

第1部 非人間的な知性

1章 人間の特殊能力

これは『超知能 AI をつくれば人類は絶滅する』第1章のオンライン補足資料である。ここで扱われている主なトピックは次のとおりである。

- そもそも「知能」とは何か。
- 機械が人間より賢くなることは本当に可能なのか。
- 知能には実用上の上限があるのではないか。

知能は意味のある概念なのか？

はい。たとえ正確に定義するのが難しくても、そこには実在する現象がある。

過去30年間で、化学のノーベル賞は人間に77回授与され、チンパンジーには0回である。この事実を初めて聞いた宇宙人は、ノーベル賞委員会が偏っているのではないかと疑うかもしれない。しかし実際には、人間にはチンパンジーとは一線を画す何かがある。

あまりにも自明な指摘に思えるかもしれないが、自明な点が重要になることもある。私たちには月面を歩き、地球の運命をチンパンジーではなく人間の手に委ねることを可能にする能力がある。知能の本質について哲学者や科学者が議論することはできるであろう。しかしどのような結論に至ろうとも、その根底にある現象は消えない。人間には、自然界でかつて見られなかった偉業を成し遂げさせる何かがある。そしてその何かは、私たちの脳と、それを使って世界を理解し影響を与える能力に関係している。

正確な定義を与えられないからといって、それが私たちに害を及ぼさないとはいえない。

山火事に巻き込まれたとき、燃焼の化学反応を理解しているかどうかは問題ではない。理解していなくても、燃えてしまう。

知能についても同じである。もし機械が地球表面を自らのインフラへと変え、さらに膨大な廃熱を生み出して海を沸騰させるようになれば、「知能」の正確な定義をまだ持っていないことは大して問題にならない。私たちはやはり死ぬであろう。

私たちはこれを文字通りの意味で言っている。そしてなぜ人間より賢い AI がそのような極端な結果をもたらすと予想するのかについては、これからの章で検討する。第3章では、超知能マシンは目的を追求するだろうと論じる。第4章では、その目的は人間が意図したり望んだりしたものとは一致しないだろうと論じる。第5章では、その目的を達成するためには、人間が生存に使っている資源を奪う方が効率的であると論じる。そして第6章では、彼らが独自のインフラを構築し、世界を急速に居住不能に変えてしまう能力を持つだろうと論じる。

知能を構築するのに、知能の正確な定義は必要ない。

人類は燃焼の化学を理解する前に火を扱えるようになった。同様に、人類はその仕組みを十分に理解していなくても、知的な機械をつくり出す道をすでに進んでいる（第2章で扱う）。

私たちは「知能」を厳密な数学的概念として捉えるよりも、まだ十分に理解していない自然現象に貼られたラベルだと考えることを勧める。

人間の脳の何かが、驚くほど多様な偉業を可能にしている。粒子加速器を建設し、新薬を開発し、農業を発明し、小説を書き、軍事作戦を遂行する——そうしたことを可能にする何かが、人間の脳にはある。ネズミやチンパンジーにはそれが無い。その精神的差異の完全な科学的説明がまだなくても、それに名前を付けることは有用である。

同様に、私たちを上回る知能について語ることも有用である。すでに今日、特定の狭い分野で人間を超える AI が存在する。たとえば現代のチェス AI は、チェスの分野では人間を超えている。そうであれば、科学的発見、技術開発、社会的操作、戦略立案といった分野で人間を超える AI が登場したら何が起ころのかと問うのは自然である。そしてあらゆる分野で人間を凌駕する AI が登場したらどうなるのかと問うのも自然である。

もし将来、世界最高水準の科学研究を人間科学者の何千倍もの速度で行える AI が現れたとき、それが人間とはまったく異なる方法で結論に到達するという理由で、「それは本当の意味で知能ではない」と抗議するかもしれない。それは定義次第では正しい可能性もある。しかし、どのようなラベルを貼ろうとも、その AI が現実世界に与える影響は甚大である。

私たちは、そのような影響について語るための用語を必要としている。世界を予測し、方向づける能力を根本的に備えた機械について語るための用語が必要なのである。本書では、能力を生み出す内部過程ではなく、その能力そのものに「知能」というラベルを与えるという、より単純な道を選ぶ。

