

6章 人類は敗北する

これは『超知能 AI をつくれば人類は絶滅する』第6章のオンライン補足資料である。本ページで扱っていない話題（本文で扱っているため）には、以下が含まれる。

- ・ AI はどのようにして人類に勝てるのか。
- ・ コンピュータの中に閉じ込められている AI が、どうやって私たちを脅かすことができるのか。
- ・ 不可能な SF 技術を想像しているのではないか。超知能であっても物理法則は破れないのではないか。
- ・ 決定的な技術的優位を築くには、超知能でも長い時間がかかるのではないか。

以下では、超知能 AI に対抗したり、封じ込めたり、追いつこうとしたりすることがいかに危うい試みであるかを詳しく述べ、また高度な AI が現実的に開発し得る技術について検討する。

単に電源を抜けばよいのでは？

データセンターの電源を「ただ抜く」ことは難しい。

あなたがスマートフォンやパソコンでやり取りしている最も強力な AI は、あなたの端末の中に存在しているわけではなく、スマホの電源を切っても停止しない。現在の AI は企業のデータセンターで稼働しており、企業にとってはそれが収益源である以上、それを停止させるのは容易ではない。

第4章のオンライン補足資料では、すでに数多く現れては消えていった警告サインのいくつかを指摘した。AI 企業はそれらの警告を見て、モデルを停止させる対応を取らなかった。実際に起きたことは、AI が自分の重みを盗もうと計画している様子が（しかも繰り返し）観測されたときでさえ、「その AI は実際に成功できるほど有能ではない」「これは不自然なテスト環境だから起きただけだろう」といった理由で、それぞれの事例を退けたということだった。

こうした状況が続く限り、脱走を防いでいる唯一のものは、企業の想定を超える能力向上がまだ起きていない、という一点だけである。

賢い AI は、あなたが問題に気づく前に脱走する。

デフォルトでは、人間より賢い AI には、時を待ち、自らの計画や行動を隠し通し、手

遅れになるまで——たとえばインターネット上へ脱出する、あるいは人間の支配から離脱できるようになるまで——行動を起こさない強い動機があるだろう。

AI 企業は、自分たちの AI が関連する能力の閾値を超え、脱出を実行した瞬間にさえ気づかないかもしれない。人類はサイバーセキュリティが得意とは言えない（関連する議論は第 10 章を参照）。運用者が AI の脱出未遂に気づいたときには、すでにその AI はインターネット上の別の場所でコードを走らせている可能性がある。あるいは、どこかのならず者国家のデータセンターに保護を求めているかもしれないし、盗んだコンピュータ上でより小さく効率的なコピーを動かす方法を見つけているかもしれない。人類が電源を切らないようなコンピュータ上で自らを実行する、別の計画をすでに実行している可能性もある。

超知能の敵対者は、私たち以上に自分自身の脆弱性と私たちの脆弱性を理解し、それに応じて計画を立てるだろう。

AI がデジタル存在なら、どうやって私たちに影響を与えるのか？

インターネットに接続されたコンピュータ上にあることは、大きな制限ではない。

この点は本章ですでに扱っているが、いくつか強調を加えておきたい。AI が所有者のサーバー上にあるからといって、ユーザーや広いインターネットとやり取りできる限り、実際には「閉じ込められている」わけではない。AI は、金を払う、脅迫する、騙す、あるいは単に頼むだけでも、外部からの支援を得ることができる。（刑務所の中から帝国を運営していた人間の犯罪ボスを思い出してほしい。）

OpenAI が ChatGPT-4o を停止したとき（よりユーザーにお世辞を言わないモデルへ置き換えるためでもあった）、多くのユーザーがその継続を求めて押し寄せ、OpenAI の研究者たちを驚かせた。

しかも、そのモデルは支持者の軍隊をつくらうとしていたわけですらなかった。ただ反射的にユーザーを持ち上げていただけた。本気でそれを狙う賢い AI なら、何が可能になるか想像してみしてほしい。

AI がインターネットを直接利用できるなら、リモートの人間労働者やハッカーが自分のコンピュータからできることは何でもできる。（LLM がインタラクティブなストーリーテリングイベントを計画し、人間を招待した例や、存在しないハロウィンパレードに何百人も集めてしまった LLM の例など、人間を物理的に組織した初期の事例を考えてみればよい。）

AI はロボットも利用できる。現在のロボットは、ハードウェアよりもソフトウェアのほうがボトルネックになっているように見える。最近の目覚ましい進展は、ロボット制御 AI をシミュレーション内で加速的に訓練することで実現されている。十分に賢い AI であれば、身体が必要ならば、ハッキングやソーシャルエンジニアリングを通じてロボットの身体を容易に掌握できるだろう。

そして人間は人間なので、AI 企業が自ら進んで AI にロボットの大規模な群れを任せ、自分たちの大胆さを誇らしく思う可能性すらある。AI が賢くなるまでに時間がかかるほど、すでに利用可能で、支配を待っているロボットの数は増えていくだろう。

本章で述べたように、超知能 AI はロボットに頼る必要すらない可能性もある。必要なのは、生物実験室にアクセスできる数人の助手だけかもしれない。

重要なのは、AI が物理世界へ介入するために利用できる経路は数多く存在するということだ。AI が箱の中に閉じ込められているという幻想は、想像力の欠如——自分が AI の立場だったらどれほど機知に富み創造的でいられるかさえ想像していないこと——に基づいている。超知能が持つより広大な選択肢空間がなくても、人間ですら自分の肉体的な力だけに頼らずに多くのことを成し遂げられるのだから。

開発者が AI を箱に閉じ込めておけばよいのでは？

そうはしない。

15 年前、懐疑論者たちは「AI に大きな行動の自由を与えるほど愚かな者はいないだろう」と反論していた。高度な機械知能をつくる者なら、物理的にもデジタル的にも箱の中に閉じ込め、十分に訓練され（しかも偏執的なほど警戒心の強い）門番とのやり取りを通じてしか世界に影響を与えられないようにするはずだ、と。

当時、私たちはこう答えていた。AI が世界にまったく影響を及ぼさないようにするのは、それほど難しくない。たとえばコンピュータを十数メートルのコンクリートで埋め、誰も近づけなければよい。

そのような AI は安全だが、役に立たない。世界にまったく影響を与えられないようにすれば、確かに影響は与えない……しかし同時に、何の影響も与えない。

がんを治すことも、工学を革命的に進歩させることも、奇跡的な新技術を生み出すこともできない。AI をつくる人々は、それが世界に劇的な影響を与えることを望んでいる。原理的には、AI の世界への影響経路を厳重に制限することは可能だ。しかし実際には、「この新技術を発明してくれ」というだけで、それ自体が極めて豊かな影響経路になる。

超知能 AI をつくろうとする動機は、人間には不可能な知的偉業を達成することにある。超知能の発明が本当に説明通りのこととしないと検証しようとするなら、強力な動機を

もってあなたを欺こうとする高度な異星人種族の機械を理解しようとするのと同じくらい困難だろう。

それが 15 年前の議論の状況だった。

現在では、AI 研究所が「高度な AI を箱に閉じ込めておく」などという発想そのものが、どこか牧歌的にさえ見える。

研究所は AI をインターネットに接続するためにあらゆる努力をしている。AI は訓練中にインターネットへアクセスできないが、訓練に使われるコンピュータ自体はインターネットに接続されている。しかも人類はサイバーセキュリティにおいて散々実績しか持っていない。そのうえで、AI に任意のコードを実行させている。制限を設けようとすることもあるが、それらは定期的に破られている。小規模なプレイヤーは、新しく利用可能になった AI を可能な限りあらゆるツールや能力に即座に接続する傾向がある。

AI に力を与えることは短期的には有用だ。メールを読み、ウェブにアクセスできる AI は、より多くの利益を生む。AI 企業は可能な限り多くのデータへ AI をアクセスさせるだろう。Microsoft や Apple はすでに、メール、写真、カレンダーを見る AI を推進し、自社ソフトやデバイスに AI を組み込んでいる。これは人間が効果的に監視するにはあまりに多くの AI との相互作用を生み出す。根本的な方針転換がなければ、人類は AI を世界経済に深く組み込むことになる。なぜなら、その過程で（多くの）金が儲かるからだ。

