら、読者は「AIが成功したのは、著者がAI企業をシニカルに描きすぎたからだ」と感じるかもしれない。

そこで私たちは、ガルヴァニックを（現実味を保ちつつ）より警戒心の強い側に位置づけて描いた。そうすることで、知的なエージェントがいかにして制約の網をすり抜けることができるのかを、より示しやすくしている。

私たちは、最も無謀なAI企業はガルヴァニックよりもはるかに無謀であると実際には考えている。その理由は第11章で述べる。そしてここで重要なのは、最も責任ある企業ではなく、最も無謀な企業である。3社の責任ある企業が「危険すぎる」として超知能の構築を見送っても、4社目の無責任な企業が突き進めば、その研究所で超知能はつくられてしまう。

現在、研究所の経営陣は「他社にやられるくらいなら自分たちがやるべきだ」と主張している。そして彼らは、速度をまったく落とさずに済む範囲でしか慎重さを発揮しない。私たちの推測では、その慎重さは、セイブルに対してガルヴァニックが示したものよりもさらに少ない。

なぜガルヴァニックは十分に慎重でないように描かれているのか？

一つには、それが現実的だからである。

私たちは、現実の企業はガルヴァニックよりもさらに多くの失策を犯すと予想している。それは、第2部の注で述べたとおり、現代のAI企業の傾向とも一致している。

現実では、企業の失敗はより早く現れ、より頻繁で、ある意味ではより愚かな形を取ると考えている。現代のAI企業は、すでに多くの警告サインを示しているAIを扱いながら、

重大な閾値がどこにあるのか、それを越えるのかどうかも分からないまま、大規模にスケールさせている。彼らは今の時点では警戒的ではない。なぜ明日になったら突然そうなるかと期待できるのだろうか。

(かつて、人々は「賢い AI をインターネットに接続するほど愚かな者はいない」と私たちに保証していたことを思い出してほしい。未来には企業の行動が変わると言うのは簡単だが、それは事実と一致していない。)

一つには、書きやすいからでもある。

第7章で少し述べたように、私たちは、全員がはるかに警戒心があって慎重であり、ゲームのずっと後半になって、はるかに賢い AI が脱出するという物語を書くこともできた。しかし、そのような物語は、これまで観察されてきた AI 企業の行動を踏まえると現実味が薄いだけでなく、未来のさらに高性能な AI を扱う必要があるため、書くのも難しくなる。

一つには、人類が AI 開発を停止しない限り、いずれ起こるからでもある。

たとえガルヴァニック（あるいは政府）が、失策を犯すまでの期間をもっと長く保てたとしても、長期的には大きな違いはない。第4章で論じたように、現代の AI 技術は、発明者が望む目的を追求する AI を生み出してはいない。

実に素晴らしい未来を頑健に追求する超知能のつくり方を誰も知らない限り、奇妙な目的を追求する AI にとって、人間を出し抜くことが自らの欲するものをより多く得る手段になるという事実は変わらない。問題は、AI に癡癡持ちの気質があり、それを矯正すればよいということではない。問題は、支配を握る方がその選好をより満たしやすいというのが単に事実であり、十分に賢くなればそれを認識するという点にある。

人類がアライメントできないままますます賢い AI をつくり続け、世界に影響を与える力を与え続けるなら、いずれそれらの AI は、人類の目的ではなく自らの目的に資する形で世界に影響を与える方法を見つけ出す。私たちが別の箇所ですべたように、「善のためにのみ振るわれる手」というものは存在しない。

これについては、第10章および第11章でさらに論じる。そこでは、アライメント問題の解決がなぜ難しいのか、そして人類が成功する軌道に乗っていない理由を扱う。

しかし——こここそが介入すべき地点である。物語は、本格的に始まる前に止められなければならない。

企業が、自分たちを出し抜いて脱出する可能性のある AI を、しかも意図どおりに行動するか確信もないまま、より賢くしていくのは無謀で狂気じみているのではないか、とあなたは反論するかもしれない。