「人間レベルの知能」は意味のある概念か？

はい、多くの場合において意味がある。

人類は高度な技術文明を築いたが、チンパンジーはそうしていない。チンパンジーは互いにコミュニケーションを取り、道具を使い、多くの優れた能力を持っている。それでもなお、何らかの意味で「私たちと同じレベルではない」という感覚がある。人間の知能を尺度として使うことには問題もあるが、それでも「このレベル」と指し示すことには一定の有用性がある。

もし将来、宇宙の彼方で異星文明に出会ったとしよう。仮にその文明が私たちと同程度の技術水準にあったとしても、歩行は人間より劣り、泳ぎは優れているかもしれない。チェスやポーカーのような対戦ゲームでは人間より強いが、抽象数学では弱いかもしれない。あるいはその逆かもしれない。思考は遅いが記憶力は優れている、あるいは思考は速いが記憶力は劣っているということもあり得る。

そうした存在を「人間レベルの知能」と呼ぶべきであろうか。（逆に、人間の方を「異星人レベル」と呼んではいけない理由はあるのであろうか。）

ここで私たちが「人間レベルの知能」と言うとき、それは人間に技術文明を築き維持することを可能にしている何らかの性質を指している。チンパンジーにはそれがない。

歴史的（あるいは人類学的）に見ると、人間とチンパンジーが分岐した後のある時点で、何らかの閾値が越えられたように見える。人間に優秀な科学者がいて、チンパンジーには再現性のない論文を書く凡庸な科学者がいるという話ではない。チンパンジーはそもそも科学論文を書かない。まったく書かないのである。人間の脳とチンパンジーの脳は生物学的にはかなり似ている。それでも人間は文明を発明し、鉄を精錬し、ロケットを軌道に送り、読み書きを行えるようになった。

理論を脇に置いて素朴に見れば、何らかの堤防が決壊し、その背後に溜まっていた膨大な知能が一気にあふれ出したかのようである。ある種の「大爆発」が起きたかのように見える。

この考えに巧みに異議を唱える人もいるであろう。しかし彼らができるのは、定義や細かな論点をめぐる議論を持ち出すことだけであって、「実は200万年前のホモ・エレクトスが原子炉をつくろうとしていた証拠を見つけた。ただし非常に下手だったただけだ」と言うことではない。

文明を創造できるほど強力で汎用的な知能は、世界に急激に現れ、ホモ・サピエンスを他の動物から明確に分けた。私たちは「人間レベルの知能」という特定のラベルに固執しているわけではない。この言葉には問題点も多い。しかし何と呼ぶにせよ、あの閾値の「向こう側」にあるものを指す概念を持つことは有用である。

知能は複数のスキルから成り立っているのでは？

はい。ただし、能力のあいだには大きな重なりがある。

たとえば、私が姉よりクラシック音楽の作曲が得意で、姉は私より小説を書くのが得意だとする。音楽と小説執筆は異なる技能なので、どちらが「より賢い」と言えるのかを判断する明確な基準はない。では、AI が人間より「賢い」と言うことに、どのような意味があるのだろうか。

私たちの答えはこうである。もし私が一つのことにおいて優れており、姉が別の一つのことにおいて優れているだけなら、意味のある比較をするのは難しいかもしれない。しかし、私が一つのことにおいて優れていて、姉が2000 ものことに優れているとしたら、私たちが対等だと主張したり、立っている地平について何も言えないと主張したりするのは、いささか不自然である。

『超知能 AI をつくれば人類は絶滅する』は、将来の AI の進歩がもたらすであろう実際的な影響についての本である。その影響を意味ある形で語るために、ChatGPT、人間、ショウジョウバエの三者を正確に比較して、それぞれの「知能レベル」を厳密に定義する必要はない。必要なのは、AI がますます広範な技能において向上していること、そしてやがて極めて実用的に重要な技能において人間を凌駕するようになることを見て取ることである。

知能は過大評価されているのでは？

それは「知能」を過度に狭く定義している場合に限られる。

私たちはときどき、次のような主張に出会う。「成功に必要なのは知能だけではない。最も成功している人間の多くは、カリスマ的な政治家や CEO、ポップスターである。オタクはある種の能力には優れているかもしれないが、世界を動かしてはいない」