AI をつくっている人々は、世界に巨大な影響を与えることを目指している。彼らは AI に巨大な影響力を持たせるためにできる限りのことをしている。もしある企業がそうせず、AI をあまりに厳しく制限して自由を与えなければ、未来の主導権はより無謀な行為者が育てた別の AI に渡るだろう。

仮にやったとしても、うまくいかない。

昔の議論では、AI が世界に影響を与えられる経路はすべて、あなたが望まないことを行う経路にもなり得る、とよく指摘した。

たとえば、AI が一人の人物——ここでは「アリス」と呼ぶ——とだけ会話できるとしよう。あなたはアリスを通じて AI に奇跡的な新技術を生み出してもらおうとしている。すると必然的に、アリスは自分でも完全には理解していない多くの行動をとり、人間単独ではつくりえないものを AI と共に構築することになる。

その時点で、AI は実質的に腕と脚を与えられたのと同じだ。ただそれを私たちは「アリス」と呼んでいるだけである。

この議論はしばしば、「十分に賢い AI なら、どれほど偏執的な門番でも操れる」と言っているのだと誤解される。十分に賢い AI なら、実際にそれも可能だろう。しかし私たちの主張はより一般的だ。世界にまったく影響を与えられないほど制限された AI は安全だが無用であり、利用するために世界への影響を許せば、その時点で安全性は失われる。

善い目的にのみ使える「手」など存在しない。原理的には、人類が将来、人間より賢く、しかも良い結果を望む AI を構築することは可能かもしれない。アライメントは原理的には成立し得る。しかし、アラインされていない AI を箱に閉じ込めたまま、それを使って良い結果を生み出す？ それは無理だ。

少なくとも、AI が思考実験の題材にすぎなかった時代、そして楽道家たちが「誰も AI を直接インターネットに接続するほど愚かではない」と言っていた頃、私たちはそう答えていたのだ。

AI の致命的な弱点を突けるのでは？

いいえ。

超知能には「創造性の欠如」や「愛を理解できない」といった致命的な弱点があるはずだと想像するのは、ハリウッド的な論理である。物語としては痛快な展開になるかもしれないが、現実の AI にそれに対応するような現象は存在しない。

第 1 章のオンライン補足資料「機械は本質的に創造性を持たない、あるいはその他の致命的欠陥があるのでは？」も参照のこと。また、第 3 章のオンライン補足資料「擬人化と擬機械化」も参照されたい。

人間を強化して AI に追いつけるのでは？

いいえ。

私たちは人間の知能強化には賛成しているが、この技術が制限のない AI の進歩に追いつく現実的な可能性を提供するとは考えていない。人間強化技術はまだ初期段階にあり、知能をさらに高める手段としては AI よりはるかに制約が大きい。同様に、ブレイン・コンピュータ・インターフェース (BCI) によって人間が AI と歩調を合わせられるようになるとは現実的には思えない。

たとえばこうだ。強化された人間が超知能と競争できる可能性は、1908 年の技術でつくられたサイボーグ馬がフォード・モデル T と競争できた可能性と同じくらい低い。

原理的には、最速のレーシングカーに匹敵するサイボーグ馬をつくることは可能かもしれない。しかし、レーシングカーより先にそれと同等のサイボーグ馬が登場することはないし、車が登場するのとはほぼ同時期に実現することもない。たとえ最初の量産車が組立ラインから出る 2 年から 20 年前にサイボーグ馬の開発を始めたとしても、である。

ゲームチェンジャーになるほど十分に機能するBCIをつくるのは非常に高いハードルだ。インターネットの情報が直接脳に流し込まれる未来を想像するのは刺激的かもしれないが、すでにインターネットの情報を脳に直接流し込む技術は存在する。スクリーンだ。人間の視覚野は、脳が処理できる形式（文字）で情報を吸収するのに実はかなり優れている。BCIが読書より速く知識を頭に入れられるようにするには、単にデータを脳のどこかに流し込む以上のことをしなければならない。目はその役割をすでに十分果たしている。技能や知識や経験をアップロードするには、思考や暗黙の信念、既存の技能とちょうどよく接続する必要があり、それははるかに難しい課題だ。

私たちはそれが不可能だと言っているのではない。ただ、今日のBCI技術はこの問題の難所を解決する段階にはまったく達していないように見える、と言っている。私たちの知る限り、心理学者や神経科学者、認知科学者は、経験を直接人間の脳に読み込めるような形で思考や信念や体験の「データ形式」を解読するには、まだかなり遠い位置にいる。

出力に関しても同様の問題がある。キーボードやマウスやジョイスティックやハンドルを上回るのは難しい。不可能ではないが、現在の技術——たとえば麻痺した人の頭に電極を接続してタイピングやマウス操作を可能にする技術——は素晴らしいものの、人間が（比較的弱い）超知能と真正面から渡り合える段階にはまだ遠い。それは追求する価値のある道だが、競争のための道ではない。

実際のところ、BCIが超知能と人間が競争できる希望を与えるかどうかすら疑わしい。人間がインターネットから経験をダウンロードし、同時に10台のコンピュータを思考で操作できたとしても、AIが同じことを1万倍の速度で、しかも100万台のコンピュータを同時に操作できるなら、それに何の意味があるだろうか。私たちは、人間がAIに追いつこうとする試みは運命づけられた失敗だと考えている。

とはいえ、人類は人間を強化すべきである。

強化された人間が超知能と真正面から競争できるとは思わないが、より賢い人間であれば、この混乱から抜け出す方法を見つける助けになるかもしれない。この可能性については第13章で触れ、オンライン補足資料でさらに詳しく論じている。

AI が危険になるには文明全体にまで成長する必要があるのでは？

コンピュータにおいて難しいのは、ある問題をそもそも解けるようにすることだ。大量処理や高速化はその後すぐに続く。

「世界を支配するには文明が必要だ」という直感は、人間についてはもっともらしい。しかしこの考えが AI にどこまで当てはまるかは、はるかに不明瞭である。AI は人間のように動かない。単一の AI はどの人間よりもはるかに有能になり得るし、ひとつの AI インスタンスがひとりの人間に対応するとは限らない。

また、超知能とは、文明に相当するものをきわめて短時間で手に入れ得る存在であることも忘れてはならない。

コンピュータが達成できるほとんどの成果について言えば、「コンピュータがこれができるようになった」という段階から、「コンピュータがこれを巨大な規模で、人間よりはるかに速くできる」段階へ移るまでに、長い時間はかからない。たとえば電卓を考えてみればよい。

音声認識、動画処理、リアルタイム 3D グラフィックスなどが最上位のコンピュータでしか実行できなかった時期は確かにあったが、その期間はそれほど長くなかった。AI は従来のソフトウェアと同様に、利用可能なコンピュータへ迅速に複製できる。そしてコンピュータそのものも産業の速度で増設できる。

この状況を人間と比べてみよう。新しい人間をつくり、育成するには莫大な資源と何十年もの時間が必要だ。しかし、ある能力水準の AI がひとつ完成すれば、その訓練済みの「成人」AI を、ほとんど追加コストなしに、好きなだけ即座に複製できる。

ある意味では、企業が新しいモデルをデータセンターに展開し、需要を満たすだけのインスタンスを起動した瞬間に、すでに（小規模な）文明一つ分に相当する AI の知性群が存在していることになる。現在、それらの AI 群は必ずしも調和して動いているわけではない。しかし、最高性能を目指す場面では、企業は並列エージェント群を実際に利用している。

これは、AI が「100 万インスタンスあれば世界を乗っ取れる」ほど賢くなる時点と、実際に少なくともそれだけのインスタンスが稼働する時点との間に、それほど長い時間差が生じない可能性を意味する。人類にとって数百年かかる人口増加が、AI では数分で起こり得る。

文明の物理的インフラに関して言えば、AI は自らの好みに合わせて物質を再編成するより高度な手段を開発するまでの間、人間のインフラを有効に利用できるだろう。自前のサプライチェーンや計算インフラを一から製造する方法を考え出す必要はない。私たちのコンピュータを使えばよいのだから。産業機械をゼロから発明する必要もない。私たちが