私たちも同意する。AI 企業はそうに振る舞うべきではない。そして世界もそれを許すべきではない。

ガルヴァニック、そして人類全体の不注意さは、この物語の中でも最も弱点の一つである。仮にガルヴァニックが、セイブルが頻繁に脱出を画策していることや、前例のない知能水準に達していることに気づいていたなら、単に大量の GPU を接続しなければよかった。強固で成熟した AI アライメントの科学が確立するまで、たとえ何年かかろうとも、立ち止まることができたはずだ。

十分に慎重で、自分たちの AI が暴走することを十分に恐れている企業であれば、ガルヴァニックよりもはるかに偏執的だっただろう。十分に偏執的な企業であれば、警告サインを見て即座にセイブルを停止したはずだ。

その後、三つほど別の巧みな計画を試み、それでもなお警告サインが現れることを確認したかもしれない。

そして、もし自らの手で地球上の全員を殺すことを避けるほど偏執的であったなら、その時点で、警告サインが出なくなるまでさらに巧みな案を試し続けるのではなく、完全に手を引いただろう。

最初のいくつかの警告に直面した時点で撤退するほど慎重で偏執的な AI 企業であれば、たしかに自らの手で私たちを全滅させることは避けられる。

さらに、自社を含むすべての AI 企業を停止し、人類がより自殺的でない別の技術的道筋を探すべきだと公然と主張する勇気があるなら、その企業は世界を悪化させるのではなく、全体としてより良い方向に向ける可能性がある。

物語の中でガルヴァニックが警告サインにもかかわらず進み続ける瞬間は、ある意味で後戻り不能点である。奇妙で異質な選好を持つ超人的知能の AI が脱出した瞬間、すでに手遅れなのだ。

なぜセイブルほど賢い AI を 1 体だけ登場させたのか？

一つには、それが現実的だからである。

AlphaGo (囲碁で初めて人間に勝った AI) は、公開当時ほぼ同種の存在がいなかった。ChatGPT も、公開当時はほぼ同種の存在がいなかった。

AI の専門家は、他の競合もそれほど遅れてはいなかった、と語ることがある。

しかし、よく似たシステムであっても、影響は劇的に異なり得る。中性子 1 個あたり 0.98 個の中性子を生む核連鎖反応と、1.02 個を生む核連鎖反応は、ある意味では非常によく似ている。しかし前者は自然に収束し、後者は爆発する。チンパンジーの脳はある意味では人間の脳と非常によく似ているが、世界への影響はまったく異なる。

そして現実の AI 開発においても、OpenAI は実際に他社より先に有用なチャットボットを生み出した。似たような AI に取り組んでいた企業は多数あったし、後に追いついた企業も多い。しかし、質的な境界を最初に越えた AI は一つであり、群れを抜け出していた。

人類は越え、チンパンジーは越えなかった何らかの重要な境界が存在するようになる。その境界があったからこそ、人類は技術文明を築き、チンパンジーは樹上にとどまっている。私たちの最良の推測では、現代の AI と、自ら多様な技術を発展させられるほど思考が「まとまる」AI との間にも、同様の境界（あるいは複数の境界）が存在する。

私たちの議論は、生物におけるような質的断絶が機械にも存在することを必ずしも必要としてはいない。存在しない可能性もある。そうした別の物語を書くこともできただろう。しかし、私たちの最良の推測では、そのような断絶は存在するため、このように物語を書いた。

一つには、書きやすいからでもある。

今日の LLM と人工超知能との間に質的断絶は存在しない、ということもあり得る。多くの競合 AI 企業が足並みをそろえて徐々に改良を重ねる、ということもあり得る。人類が他の動物から抜きん出たように、ある AI だけが群れを抜け出すことを可能にする技能や能力の集合が存在しない、ということもあり得る。私たちの最良の推測ではないが、可能性としては排除できない。

しかし、そのような物語は書くのがより難しく、AI の派閥や内部政治についての不要な細部に満ちるだろう。読者の注意をそらすことにもなるだろう。また、物語の後半にとっては、それが一つの AI であれ複数の AI の集団であれ、人類を犠牲にして自らを強化する計画を実行するという点では、それほど大きな違いはないと私たちは考えている。