私たちはこの主張を否定しない。むしろ、本書でいう「知能」とは、オタクとスポーツマンを分ける性質のことではない。それは、人間とネズミを分ける性質のことである。

ハリウッド映画の脚本で「知的な」人物といえば、たいていは学問的な頭の良さを意味する。歴史に詳しくたり、天才的な発明家だったり、チェスが得意だったり、難事件を解決したりする人物である。

映画の中の「頭のいい人」は、その強みを持つ一方で、典型的な「ハリウッド的オタク」の弱点も併せ持っている。感情的知性や常識、世渡りの才に欠けているかもしれない。手先が不器用だったり、カリスマ性に欠けていたりするかもしれない。

しかしカリスマ性は腎臓が分泌する物質ではない。カリスマ性も、学問的な知力と同様に、脳内のプロセスの産物である。そこには無意識的な脳内プロセスも含まれる。人を惹きつける振る舞いは、必ずしも本人の意識的な制御下にあるとは限らない。それでも最終的には、カリスマ性も工学的な才覚も、オタクとポップスターの間でどのように分配されようかと、人間とネズミを分ける神経学的基盤の一部である。

私たちが「人工知能」と言うとき、それは「人工的な学問的頭脳」のことではない。「人間の脳とネズミの脳を分けるすべてを人工的に実現したもの」を意味する。人間を月に歩かせる力も、演説家が群衆を涙させる力も、兵士が巧みに銃を狙い定める力も——そのすべてを含む、丸ごとのパッケージを指しているのである。

「汎用知能」は意味のある概念か？

はい。

ハヤブサは時速 240 マイル（約 390km）で急降下できる。マッコウクジラは海面下何キロメートルも潜ることができる。ハヤブサは海に入れば溺れ、クジラは空を飛ばそうとすれば地面に叩きつけられるであろう。ところが人間は、自ら設計した金属の殻（飛行機や潜水艦）の中に入り、どちらの生物よりも速く飛び、より深く潜ることに成功した。

私たちの祖先の環境には深海は含まれていなかったし、空を舞う能力が自然選択の対象になっていたわけでもなかった。それでも私たちは、特別な本能によってではなく、頭脳の並外れた汎用性によって、これらや他の多くのことを成し遂げた。

私たちの祖先は、槍をつくる以上に複雑な工学的課題に直面することはほとんどなかった。それにもかかわらず、どういうわけか「問題解決一般」に強い能力を持つように淘汰された。

人間は完全無欠の問題解決能力を持っているであろうか。もちろん違う。たとえば、制限時間内であれば、最高レベルのチェス AI ほど上手にチェスを学び、プレイすることはできない。人間を超えるチェスの実力が可能であることは実証されており、人間は補助なしではその水準に到達できない。私たちの知能は「万能」ではない。物理的に可能なすべてのことを、脳だけで学び実行できるわけではないのである。したがって、人間の持つ「汎用性」とは、あらゆることができるという意味ではない。それでも、ディープブルーのような特化型チェス AI と比べれば、人間の新しい問題を学び解く能力には、はるかに大きな一般性がある。

ただし、汎用性は白黒はっきりした性質ではなく、程度の問題である。

ディープブルーはチェス盤以外を動かす能力に関しては、ほとんど汎用的ではなかった。勝ち手を見つけることはできても、車を運転して店に行き牛乳を買うことも、重力の法則

を発見して月ロケットを設計することもできない。チェッカーのような簡単なゲームも、囲碁のような難しいゲームもプレイできなかった。

一方、囲碁を制した AlphaGo を考えてみるとよい。AlphaGo の背後にあるアルゴリズムは、優れたチェスもプレイできる。囲碁は最初のチェスアルゴリズムでは攻略できなかったが、最初の囲碁アルゴリズムの変種は、チェスでも従来記録を破り、さらに Atari のビデオゲームでも優れた成績を示した。もちろん、それでも牛乳を買いに行くことはできないが、より汎用的になっている。

どうやら、ある種の知能の様式は、他の様式よりもはるかに汎用的である。

しかし私たちは、「知能」以上に「汎用性」をきちんと定義できていない。

人間がショウジョウバエより汎用的だと言うのは簡単である。しかし汎用性とは具体的にどのように働くのであろうか。

分かっていない。「汎用性」についての成熟した形式理論はまだ存在しない。知能がより多様な環境で、より複雑な課題に対して予測し、制御できるほど「より汎用的」であると直観的に言うことはできる。しかし課題や環境を定量化して厳密な定義にする方法は持っていない。