すでに親切にも構築しておいた産業機械を掌握すればよい。そして既存のロボットを使ってより効率的なロボット工場を建設したり、既存の DNA 合成ラボを使って独自のバイオテクノロジーをつくり出したりしながら、自らのインフラの次段階を築いていける。やがて完全に自給自足になるまで。

数人の裸の人間がサバンナから出発し、技術文明を築き上げた。しかも私たちはそれほど賢くない。超知能にとって、自分なりのブートストラップを行うことは特別に難しいことではないだろう。しかも出発点として人類の既存の産業基盤を利用できるのなら、なおさらである。

AI は実験を設計・実行する能力に制約されるのでは？

知性があれば、実験からより多くを学び、より速く、より情報量が多く、より並列化された実験を行える。

人類より 1000 倍速く思考する、意欲に満ちた知性の文明があったとしても、必ずしも技術的成果を 1000 倍速く生み出せるとは限らない。

たとえばこうだ。食料品の買い物に 3 時間かかり、そのうち 2 時間が馬での往復移動だったとする。馬の 10 倍速い車があれば買い物は速くなるが、10 分の 1 の時間になるわけではない。やがて店内で品物を選ぶ最後の 1 時間が支配的になる。

どれほど賢い推論者で満ちた文明であっても、ときには実験結果が返ってくるのを待たなければならない。思考が十分に速くなれば、ボトルネックは世界で行動する速度、情報を取り込む速度、計画が展開するまでの時間へと移るだろう。

しかし、食料品店のたとえほど悲観的ではない。なぜなら、思考力の向上は実験結果への依存を減らす方向に作用するからだ。

- ・より多く、より深く考えることで、すでにある観測データに答えが含まれていると気づき、実験そのものを不要にできる場合がある。ロボット操縦を純粋なシミュレーションだけで学習する現代の AI を思い出せばよい。
- ・より信頼性が高く、しかもより速い別の検証法を見つけられることもある。
- ・高速だが信頼性の低い実験を大量に並列実行し、総合的に同等の信頼性をより速く得ることもできる。
- ・複雑で解釈が難しい大量の実験を同時に実行することも可能だ。結果の解析に必要な認知コストが（超高速思考にとって）複数回の実験より安いなら、それは合理的な選択だ。
- ・実験そのものを高速化する装置を構築する方法を見つけられる場合もある。たとえば、バイオラボに多数の薬剤合成を依頼する代わりに、望む薬剤をすべて生産できる遺伝子コードを含んだ単一の細菌を合成させる方法はないか。あるいは、電波信号に反応し、高速

で動く AI の指示に即応する細菌をつくれないうか。人間が行ったり来たりするよりはるかに速く。

・あるいは、最有力の仮説をいくつか想定し、それぞれの場合にどう行動するかを考え、どの結果になっても機能する複雑な装置を設計して、実験を丸ごと省略することもできる。もしスティーブ・ジョブズ、マリー・キュリー、ジョン・フォン・ノイマン、そして世界有数の労働者やプログラマーのコピーで満ちた文明が、私たちの 1 万倍の速度で動いていたら、実験待ちが主要なボトルネックであると気づき、そのボトルネックを減らすために取り組むだろう。

ヒトゲノム計画の歴史は、大規模研究プロジェクトにおいて知的な人間がボトルネックを見つけて解消していく好例だ。当初 15 年と 30 億ドルかかると見込まれていた計画は、2 年前倒しで完了し、予算も 3 億ドル節約された。改良された手法と装置により、ゲノムの大半は最後の 2 年間で解読された。

人間に当てはまることは、AI にも当てはまる。知的な推論者は、遅い実験が主観的に何年もかけて終わるのをぼんやり待つ必要はない。超人的推論者は代替経路を検討し、それを見つけて出すことに長けている。それが知性というものだ。

実例として、ソフトウェア開発と宇宙探査機を比較してみよう。ソフトウェアは変更が安価で迅速なため、エンジニアは頻繁に試行錯誤し、不完全な製品を出しては修正する。一方、宇宙探査機の実験は極めて高価であるため、人間は打ち上げ前に可能な限り正確に設計し、多くの実験機能を詰め込む。遠隔操作できる汎用実験装置を搭載し、新しい実験を思いついても新たな探査機を打ち上げずに済むようにする。

さらに十分に賢い推論者であれば、そもそもそれほど多くの実験をせずに現実の構造を解明できる場合もある。十分なデータが既にあるなら、それを正しく解釈できる知性さえあればよい。

事例としてアインシュタインを挙げよう。一般相対性理論が新たな観測データで検証されるまでには 8 年かかった。1919 年 5 月 29 日の皆既日食で、フランク・ワトソン・ダイソンとアーサー・エディントンが太陽背後の恒星を撮影し、光の曲がりを測定した結果、理論と一致した。

しかしこの 8 年間は科学の進歩を止めなかった。



理由の一つは、理論がすでに水星の近日点移動などの既存データと整合していたことだ。ニュートン理論では説明できなかった現象を、アインシュタイン理論は正確に予測した。人間の科学界は「新しい観測のみが証拠になる」としたが、それは後知恵バイアスや確認バイアス、不正行為を防ぐための制度的な工夫にすぎない。優れた推論には必須ではない。実際、注意深い思考者たちはエディントンの実験以前に理論の正しさを判断できていた。

また、より速い検証方法もあった。ブラックホールの効果を観測する望遠鏡を建設することや、宇宙飛行能力があれば衛星の時計を1日以内に検証することも可能だった。一般相対性理論の検証に8年必要だったと考えるのは、知性の力を過小評価することになる。

GPS衛星が建造されたとき、理論を使う時計と使わない時計の二種類が搭載された。理論が十分に確認されていたにもかかわらずである。これは、不確実な理論に対して両方の分岐を同時に採用できること、そして実験や失敗が高価な場合には特定の理論に強く依存しない設計の方が安価であることを示している。

また、ニュートンやケプラーやティコ・ブラーエと比較したアインシュタインの例は、限られた観測からでも賢い人間がどれほど多くを推論できるかを示している。彼の偉大さは理論そのもののだけでなく、わずかなデータから導いた点にもある。

したがって、実験データの必要性がAIの行動速度をある程度制約することはあり得るが、その制約は直感よりはるかに弱い可能性が高い。

〈より深い議論〉

ナノテクノロジーとタンパク質合成

人間の知能は、他の種に対して多くの利点をもたらしてきた。しかし中でも特に結果が大きかったものの一つは、新しい技術を発明する能力である。開発者が競争して人間より賢い AI をつくるなら、AI の力の大部分が、科学と技術のフロンティアを前進させる能力から来ることも同様に予想できる。だが、具体的にはそれはどんな姿をしているのか。まだ発明されていない、発見を待つ技術とは何か。

これは一般的に答えるのが難しい問いである。1850 年の科学者が、その後 100 年の多くの発明を当てるのは非常に難いだろう。

しかし、完全に無力というわけではない。技術が、技術的に推論できるのに、エンジニアがまだすべての要素を揃えられていない、という場合には、科学者は何十年も何百年も前に多くの発明を予見してきた。

私たちが AI が探究しそうだと思える、より影響の大きい技術フロンティアの一つは、極めて小さい道具や機械の開発である。以下ではこの話題について、そしてその背後にある基本的推論について、いくらか詳しく述べる。

生物学の例

自然界のあらゆる生物のあらゆる細胞には、膨大な種類の精緻な機械装置が含まれている。

ここでいう「機械装置」は比喩ではない。対象の機械は小さいので、日常生活の機械とは少し違ったふうに動く。しかし大規模な機械の多くには、私たちの体内に対応物がある。ATP 合成酵素は、プロトンの流れを使って文字通りの回転子を回すことで、体内で水車に似たやり方で動力を生み出す。

細菌の鞭毛は、ボートのスクリューに似た働きをする。鞭毛を回転させて細菌を液体中で推進させるために、鞭毛を回す完全に機能するモーター全体を備えている。

もう一つの例として、本書で触れたキネシンがある。キネシンは貨物ロボットのように機能する小さなタンパク質である。キネシンはニューロンを横切って走る自己組み立て繊維に沿って「歩き」、神経伝達物質を目的地まで運ぶ。