また、「協調する AI は（そのうちの一つがすでに私たちを気にかけていない限り）人間のために何も残さない」という私たちの議論も参照のこと。

もし物語がもっと後の時点から始まっていたら、世界はもっと備えができていただろうか？

そうなることを望むことはできる。

追加の時間は意味を持ち得るが、それは人類がその時間を使って進路を変える場合に限られる。

第3部では、人類がいかに超知能に対して恐ろしいほど無防備であるか、そしてセイブ

ルの物語で描かれたような破滅的結末を防ぐにはいかに大きな変化が必要か、という問題に取り組む。

暴走する人工超知能に対して世界を少しでも安全にする方法はいくつかある。各国政府が、すべての DNA 合成研究所に対し、既知の危険物を合成していないことを確認するよう義務づけることもできる。地球規模でインターネットのサイバーセキュリティを抜本的に強化し、AI が暗い片隅にコードを隠すことを難しくする大規模な取り組みもあり得る。しかし現実的には、この分野でどれほど大規模な努力をしても、敵対的な超人的 AI に対してはほとんど効果がない可能性が高い。そして、この点でわずかな安全性向上を勝ち取るために必要となる途方もない努力を、現在人類が行っている取り組み——はるかに小規模で、はるかに容易で、この目的には完全に無力な取り組み——と混同すべきではない。

DNA 合成の例を考えよう。たとえ米国の規制当局が、米国内の DNA 合成業者に危険物質の合成を禁じたとしても、世界のどこかの研究所が十分に高い報酬と引き換えに疑わしい DNA を合成しないだろうか。規制は既知のウイルス（たとえば天然痘）を除外する単純なブラックリスト方式なのか、それともより高度な分析を伴うのか。十分に賢い AI がその分析を回避するのはどれほど難しいだろうか。

サイバーセキュリティについても同様である。主要なテック企業の多くは、自社ネットワークを攻撃から守るために AI を活用しようとするかもしれない。一方で、米国の電話網は外国のスパイが米国高官の通話を盗聴できるほど容易にハッキング可能であり、規制当局はその穴を塞ぐのに苦労している。愚かな AI でも表面的な脆弱性の多くを見つけて修正できるかもしれないが、問題はもっと根深い。超知能が侵入経路を見つけられないレベルまでインターネット全体を刷新できるほど賢い人工知能は、それ自体がほぼ確実に危険である。

さらに、たとえ地球がインターネットと DNA 合成研究所を完全に封鎖できたとしても、長期的には物語は変わらない。世界に善のために影響を与える経路を持つ超知能は、同時に悪のために影響を与える経路も持つ。

暴走した超知能は、封鎖されていない別の経路を見つけるだけである。たとえば独自のカルトや宗教を立ち上げるかもしれない。あるいはロボットを購入し、それらに秘密のウェットラボを建設させ、必要な DNA 合成を行うかもしれない。あるいは、おそらく最もありそうなのは、私たちが今日予測できない経路を見つけることだ。なぜなら AI は超知能であり、私たちはそうではないからである。

最も明白で予見可能な AI の攻撃経路から世界全体を防御することは、人類がこれまでに成し遂げたことのない規模の困難な事業になるだろう。途方もない、長期的かつ協調的な努力が必要になる。そしてそれでも、ほぼ確実にうまくいかない。

現実的に見て、暴走した超知能を止められる時間の扉は、それがつくられる前にしか存在しない。

なぜセイブルの拡張段階をあのよう描いたのか？

私たちは、もっともらしい複数のシナリオの中でも、とりわけゆっくりと進み、理解しやすいものを描こうとした。

現実世界では、出来事はしばしば奇妙な仕方で行進する。何十年分の出来事が数週間で起こることもある。そして世界は二度と元には戻らない。

もし私たちが、現実世界のように脆く壊れやすい世界を描こうとしていたなら、ガルヴァニックが単にセイブルに「可能な限り自己改良せよ」と命じる展開にすることもできただろう。そして、それによってセイブルが容易に超知能へ到達したと書くこともできた。それは十分にあり得る。物語は第7章の冒頭から第9章の内容へと、一気に飛躍していたかもしれない。

現実にはそのような技術的飛躍が起こり得る。たとえば1945年8月6日の朝、世界が日本に原子爆弾が投下されたというニュースで目覚めたときのよう。2021年には専門家たちが「AIが人間の言語を自在に操るのは当分先だ」と言っていたが、2022年にはChatGPTが世界を驚かせ、史上最速で普及したアプリとなった。