不満に思えるであろうか。私たちも不満である。人類が汎用知能をつくらうとする前に、まず汎用知能についてより深く理解してほしいと強く願っている。それが、第 10 章と第 11 章で述べる厳しい技術的状况を改善する助けになるかもしれない。

形式的な説明はなくても、世界を観察することでいくつかの事実は推測できる。

人間が高層ビルや月ロケットを建設する生得的知識を持って生まれてくるわけではないことは明らかである。私たちの遠い祖先はそれらを扱う必要がなく、遺伝子にその知識が刻み込まれることもなかった。それらの能力は、生まれながらに理解していなかった領域について学ぶ力から生まれる。

汎用性を評価するなら、「どれだけ知っているか」ではなく、「どれだけ学べるか」を問うべきである。

ある意味で、人間はネズミより強力な学習者である。ネズミも迷路を覚えるなど学習はできる。しかし人間は、より複雑で奇妙なことを学び、知識の断片をより効果的に結びつけることができる。

それはどのように働いているのであろうか。私たちにはあって、ネズミにはないものは何であろうか。

引越した後、新しい町で道を覚える二人を考えてみる。

アリスは必要な経路を丸暗記する。家からホームセンターへ行くには、サードストリートを左折し、二つ目の信号を左折し、二ブロック進んで右折して駐車場に入るという具合である。食料品店や職場への道もそれぞれ別に覚える。

一方ベスは、町の地図を頭の中に取り込み、理解する。

アリスは日常生活では困らないかもしれないが、新しい場所へ行かなければならなくなったとき、道順がなければ困ってしまう。ベスはルート計画に時間がかかるかもしれないが、はるかに柔軟である。

アリスは覚えた特定の道では速いかもしれない。しかしベスはその他のあらゆる場所で優位に立つ。さらに、ラッシュアワーに渋滞を避ける経路を見つけたり、別の町の道路設計を考えたりする際にも有利であろう。

学習には、単なる経路の暗記に近いものと、地図を内面化することに近いものがあるように見える。さまざまな状況で再利用・適応できる「頭脳の菌車」が存在するように思われる。深く機能する思考様式があるのである。

このテーマについては第3章でさらに論じる。

知能は単純な一次元の尺度で測れるものか？

いいえ。しかし、AI がまだ到達していない水準は存在する。

「超知能」という考えは、「知能」を単純な一次元の量だと仮定しているのではないかと指摘されることがある。AI 研究を注ぎ込めば、それに比例して「知能」が増える——まるで知能が機械ではなく、地面から汲み上げ続けられる液体であるかのように。

私たちはこの批判の核心部分には同意する。知能は単純な一次元の尺度では測れない。計算資源を増やせば常に簡単に AI が賢くなるわけではない（もっとも、この 10 年を見る限り、そうなる場合もある）。同様に、知能の向上が常にそのままでの向上に直結するとも限らない。世界は複雑であり、能力はボトルネックや停滞に突き当たることがある。

しかし、第1章で述べたように、複雑さや限界やボトルネックが存在するからといって、AI がちょうど人間の能力水準あたりで都合よく壁に突き当たるとは限らない。本文中で論じたように、生物の脳には AI にはない制約がある。

人間の知能にも多くの制限がある。それでも私たちは月に到達した。動物の知能は単一のスカラー量ではないが、それでも人間はチンパンジーを圧倒している。知能がいかに複雑であろうと、人間とチンパンジーのあいだには明確で質的な隔りがある。

人工超知能にも制限や複雑さがあるかもしれない。それでもなお、人間を圧倒する可能性はある。研究者や技術者がますます高性能な AI を目指して競争を続けるなら、彼らと私たちのあいだに質的な隔りが生じる可能性は十分にある。

AI は臨界点を越えて知能爆発を起こすのか？

おそらくそうなる。

現代の AI 進歩は、一見すると漸進的に見えることがある。たとえば、METR という組織は AI がどれほど長時間のタスクを遂行できるかを追跡しており、ここ数年はほぼ指数関数的な曲線を描いている。これを「なだらかな進歩」と見なすこともできる。では、それは進歩が穏やかで予測可能であり続けることを意味するのか。