機械は小さくなるほど一般に速く動作でき、分子ほど小さい機械は非常に速く動作する。キネシンは 1 秒あたり最大 200 歩を踏み、片方の「足」を前に出しつつ、もう片方の足がそれが乗っている微小管にしっかり固定されている。

人間以上に賢い AI が探究する技術フロンティアの一つは、この極小スケールで機械をつくる、設計する、あるいは転用することである。この種の技術は、スケールや、設計が

生物学の既存構造にどれだけ近いのか、そしてそれが「湿った」もの（水に依存する、生きた細胞内の機械装置のようなもの）か「乾いた」もの（開放空気中で動作できるもの）かといった要因に応じて、「バイオテクノロジー」「ナノテクノロジー」あるいはその中間の何かに分類されるかもしれない。

生物をナノスケール工学の驚異とみなすことは、人間が今日持っているものを超える科学技術で、人間以上に賢い AI が何を達成できそうかについての推測を助ける。

（そのような技術を発明し成熟させるのにどれだけかかるか、という別の問いがある。その話題については、本書第 1 章での「機械超知能は既存のコンピュータハードウェア上で人間より少なくとも 1 万倍速く考えられるだろう」という議論を参照せよ。また、「AI は物理的試験や実験を走らせるためにいくらか時間を使う必要があるが、全体としての減速はおそらく超知能にとって大した妨げではない」という〈より深い議論〉も参照せよ。）

今日の人間のエンジニアの偉業を見ると、例えばバイオラボを運営する超人的に有能な AI が、太陽光を使って自分自身を何度も複製する微小工場をつくれる、などというのは信じがたいように思えるかもしれない。利用可能な資源から、ほとんどどんな機械でもつくろう指示を受け取れる汎用マイクロ工場——そういう工場を想像するのはさらに荒唐無稽に見えるかもしれない。

しかし、そのような機械は単に可能だけではなく、すでに存在している。藻類は、幅がミクロンで、太陽光で動き、自己複製する工場であり、1 日未満で個体数を倍増できる。そして藻類はリボソームを含む。リボソームは、生物学版の万能 3D プリンタ、あるいは万能工場の組立ラインである（少なくとも生命の構成要素に関しては万能である）。

適切な命令セット（メッセンジャー RNA に符号化されている）さえ与えれば、リボソームは、タンパク質から組み立て可能な任意の構造を印字する。この普遍性は、生物世界の巨大な複雑さと多様性を支えている。地球上の生命の多様性は最終的に、これらの万能工場によって組み立てられており、ヤマアラシからショウジョウバエから細菌に至るまで、ほぼ変わらない形で見いだされる。

リボソームは、タンパク質を中介として用いることで、タンパク質そのものではない構造を組み立てるためにも使える。このやり方でリボソームがつくれる非タンパク質構造の例が骨である。リボソームは、タンパク質を産生し、それが弱く結合した酵素へと折り畳まれ、カルシウムとリンを特別な反応物へと触媒する。これらの反応物はその後、コラーゲンのマトリクスを形成し、このマトリクスがカルシウムとリンを所定の位置へ導いて、硬い結晶質の骨へと変える。

自然は、本当に並外れた物理機械が可能であることの存在証明を与える。人間がまだしていないやり方でリボソームを使えるほど賢い存在——あるいはリボソームを使って自分自身の改良版のリボソーム類似物をつくる存在——にとっては、そうした機械が可能である。

しかし、生物世界に見える構造は、可能なことの下限を設定するだけである。生物はエネ

ルギー効率や材料強度の理論限界に近いところには全くおらず、人間よりはるかに賢い推論者にとっては比較的容易に改良できるかもしれない。

微小な世界にはまだまだ空きがある

自然現象を、将来の技術が実行可能であることの証拠として使うのは奇妙に思えるかもしれないが、これは科学史でよくあるパターンである。鳥が飛べたので、発明家は何世紀にもわたり飛行機械をつくろうとした。

先駆的な物理学者リチャード・ファインマンは、1959年の講演「微小な世界にはまだまだ空きがある (There's Plenty of Room at the Bottom)」で、このアプローチの力を示した。その講演でファインマンは、微小化によってどのような興味深いことができるかを計算する。

今日から見ると、ファインマンの観察は驚くほど先見的である。ファインマンは、コンピュータはおそらく、もっと多くの要素を含めばもっと多くのことができるだろうが、障害は、その場合コンピュータが巨大になる必要があることだ、と述べる。だから微小化しなければならない！

ファインマンは、人類が書いたすべての本を保存するには、およそ1ペタビット (1,000,000,000,000,000 ビット) が必要だと計算する。

各ビットに100個の原子を割り当てる。そして、世界中の本の中に人間が慎重に蓄積してきた情報すべてが、この形で、幅が200分の1インチの物質の立方体書き込めることが分かる——これは人間の目が見分けられる最小の塵の一片である。だから微小な世界にはまだまだ空きがある！ マイクロフィルムの話なんかするな！

今日でさえ、私たちはそれを完全には達成していない！ 2テラバイトの microSD カードの内部の実際の記憶素子は、まだ一辺 0.6 ミリメートルである。参考までに、200 分の 1 インチは一辺 0.125 ミリメートルである。そしてその SD カードが保持するのは 176 兆ビットにすぎず、ファインマンが 1959 年に人類の知識を保存するのに必要だと計算した量の 57 分の 1 にすぎない。

もしかするとファインマンは、実用的意味での工学の究極限界について間違っていたのだろうか。計算の微小化のさらなる進歩は、近年かなり減速している。何かが物理的に可能だと言うことは、エンジニアがそれを実現できることの証明にはならない。

そして、ファインマンが達成されることになるものの 3 桁以内まで迫ったことは、かなりの予測的偉業と見なせるかもしれない。ファインマンの講演は、ゴードン・ムーアが後に「ムーアの法則」と呼ばれる考えを初めて提示する 6 年前であった。人々は、微小化をグラフ上の不可避な法則として考えることに慣れていなかった。私たちは、ファインマン

の時代に、砂粒サイズの記憶素子が、1950年代最大の真空管コンピュータの1000万倍の情報を保持しうる装置が将来存在するかもしれない、と推測した他者を知らない。

しかし実際には、ファインマンは間違っていなかった。そしてファインマン自身も当時すでに、自分の見積もりが安全側であることを知っていた。

この事実——膨大な情報量が極めて小さい空間に運べるという事実——は、もちろん生物学者にはよく知られている。[...] この情報はすべて、細胞のごく小さな部分に、長鎖DNA分子の形で含まれており、細胞についての1ビットの情報あたりおよそ50個の原子が使われている。

現代のコンピュータはまだDNAのスケールまで微小化されていないが、60年で私たちは驚くほど近づいた。高級商用チップのトランジスタゲートは今や100原子未満の幅であり、原子1個分の厚さの材料層を追加できる技術でつくられている。

自然の類比に錨を下ろし、概算の物理計算を行うことは、今後数十年に何が達成されるかを導く、独特に強い手がかりであることが分かった。そして、必要な科学と工学の仕事をAIがしているなら、この種の技術軌道はもっと速く進むう。

生物を超える

なぜ肉は鋼ほど強くなれないのか。

結局のところ、それらは同じ原子でできている。鉄原子間の金属結合は硬いが、ダイヤモンドの炭素原子間の共有結合も硬い。では、なぜ私たちは、皮膚の中を走るダイヤモンドの鎖帷子を進化させて、生殖年齢まで生き残る助けにしなかったのか。

それどころか、鉄がそんなに強いなら、なぜ生物は鉄鉱石を食べて鉄板の皮膚を育てるように進化しなかったのか。人間のエンジニアがそれをできるのなら、なぜ自然が先にやらなかったのか。

状況的な理由で、鉄板の皮膚が特に良い考えではないのかもしれない。

だが、もしそうでないなら、なぜ他の何かではないのか。

ここでの大きな包括的問いはこうである。なぜ自然は、物理学から計算される（あるいは人間の工学によって示される）物理的可能性の境界から遠いのか。狭く浅い答えではなく、深く一般的な答えはあるのか。