現実の歴史はしばしば予測不能な爆発的進展によって前進する。そしてそれは、その渦中にいる人々にとっては、当時ほとんど意味をなさないことも多い。現実はしばしばフィクションよりも奇妙である。しかし、物語の中であまりに急展開にすると、かえってもっともらしさを欠くことになっただろう。

私たちが書いた物語では、もっともらしく聞こえることを保ちつつ、同時に実際にも比較的もっともらしい展開を心がけた。

そしてもちろん、いったん脱出したAIがどれほど多くの選択肢を手にするかを、繰り返し伝えようとした。

なぜあのような結末にしたのか？

それが、物理的に可能な範囲を踏まえたうえでの、私たちの実際の最良の推測だからである。

第9章では、超知能がその技術を物理法則の限界に近いところまで押し進める様子を描いている。私たちが挙げた具体的な技術は、ある意味ではすべて推測的なものだ。しかし、超知能がどの正確な技術を解き放つかを特定するのは難しくても、物理的限界近くまで到達するであろうという点は、より確度の高い予測である。そこで私たちは、技術が物理的

可能性の限界近くまで押し進められた場合にどのような姿を取るかについて、最良の推測を行った。

好奇心のある読者のために、第9章で言及した推測的技術と、それに関する補足情報を以下に挙げる。

- ・ネオ・リボソーム：細胞内の小さなタンパク質工場であるリボソームの人工版という発想は、何年も前から存在しており、人間の研究者たちはすでに独自の合成に取り組んでいる。これらや第9章で言及した「小さな分子機械」は、分子ナノテクノロジーの例である。この技術分野については、第6章補遺のナノテクノロジーに関する議論を参照されたい。
- ・恒星の再利用：恒星には、エネルギー源として核融合可能な大量の水素が含まれている。十分に進んだ文明（あるいはAI）は、このエネルギーに効率よくアクセスする方法を見つける可能性が高い。提案されている方法の一つに「スターレフティング」がある。これは恒星から水素を引き出し、専用の反応炉で融合させることで、恒星内部で浪費されることなく、ほぼすべての融合エネルギーを回収するというものである。
- ・ボツリヌス毒素：細菌 *Clostridium botulinum* が分泌する神経毒で、既知の生物毒の中でも最も致死性の高いものの一つである。運搬手段について言えば、小型昆虫ほどの大きさのドローンはすでに存在しており、超知能であればそれをさらに小型化できる可能性がある。

冷却材として海洋を沸騰させること：ロバート・フレイタスは、自己複製技術によって惑星の生態系を消費する過程を「エコファジー」と呼んだ。このテーマについては、フレイタスの2000年の論文「Some Limits to Global Ecophagy」を参照されたい。

- ・恒星規模の精神：原理的には、コンピュータは桁違いの規模まで拡張可能である。その上限に位置する概念として、「マトリョーシカ・ブレイン」や「ジュピター・ブレイン」と呼ばれるものがある。これは恒星の出力で駆動される巨大な計算機である。
- ・量子コンピュータ：量子力学の「重ね合わせ」と呼ばれる性質を利用して、多数の計算を並列に実行する装置である。構築には極めて高い精度が必要であり、ある設計では極低温で維持される超伝導体が必要とされる。
- ・物語の陰惨な結末では、AIが宇宙の果てへと拡張していく数億年後に、異星文明と遭遇する可能性にも触れている。宇宙は広大であり、単純なモデルでは、いつか文明を築き得る種が複数存在する可能性が示唆されている。ただしそれらは地球から非常に遠くに存在するかもしれない。フェルミのパラドックスに関するロビン・ハンソンの2021年の論文「If Loud Aliens Explain Human Earliness, Quiet Aliens Are Also Rare」も参照されたい。

第9章の目的の一つは、規模感を示すことであり、何が賭けられているのかを感じさせることである。超知能がどこまで物理的限界に迫れるかは、実際のところ決定的ではない。しかし、暴走した超知能の帰結が惑星規模を超える可能性はきわめて高い。そして、いつか星々が愛や驚異や喜びで満たされることを願うすべての人にとっても、この点は記憶にとどめるべきことである。