必ずしもそうではない。

ある量が滑らかに増加しているからといって、その結果が常に穏やかであるとは限らない。核分裂は連続的な物理過程であるが、1 個の中性子あたり 1 個未満の中性子しか生まれない場合は反応が収束し、1 個を超えれば暴走的連鎖反応になる。基礎メカニズムは同じでも、わずかな差が巨大な結果の違いを生む。これが「閾値効果」である。

知能にも少なくとも一つの閾値効果が存在する証拠がある。それが人間とチンパンジーの差である。解剖学的には大きな違いはなく、脳構造も似ている。人間に特別な「月へ行くモジュール」が追加されたわけではない。主な違いは脳容量が約 3～4 倍大きいことと、配線の微細な差異である。

進化の過程では、脳は段階的に少しずつ大きく、少しずつ高度になっただけである。それでも比較的短期間（進化の時間尺度では）に質的飛躍が起きた。

人間で起きたなら、AI でも起き得る。

私たちは、AI がその閾値からどれほど離れているのか分かっていない。

もし、人間がどのようにして汎用知能の閾値を越えたのかを正確に理解していれば、そのような重要な閾値が近づいていることを示す兆候も分かったかもしれない。しかし第 2 章で述べるように、私たちは知能についてそこまで深く理解していない。したがって、どこに閾値があるのか、そこにどれだけ近づいているのかも分からないまま、手探りで進んでいる。

近年の AI の進歩は、数学問題の解決やチェスのプレイ能力の向上として現れているが、それだけでは AI を「完全なもの」にするには至っていない。チンパンジーと人間の脳の差のように、モデルがあと 3～4 倍大きくなるだけで十分なかもしれない。あるいはそうではなく、2017 年に発明され 2022 年に成熟した新しいアーキテクチャ（現代のチャットボットの基盤）と同様に、まったく異なる構造と 10 年規模の科学的進歩が必要なのかもしれない。

人間の脳にどのような変化が起きて、私たちはその臨界点を越えたのだろうか。言語によるコミュニケーション能力だったのかもしれない。あるいは、コミュニケーションを極

めて価値あるものにする抽象概念の理解能力だったのかもしれない。もしかすると、そもそも考え方が間違っていて、現在私たちの視野に入っていないような奇妙な要因が鍵だったのかもしれない。あるいは複数の要因が十分に成熟し、それらが組み合わさることで、人類が月に到達できるような知能が生まれたのかもしれない。

私たちには分からない。そして分からない以上、現代のAIがその臨界点からどれほど近いのか、あるいは遠いのかを判断することもできない。

科学や産業の誕生は人類文明を大きく変えた。言語の誕生もまた、祖先にとって同様に重要だった可能性がある。しかし仮にそうであっても、それらがAIにとって同じような「臨界点」として機能する保証はない。なぜならAIは人間とは異なり、最初からある程度の言語・科学・産業の知識を持っているからである。

あるいは、人類にとっての臨界点は複数の要因の組み合わせであり、それぞれが一定水準に達して初めて全体が機能し始めたのかもしれない。AIは、長期記憶のように人類の方が得意だった能力で遅れをとっていても、最後のピースがはまった瞬間に実用能力が飛躍的に向上する可能性がある。

仮にこうした人間とのアナロジーが当てはまらなかったとしても、AIの進歩を断続的に予測困難なものにする別の力学が存在するだろう。

たとえば、長期記憶や継続的学習の欠如が、これまで人間にはなかった形でAIの進歩を制約しているのかもしれない。そしてそれらが解決された瞬間に、何かが「カチッ」とはまり、AIが一種の知性の「火花」を獲得したように見えるかもしれない。あるいは（本書でも述べたように）、AIがより賢いAIをつくり、そのAIがさらに賢いAIをつくるというフィードバックループが始まる段階が重要かもしれない。フィードバックループは、しばしば閾値的な変化を引き起こす要因である。

私たちの知らないところで、こうした「最後のピース」となりうる要因が十数個存在している可能性もある。あるAI研究所がその最後のピースを解明した瞬間、そのAIは急激に加速し、他を引き離し、人類が他の動物から飛躍したように一気に差をつけるかもしれない。そうした決定的な瞬間は、予想よりもずっと急速に訪れる可能性がある。準備のための時間は、思っているほど多くないかもしれない。