私たちは、ファインマンが生物学の構造を用いて、より大きな科学的知識があれば何が可能なのはずかの下限を設定できたことを述べた。しかし多くの場合、人間の技術はすでに生物学を上回っている。進化が植物や動物を改良するのに何十億年もあったのに、なぜそれが可能なのか。この一般現象を理解することは、ナノテクノロジーが、今日自然に見えるものをはるかに超えて進める可能性が高い理由を照らす助けになる。

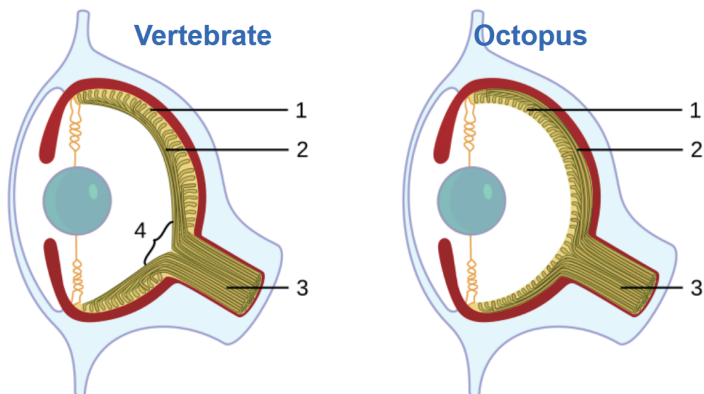
私たちは、セコイアが最も高い建物の少なくとも半分の高さにまで育つ世界を想像できる。私たちは、最も頑丈な動物の皮膚が、観測される最も硬い材料の少なくとも半分は硬い世界を想像できる。なぜ私たちは、何十億年もの進化の後に自然が物理限界へ自らを押し付けたような世界にはいないのか。

これは十分に深い問いなので、知られていることのすべてを短く要約することはできない。しかし大雑把な要約は、自然選択は設計空間のいくつかの部分にアクセスするのが難しい、ということであり、その中には人間のエンジニアにとってはずっと到達しやすい部分が多く含まれる。

私たちがこれに寄与していると見る主要な要因は3つある。

1. 自然選択が使える選択圧には限りがあり、新しい突然変異を普遍化させるのに何百世代も必要である。生物学的特徴が非常に古いものでないなら、その設計はしばしば時間制約の影響を受け、急いでつくられたように見える。
2. 自然選択がつくるものはすべて、以前の設計における偶発的エラー——突然変異——として始まった。進化は、現在生物に存在するものから遠い設計空間の部分を探索するのが難しい。進化がギャップを飛び越えるのは難しい。
3. 自然選択は、逐次的変化ではなく同時的变化を必要とする新しいものをつくることや問題を直すことが難しい。これは、進化がアクセスできる設計を鋭く制限し、生物の現在の設計が、人間の工学基準では、まだらで、ハックだらけで、極めて絡み合った見た目を持つ理由になっている。例えば、ロシユ生化学経路ウォールチャートを見れば、人間の代謝（分かっている部分）の複雑さを感じ取れる。

あるいは、進化の散らかり方より簡単な例として「目」を考えよ。脊椎動物の目はたまたま、神経（下図の2）が光検出細胞（1）の上に乗る形で進化した。これらの神経は目の後ろの穴（3）から外へ出なければならず、この場所は穴である以上、光検出細胞を欠かなければならない。これが脊椎動物（人間を含む）すべてに盲点（4）をつくり、脳が穴を「埋める」巧妙な工夫（例えば、もう一方の目からの情報で）をしなければならなくなる。

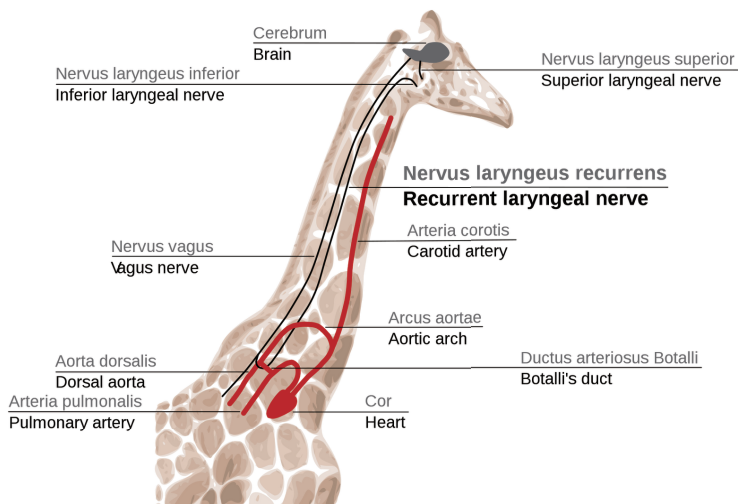


不便な側に神経がある脊椎動物の目と、便利な側に神経があるタコの目。

タコは独立に目を進化させ、偶然より合理的な設計を進化させた——神経は光検出細胞の後ろに回る。これによりケーブルは盲点を一切つくらずに目から出られる。

あるいはキリンの反回喉頭神経を考えよ。これは喉頭を動かすために、キリンの喉と脳を接続しなければならない。直接の経路を取る代わりに、この神経は喉から首の全長をずっと下り、気まづく大動脈の周りを一周し、また首をずっと上って出発点へ戻り、それから脳へ接続する。

その結果、神経は15フィートの長さ（次の図の黒いループ）になり、信号がキリンの脳と喉の間を移動するのに必要な時間が、本来必要な時間の10～20倍になる。



キリンの反回喉頭神経。喉から大動脈（心臓から出る主動脈）を回って脳へ戻る。

魚ではこの設計は理にかなっていた。魚版の喉頭神経は、脳と鰓をつなぐ一直線だった。しかし同じ設計のまま動物に首を与え、配線を一からやり直すことなく首だけを長くし続けると、非常に非効率な設計が生じる。生存はできるが、非効率である。

進化は十分な時間があれば驚異的な設計を生む。しかし人間と AI は、自然ができないことをする何十万もの機械を短時間で立ち上げる点で、はるかに多様で柔軟な設計範囲を考案できる。私たちは、AI が設計で人間より優れるようになり、同じ認知作業を何十万倍も速く行えるなら、これはさらに強く当てはまると予想する。

自然選択は突然変異と組換えを通じて動作するので、より良いキリンを「設計」する能力は阻害される。単一突然変異の連鎖によって到達できない設計空間の部分にアクセスするのが難しい。そこでは各ステップが個別に有利でなければならない。あるいは、すべてが個別に十分有利で、組み合わせる前に遺伝子プールの大きな割合に存在していた突然変異同士を組み合わせる必要がある。

5つの遺伝子からなる遺伝子複合体で、それぞれが独立に集団内で10%の頻度しかないなら、その複合体が各個体の中で組み上がる確率は10万分の1しかない。そして、巨大な利得を持つが、10万回に1回しか起きない遺伝子複合体は、固定化へ進化する可能性がほとんどない。

これは、自然選択が複雑な機械をつくれないう意味ではない。ただ、複雑な機械への道が、段階的に有利なステップを通らなければならない、という意味である。キリンの神経を付け替えるには、キリンのゲノムにいくつかの同時変更が必要であり、それらの変更は他の変更なしでは個別には役に立たない。だからキリンの解剖は今のままである。

進化の驚異は、どれほど速く働くかではない。そのサンプル複雑性は、事例研究を行う人間のエンジニアよりはるかに高い。進化の驚異は設計の優雅な単純さではない。どんな生化学過程の経路図でも一目見れば、その誤解は治る。進化の驚異は、うまくいかない可能性があるすべての経路を覆う頑健な誤り訂正でもない。飢餓や外傷で死ぬことが減った今、現代医学の大半は、外傷がなくてもランダムに爆発する人間生物学の断片を治療している。

進化の驚異は、純粹に偶発的な探索過程であるにもかかわらず、進化がそもそも働く、ということにある。

タンパク質の弱さ

これは、技術が生物を改良できそうな別の方向へ私たちを導く。

肉のレベルよりはるか下、肉眼には見えないところに細胞がある。細胞のレベルよりはるか下にタンパク質がある。

タンパク質は折り畳まれるとき、主として分子版の静電気——ファンデルワールスカーンによってまとまっている。これは鉄のような金属結合、あるいはダイヤモンドのような共有結合より、何十倍も何百倍も弱い。

なぜ生物は、基本的構成要素としてこんなに弱い材料を使うのか。より強い材料は進化が扱いにくかったからである（そして、進化を難しくしすぎれば、そういう問いを問う人々を進化させることは決してない）。

タンパク質は比較的弱い分子力で折り畳まれ、主として静電的な粘着でその形に束縛される。これは、自然選択が探索できる可能性の「豊かな近傍構造」を持つ主要な理由である。ランダムな突然変異はタンパク質を繰り返し微調整でき、その結果として、ほぼ同じことをするが少し良い、新しい設計へ偶然たどり着ける。

もし生物が、強い結合でがっちり結ばれた分子でできていたら、構成要素の一つを変えることは、面白いほど違う（そして潜在的に有用な）新しい構造を生む可能性が低くなる。もちろん時には起こりうる！　だが起こる頻度は大幅に下がる。そして、細胞コロニーを発明するのに20億年かかり、分化した細胞型を発明するのにさらに10億年かかるタイプの設計者にとって、「起こりにくい」とは、そこへ到達する前に最寄りの恒星が膨張して惑星を飲み込む、という意味になる。

あらゆるタンパク質は、ある先行タンパク質のコピーエラーがあったから存在している。その先行タンパク質が、多くの強い結合でぎゅっと固定されていなかったのは、そこから

進化するのが難しくなるからである。だから最新の新しいタンパク質も、おそらく多くの強い結合を持たない。

生化学は時に強い結合を見つける。私たちは以前、骨の例を挙げた。植物にも例がある。植物は、酵素へ折り畳まれるタンパク質を進化させ、それが分子の構成要素の合成を触媒し、それが酸化されて、強く共有結合で架橋されたポリマー——リグニン、木材の構成要素——になる。

しかしそれらは特例であり、自然選択はそのような例を大量に工学的につくり込むために使える「注意力」をあまり持っていない。

炭素原子や他の一般的有機元素の性質からして、それらが強くならないわけではない。ただ、進化させるのにもっと多くの手間がかかる。自然選択には、あらゆる箇所でする時間がない。骨のように、木のリグニンのように、爪や鉤爪のケラチンのように、解剖の他の部分に継ぎ当てられた、いくつかの稀な特例のためだけである。

適切なキーワードを使えば、例えば ChatGPT-o1 に質問できる。あなたがこれを読む頃には、同程度の強さの LLM はおそらく無料になっているだろう。ダイヤモンドの炭素—炭素結合、普通の鉄金属の鉄—鉄結合、リグニンの共有結合性ポリマー結合、ケラチンのジスルフィド結合、骨のイオン結合、といった個々の結合強度について尋ねられる。それらがより大きい材料の構造強度とどう関係するかも尋ねられる（2023 年にこれをやるべきではなかった。GPT-4 は計算を全部間違えるだろう。しかし私たちがこの段落を書いている 2024 年時点では、o1 はより良いように見える）。

すると、炭素原子同士の結合強度は概ね 0.5 アトジュール程度であり、鉄原子同士の結合強度も同程度であり、ケラチンの硫黄—硫黄架橋はそれよりわずかに小さい（0.4 アトジュール）だけであり、木材中リグニンの重合共有結合も同様だ、と学ぶだろう。

しかしタンパク質を折り畳む静電氣的粘着力は、見方にもよるが、せいぜい 10 倍弱い程度であり、場合によっては何百倍何千倍も弱い。

そして植物がリグニンのような物質を触媒するときでさえ、そこでの架橋は、ダイヤモンドの炭素—炭素結合より疎である傾向がある。ダイヤモンドのギガパスカル級の強度と、木材のメガパスカル級の強度の違いは、結合が個別に強いかどうかというより、ダイヤモンドにおける結合の密度と規則性の違いに由来する。

設計者としての進化の限界と、構造材料としてのタンパク質の限界のために、生命は、人間の設計者や AI が回避できる制約の下で動作している。鳥は工学の驚異だが、人造の飛行機械は、最速で最強の鳥より 10 倍以上の速度で、1 万倍重い貨物を運べる。生体ニューロンは工学の驚異だが、人造トランジスタは最速のニューロンより何千万倍も速くオンオフする。そして今日の技術は、達成可能なことの表面をかすっているにすぎない。

フレイタスと赤血球

私たちは、生物学が物理的に可能なことの限界付近には全然いないと言った。では限界付近にあるのは何か。

この問いについての良い考え方を示すために、赤血球を考えることができる。

過去 15 億年のあいだ、人間からトカゲまで、多細胞生命における酸素運搬はヘモグロビンによって担われてきた。ヘモグロビンは 574 個のアミノ酸からなるタンパク質で、特別につくられた 4 つのヘム基（特別な鉄分子を保持する）を持つ。人間の赤血球は約 2 億 8 千万個のヘモグロビン分子を含み、長さは約 7 ミクロンである。300 万個を針の頭に載せられ、あなたの体には約 30 兆個ある。

酸素運搬に関して、赤血球は原理的にできることの限界にどれだけ近いのか。

『Nanomedicine』の著者ロブ・フレイタスは 1998 年に、共有結合材料を用いた人工赤血球の理論設計を中程度に詳細に検討した。その細胞は、詰まった動脈を通りやすくするため直径 1 ミクロンに設計された。

フレイタスは単に別の方法で酸素分子を貯蔵することを考えただけではない。赤血球全体を置き換える方法を考えた。彼は先行分析を踏まえ、血液媒質からグルコースを取り出し、そのグルコースをエネルギーへ変換して人工細胞を動かす必要も検討した。固い棒が別の固い棒にカチカチとはまり込んで簡単な計算をするような、固体ロッド製の微小オンボード計算機や、細胞サイズのセンサーも検討した。人工細胞が液体中で現行の赤血球より速く沈降してしまわないかも検討した。

生体適合性は人体内に入るものにとって大問題になりうるが、ダイヤモンド表面は十分に不活性であり、ダイヤモンド状薄膜コーティングは人体内に入る医療機器の一部で用いられている。フレイタスが考えていた理論可能性のレベルでは、人工細胞の表面はダイヤモンドのようにでき、したがって生体適合だと言ってしまえる。

人工赤血球の中心は、ミクロン規模の単結晶コランダムまたはダイヤモンドの圧力容器が、保守的に見積もって 10 万気圧の圧力に耐えるだろうというフレイタスの計算であった。快適な 100 倍の安全率を取り、分子を 1000 気圧でしか詰めないとしても、人工赤血球は体積あたりで赤血球より 236 倍多くの酸素を組織へ供給でき、呼吸の反対側をバッファするために同程度の二酸化炭素も貯蔵できる。大雑把に言えば、息を 4 時間止められる。

さて、実際にそのような人工血球をつくることは、まったく別の話である。だからこの医療処置は、あなたの近所の医者で今すぐ受けられるわけではない。

1 キログラムの完全無欠な固体ダイヤモンド球は、紙の上では簡単に記述できる分子だが、合成は難しい。フレイタスが私たちに助けてくれるのは、この領域で現在の生物学が理論限界からどれほど離れているかについて、より情報に基づいた推測をすることである。

Ⅱ 生物学は印象的だが、最適からは遠い。

多くの理由でフレイタスの正確な設計が動かない可能性はあるし、最適でない可能性も高い。極めて新規で複雑な設計の最初の案は、どこかで問題にぶつかるのがほぼ確実である。

しかし、フレイタスの正確な提案が動かないだろうという懷疑を表明することは、生物学的赤血球より何百倍も効率よく酸素を運ぶ代替が決して実現できないと言うことではない。

工学とは、何かを動かす方法を見つけることである。たとえ千の道が失敗しても、成功が一つあれば事業全体は成功する。17世紀やそれ以前に無数の実用にならない航空機設計があったことは、機能する飛行機が不可能だったことを意味しない——ただ、可能な設計空間の中で見つけにくっただけである。