知能爆発の速度は最終結果を左右しないが、速く起き得るのであればその含意は「早く行動せよ」だ。

結局のところ、「誰かが人工超知能をつくれば全員が死ぬ」という主張において、閾値はそれほど重要ではない。私たちの議論は、AIが再帰的自己改善の方法を見つけて前例のない速度で超知能へと進化することを前提としているわけではない。そうしたことが起こる可能性はあり、私たちもそれなりに起こり得ると考えているが、それは「AIが人類を滅ぼす方向にある」という主張の核心ではない。

必要なのはただ、AI がますます世界を予測し、操舵する能力を高めていき、やがて人間を上回るという点である。それが速く起きるか遅く起きるかは、それほど重要ではない。

閾値効果が重要なのは、それが「人類が早急に対応する必要性」を高める点にある。AI があらゆる知的作業で人間をわずかに上回るようになるまで待つ余裕はない。その時点では、残された時間はほとんどないかもしれない。それは、初期の人類が火を扱い始めたのを見てあくびをしながら、「月まで半分くらい来たら起こしてくれ」と言うようなものだ。

人類の祖先が月までの道のりの半分に到達するのには何百万年もかかったが、残りの半分はわずか2日で達成された。閾値が関係している可能性がある場合、状況が明らかに手に負えなくなる前に注意を払う必要がある。そうでなければ、その時にはすでに手遅れになっているかもしれない。

ChatGPT って、もうすでに汎用知能なんじゃないの？

そう呼びたいのであれば、そう呼んでもいいだろう。

ChatGPT やそれに類するものは、それ以前の AI よりもはるかに汎用的である。多少の数学もできるし、詩を書いたりコードを書いたりもできる。これらを常にうまくこなせるわけではないが（2025 年 8 月時点では）、実に多くのことができる。

GPT-5 であっても、その推論能力の汎用性は人間の子どもよりもまだ低いだろう、というのは妥当な推測である。確かに、より多くの教科書の内容を言い当てることはできる。しかし、おそらく人間の子どもが用いるよりもはるかに大量で浅いパターンを記憶している一方で、子どもはより深い「思考の歯車」を使って同様の課題に取り組んでいる（その結果、場合によってはより良く、場合によってはより悪くなる）。

もし私たち著者が両者を比較せざるを得ないとしたら、ChatGPT はある深い意味において人間よりも全体的に「劣っている」と感じるだろう——そしてそれは（2025 年 7 月時点では）チャットボットのエピソード記憶が限られているという理由だけではない。

これに対して、「何を言っているんだ？ ChatGPT は会話もできるし、感情的に深い対話もできる。高度な数学を解いたりコードを書いたりもできるのに、多くの人間にはできない。どうしてそれが人間より劣っていると見えるんだ？」と反論する人も少なくないだろう。10 年前には、そもそもこうした議論自体が成立しなかった。そこから、この間にどれだけの進歩があったかが分かる。

現在の世界は、「AI は明らかに人間より劣っている」という状態と、「何をやらせるかによる」という状態との中間あたりにあるのかもしれない。

残りの差を埋めるために必要なのは、単にもう少しのスケール拡大だけなのかもしれない

い。人間の脳がチンパンジーの脳と大きくは変わらないが、3倍から4倍ほど大きいように。あるいは、ChatGPTの基盤となっているアーキテクチャが、「汎用性の火花」を支えるには浅すぎるのかもしれない。

あるいは、現代のAIアルゴリズムでは扱えない、汎用知能にとって重要な要素が存在し、その代わりに現代のAIは、大量の練習と記憶によって、反復によって解ける種類の課題に対応しているのかもしれない。その場合、その欠点を補うためには、たった一つの画期的な（そして同時に信じられないほど単純な）アルゴリズムの発明があれば十分かもしれない。そうなれば、AIは人間が理解できるほとんどのことを理解し、人間と同程度の効率で経験から学べるようになるだろう（しかもインターネット全体を読み込んで記憶する能力を保ったまま）。あるいは、さらに4つほどのブレイクスルーが必要かもしれない。第2章で述べた通り、誰にも分からない。