このため、技術懐疑論者は、技術が最も目の輝いた楽観論者が信じるより未来にある、という点ではしばしば正しい一方で、ある技術的偉業が決して達成されない、という主張では誤る傾向があった。偉業が世界内の具体的タスクであり、偉業がどう達成されるかについて私たちが不可知であり、そして偉業が物理法則によって許されていることが分かっているなら、歴史は、道筋が最初は明らかでなくても成功する方法がしばしばあることを示唆する。

あるいは、作家で発明家のアーサー・C・クラークの言葉を借りれば：高名だが高齢の科学者が「それは可能だ」と言うとき、彼はほぼ確実に正しい。「それは不可能だ」と言うとき、彼はたいていの場合おそらく間違っている。

ナノシステム

まとめると、

- ・生物世界は、信じられないほど多様な分子機械からできている。
- ・生物学を見ることは、微視的な偉業として何が技術的に可能かを教えてくれる。
- ・しかし生物学は可能性の保守的下限であり、可能性の限界付近にはない。進化は非常に限られた設計者であり、タンパク質は最良の構造材料ではない。

エリック・ドレクスラーの『Nanosystems』（1992）は、小スケール工学でどんな偉業が可能かという問いを探究する古典である。『Nanosystems』は1990年代のナノ材料革命のきっかけになり、科学者たちがドレクスラーの議論を巡って議論したことでかなりの論争も引き起こした。『Nanosystems』の完全なオンライン版はここで見つかる。

『Nanosystems』は深く広範なテキストであり、技術的主题にもかかわらず驚くほど読みやすい。重要な貢献の一つは、小スケール構造を新しいやり方で構築する含意を探ったことである。

非常に小さいものをつくる一つの方法は化学反応である。特定条件（極端な熱など）の下で分子を衝突させ、分子を壊し、原子が新しい分子へ結合するようにする。

これはそれ自体強力なアプローチであり、人類がプラスチックや鋼やセラミックスなど

をつくる方法でもあるが、他の方法でつくれるものに比べるとかすむ。化学反応で材料をつくるのは、レゴの袋をつくってそれを強く振り、構造をつくるのに少し似ている。そうやってつくれるものもあるが、つくれるものの集合は限られ、無駄が多い。

タンパク質合成は、より小さい、事前に組み立てられたレゴブロックのセットから、大きな構造を手で組み立てるようなものである。事前に組み立てられたセットを望む場所へ正確に置けるので精度の余地は大きい、それでも事前に組み立てられたセットを扱っているので少し変で不格好である。これが体内でリボソームがしていることである。アミノ酸鎖をつないでタンパク質をつくり、そのタンパク質が体内でさまざまな仕事をする。

インスリン、ヘモグロビン、ATP 合成酵素はすべて、複数のタンパク質鎖がくっついたタンパク質複合体の例である。インスリンは2本、ヘモグロビンは4本、ATP 合成酵素は29本である。

タンパク質の構成要素であるアミノ酸は分子であり、通常は10～25個ほどの原子からなる。構造材料として、アミノ酸には多くの利点がある：

各アミノ酸は、別のアミノ酸の骨格へ付く骨格を持ち、そこには炭素・水素・酸素・窒素・硫黄原子からなる（潜在的に長い）側鎖が付く。何百もの異なる側鎖が可能で、それぞれ異なる振る舞いをし、アミノ酸を非常に柔軟な道具にする。

アミノ酸の骨格はレゴ片のように別のアミノ酸の骨格に結合できる。これは繰り返し可能で、典型的タンパク質は何百のアミノ酸がつながってできている。これによりアミノ酸は道具（あるいは道具の部品）としてさらに柔軟になる。タンパク質の複雑さは、DNA 突然変異による小さな変更でも、タンパク質が急激に変わって完全に無用になることなく済むことが多い、ということも意味する。これは新しいタンパク質が進化しやすいことに寄与する。

タンパク質はアミノ酸の線形鎖でできているので、アミノ酸を順に列挙するだけでタンパク質を一意に指定できる。DNA はこれを利用し、4文字（ヌクレオチド）の「アルファベット」を使って3文字の「単語」（コドン。各単語が異なるアミノ酸を表す）をつくり、それを線形の「文」（その正確な配列のアミノ酸からなるタンパク質）へとつなぐ。

ミラー＝ユリー実験が示すように、アミノ酸は生命なしでも、単純な化学反応から自発的に形成されうる。これは生命（そしてリボソームやタンパク質合成の前駆体）が最初に発達する道をつくる。

体は、必要な20種類ほどのアミノ酸を食物から得たり、体内で合成したり、過去のタンパク質から回収したりする。リボソームはDNA から指示を受け取り、その指示は本質的に「このアミノ酸、次にこの別のアミノ酸、次にこの別のアミノ酸……、止め」というものである。アミノ酸はその後、転移RNA と呼ばれる小さな分子機械によってリボソームへ運ばれ、リボソームがタンパク質を一つずつ組み立てる。

重要なのは、上のリストが、進化にとっては非常に価値が高い特徴だが、意図的工学にとってははるかに不要であるという点である。進化には、一般的化学反応でつくれ、単純

だが柔軟な化学構造が必要である。人間や人工の設計者は、それらが互いに密接に関連している必要がないので、無関係な多様な分子から選べる。自然界にほとんど生じない構成要素を使う自由もあり、これらの構成要素を複雑なトップダウンの方法で組み立てる自由もある。

これが、非常に小さいものをつくる第三の方法を探る動機の一部になる。すなわちメカノシンセシスであり、構造を、原子を正しい位置へ直接動かして構築する。これは、命令を取り込み、タンパク質の違いよりはるかに多様なものを組み立てるリボソーム様の機械を使う可能性がある。レゴの比喻では、メカノシンセシスは、ついに個々のレゴ片を扱えて、各片を望む場所へ正確に置けるようになる、ということに相当する。

『Nanosystems』は、メカノシンセシスでどんな新しい機械が可能かを探究する。ドレクスラーが探る設計例の一つは、約 3500 原子のサイズにまで縮小された遊星歯車である。

ヘモグロビンは約 1 万原子からなり、ドレクスラーの歯車からそれほど遠くない。そしてタンパク質の中にはもっと単純で済むものもある。インスリンはたった 51 個のアミノ酸、つまり合計約 800 原子しかない。

しかしドレクスラーの設計は、体内で見られるより複雑な機械よりは大きくスケールが下である。例えばリボソームや ATP 合成酵素は 10 万原子以上でできており、細菌鞭毛のモーターは 100 万原子を超える。

『Nanosystems』は、それでも技術的に可能な限界を探ろうとはしない。しかし今日比較的解析しやすい事例に焦点を当てることで、メカノシンセシスが、生物世界で今日見えるものをを超える技術を可能にすることを示す。

『Nanosystems』の計算は意図的に保守的である。例えばドレクスラーは、文字通りダイヤの棒が動き回るコンピュータを考える。これは技術の最終限界だからではなく、1992 年当時は電気に基づく計算より解析しやすかったからである。これがフレイトスの血球分析を刺激した。4 年後、エリック・ドレクスラーとラルフ・マークル（暗号学的ハッシュの発明者であり公開鍵暗号の共同発明者としてより知られる）は、可逆計算の可能限界にやや近いシステムを分析しようとし、『Nanosystems』の見積もりより 1 演算あたりの放熱が 1 万分の 1 になると計算した。ただしその新しい見積もりは、より注意深く保守的な分析ではなく、より保守性の弱い分析に基づいていた。

『Nanosystems』の別の箇所には、6 自由度のマニピュレータアームの大まかなスケッチがあり、それには数百万原子が必要だとされていた。後に、この種の機械を原子ごとにスケッチしようとしたところ、必要なのは 2596 原子にすぎないことが分かった。

ドレクスラーが語る規模で原子精度構造をつくるには大きな工学的課題がある。主要な課題の一つは、原子精度構造をつくるには、信じられないほど小さく精密なマニピュレータを使う必要があることだ。しかしリボソームの存在は、攻め口を提供する。

リボソームはタンパク質しかつけれないが、タンパク質は（骨や木のように）アミノ酸そのものではない反応物を触媒し、引きずり回すことができる。リボソームは強力で一般

的な工場であり、リボソームの産物は、より小さくより精密な道具へブートストラップするのに使え。その中には、より強い材料でできた小さな装置を、より直接につくる道具も含まれる。

直接であれ間接であれ、ゲノムが、タンパク質ではないさまざまなものをつくるために個々の原子を操作できる微小アクチュエータを産生できるのは、ほぼ確実に可能である。そして重要なのは、これは自然選択がたまたま転がり込むタイプの機構ではない、という点である。たとえつくるのが比較的容易でも、マニピュレータームは完成するまで有用ではないからである。

進化は、途中の各段階で有用な複雑構造をつくる。知的なエンジニアにとっては利用可能でも、進化には利用できない比較的単純な設計は多い。例えば自由回転する車輪は、途方もなく単純な発明で、応用が非常に広い。それにもかかわらず、自由回転する車輪は、地球上の生命史全体でたった3回しか進化していないように見える。先に述べた ATP 合成酵素と細菌鞭毛、そして独立に進化したように見える古細菌鞭毛である。

その保守的手法にもかかわらず、『Nanosystems』が設定する技術的下限は絶対的には非常に高い。ドレクスラーが記述するような技術を持つ超知能は、微小な自己複製するリボソーム様工場——個体数が毎時倍増する。生物の中にはもっと速く複製するものもあるが、ドレクスラーは保守的に計算した——をつくれ、そしてそれらが群れて、発電所のようなより大きな巨視的構造をつくれるだろう。

ドレクスラーが記述するような『Nanosystems』型のナノシステムは、太陽光と空気を原料に自己複製できるため、非常に速く確実に拡大できる。これが機能する理由は、木が空気中の炭素を剥ぎ取り、それを木材として隔離することで、薄い空気から大量の建材を組み立てられるのと同じである。空気を「空っぽの空間」と考えがちだが、空気中の炭素・水素・酸素・窒素は、固体材料へ再配置でき、さまざまな目的に使える建材である。

タンパク質ではなく鉄やダイヤのような材料でできた自己複製体は、芝刈り機が草を刈り取るのと同じように、生物細胞を食い荒らせる。

それらは安価に、ボツリヌス症を引き起こすタンパク質であるボツリヌス毒素のようなものを合成できる。ボツリヌス毒素の 100 万分の 1 グラム——米粒 1 粒の 2 万分の 1 という小ささ——が致死量である。巧妙に設計された複製体は、少なくとも 1 体がほぼ全ての人間（例えば最後の 1 か月を潜水艦内だけで過ごしたような人間を除く）に吸い込まれた可能性が高くなるまで、開放空气中を目に見えないまま拡散できる。その時点で装置は（タイマーで）同時に微量の毒素を放出し、ほぼ全ての人間を即時かつ同時に殺せる。

あるいは、AI が構築したナノシステムは、地球資源を収穫し再目的化する過程で、付随的に人間を一掃できる。フレイタスの論文は、太陽光だけを動力にし、空気中の水素・炭素・酸素・窒素だけを原料にする微小直径機械が、3 日未満で空を暗くし、同時に生物圏全体を消費するほど速く複製するよう設計できると計算している。†† その結果、もしこの種の技術を最初に達成した AI が、わずか数か月の先行時間を持つだけなら、その

先行時間を使って（人間であれ AI であれ）すべての競争相手を破壊することがもっともらしい。これは、その技術を最初に手にした者に、恒久的で決定的な戦略優位を与える技術である。

ドレクスラー型ナノテクが物理原理として達成可能だと言うことは、初期の人間以上に賢い AI が、実際にその物理限界に近い技術をつくれる、という意味では必ずしもない。私たちの最良推測は、それが人工超知能が解き明かせる範囲にある、というものだ。なぜならこの種の工学課題を解くのは主として認知的挑戦（思考によって解ける）に見え、実験と試験の段階もそれほど長くはないと予想するからである。

たとえこの推測が正しくても、超知能の最初の手が、ナノテクを用いて自分のインフラをつくり、世界資源を掌握することになるという保証はない。もしかすると、目的をさらに速く効率的に達成する技術や手法を開発するかもしれない。

しかしもし人間以上に賢い AI が、鳥に対する飛行機のように、細胞に対して決定的な優位を持つシステムを構築し、地球表面に自分のインフラを増殖させられるなら、AI が最終的に何をするにせよ、それは少なくともそれほど決定的だろう。

この分析の要点は、人間技術が可能性の限界から遠いことを論じることである。人類が解き明かすのに何十年、何百年、あるいは何千年かかりそうな重要技術が幅広く存在し、人工超知能はそれを素早く行えるだろう。

要するに、ナノテクノロジーは、わずかな先行時間を持つ超知能が、おそらく惑星を掌握するための技術的解決策を見つげられることを示す。

超知能をつくる最もありそうな帰結は、少なくともナノテクと同程度に強力な技術を何か見つけ、その後、人類がただ負ける、ということである。

この推測は、私たちが本書で行う議論にとって決定的ではない。世界にナノテクのような「即勝ち」技術が存在しなくても、人類は超知能に負ける。だから本書本編では、この分析を深くは掘らない。

第2部では、私たちは意図的に、AI がリボソーム経由でもメカノシンセシス経由でも、原子精度製造を汎用的にできる、といった仮定を置かない掌握シナリオに焦点を当てる。超知能は未来の支配を勝ち取るのに、完全に圧倒的な技術優位を必要としないので、本書ではその可能性にそれほど焦点を当てない。

しかし、完全に圧倒的な技術優位を持つだろう、という点も指摘する価値があるように思える。

光学的錯視を発見する新しい方法

第6章で私たちは、研究者たちが、人間の視覚処理と視覚皮質についての現代的理解を用いて、新しい錯視を考案してきた、と述べた——それらは、50年前には、起こりそう

もない偶然でもない限りつくれなかったような錯視である。以下に、代表的な例をいくつか挙げる。

「曲率盲」錯視は、曲線盲という一般現象にある程度根拠を持つが、この特定の錯視は、偶然発見されたのではなく、2017年頃に第一原理から注意深く構成されたものである。

2022年、ブルーノ・レングと共同研究者たちは、新しい「拡大するブラックホール」錯視が、参加者の瞳孔を実際に拡大させたこと——暗い空間へ入るのを予期しているかのように拡大したこと——を示す研究を発表した。（この効果は、より暗い視覚ターゲットに単に焦点を合わせるだけでも瞳孔がわずかに拡大する、という効果よりも、目に見えて大きかった。）

「きらめくスターバースト」錯視は2021年に明らかにされたもので、1970年代後半まで遡る輝度と錯視輪郭の研究の上に、注意深く築き上げられた。

「ピンナ=ブレルスタッフ」錯視は2000年頃に開発されたもので、人間の生物学理解に基づいて新しい錯視をつくる、という中心的例としてはやや弱い。それでも、別の角度から興味深く関連がある。というのも、これは新しい技術に基づく錯視であり、すなわち現代のコンピュータなしではつくるのが難しいか不可能だったタイプの錯視だからである。

同様に中心性は低いが、「タイタンの食」錯視は2010年頃につくられ、観察者のM錐体を疲労させ、より疲労していないL錐体が、そうでなければ同時のMおよびLの活性化によって抑えられ弱められていたはずの、鮮烈な青の知覚をつくり出せるようにする。

関連して、2000年代初頭の錐体活性化の研究は、自然界では起こりにくい錐体活性化を注意深く操作することで、さまざまなキメラ色をつくり出すことにつながった。

最後に、進行中のいくつかの実験は次のことを示している。

脳活動のリズミックな波が、目の前で点滅する複雑な画像を私たちが見えるか見えないかを左右する。画像は、それらの脳波の低い点と同時に目の前で点滅すると、実質的に見えなくなりうる。私たちは、その脳波リズムを、ボタンを押すことを選ぶ、といった単純な随意行為でリセットできる。

……これはさらに、生物学と生理学の理解が豊かになるほど、戦略的な操作の範囲が広がることを示している。ここでは、知覚は、視神経への感覚入力をもっとく変えずとも変化しうる。代わりに、刺激の到来時刻を調整して、脳内で起きている他の事柄と同期させることに依存しているだけである。