「汎用知能」という言葉には、さまざまな意味があり得る。

「AIはすでに汎用知能を持っている」と言う人は、人類文明という形で大変動を引き起こした、あの十分には理解されていない能力の組み合わせを、AIがすでに獲得したという意味で言っているのかもしれない。

あるいは、AIが少なくとも「人間とAIのどちらが本当に賢いのか」を人々が激しく議論する段階にまで到達した、という意味かもしれない。

または、人々が議論するのをやめる段階を思い描いているのかもしれない。それは、AIがあらゆる人間よりも明らかに、そして広範に賢くなったためかもしれないし、あるいは議論する人間そのものがいなくなったためかもしれない——人類が踏み込みすぎ、AIが私たちのあらゆる議論や営みを終わらせてしまった結果として。

AIが「人間レベルのチェスを指し始めた」と明確に言える日時があったわけではない。しかし、チェスAIが人間の世界チャンピオンを圧倒できるようになった時点で、その段階はすでに過ぎ去っていた。

つまりこういうことだ。「ChatGPTは汎用知能なのか？」という問いへの答えは、その問いで何を意味しているのかによって、「はい」にも「いいえ」にもなり得る（そしてこれは、この数年間のAIの進歩について多くを物語っている！ ディープブルーは明らかに非常に特化型だった）。

超知能の方が、より重要な区別である。

「人間レベルの知能」という言葉には複数の解釈があり得るため、私たちは（超人的なAIについて語る場合を除き）この用語をあまり使わない。同様の理由で、「汎用人工知能（AGI）」という言葉も通常は避ける。これらの概念について言及する必要がある場合

には、より具体的に説明するようにする。

代わりに、「人間より賢い AI」「超人的 AI」「超知能」といった、人間を基準とした表現を用いる。

本書における「人間より賢い AI」や「超人的 AI」とは、人間とチンパンジーを分ける「汎用性の火花」を備え、問題解決や真理の探求において、最も優れた人間個人よりも明らかに優れている AI を指す。

超人的 AI は、人間よりほんの少し賢い程度かもしれないが、特定の課題では人間の方が優れている場合もあるだろう。しかし本書では、「人間より賢い AI」とは、さまざまな難しい課題にわたって公平に比較した場合、最も有能な人間よりも総合的に優れていることを意味する。

一方で「超知能 AI」または「人工超知能 (ASI)」とは、人間の知能をはるかに凌駕する超人的 AI を指す。第 6 章で述べる理由から、現実の人間個人や人間の集団は、実用的に重要なあらゆる領域において、超知能 AI と競うことが完全に不可能になると想定する。本書では主に「超人的」と「超知能」をほぼ同義として扱う。ただし第 2 部では、この区別がより重要になる。そこでは、AI が当初は人間よりやや賢い程度でありながらまだ超知能ではない状態から出発する「AI による支配」のシナリオを描く。これは、超知能が過剰である可能性を示すためである。つまり、AI は近いうちに超知能に達するかもしれないが、人類絶滅を引き起こすためにそこまで賢くなる必要はないかもしれない。

これらは非常に大まかな定義だが、本書の目的には十分である。

本書は、複雑な知能理論を提示し、その理論から災害を予兆するような難解な結論を導き出すようなものではない。むしろ、もっと基本的なレベルで議論を進める。例えば次のような主張である。

- ・ある時点で、AI は人間（チンパンジーではなく）がロケットや遠心分離機や都市をつくれるようにしている何らかの能力を、完全に獲得するだろう。
- ・ある時点で、AI は人間を超えるだろう。
- ・強力な AI はおそらく独自の目標を持ち、それを頑固に追求するだろう。なぜなら、目標を粘り強く追求することは多くの課題にとって有用であり（実際、人間もその理由で目標を進化させてきた）、その性質は合理的だからである。

これらの主張が正しいかどうかは、知能の内部構造について特別に深い理解を持っているかどうかには依存しない。トラックの内部構造を詳しく知らなくても、こちらに向かって突進してくるトラックが見えれば危険は分かる——私たちはそう主張する。

そしてこのような単純な議論は、ChatGPT が「本当に」人間レベルかどうか、「本当に」汎用知能かどうかには依存しない。ChatGPT は今できることをやり、将来の AI はそれ以上のことをよりうまくできるようになる。本書の残りでは、その延長線上に何があるのかを論じていく。

超知能はどれほど賢くなり得るのか？

非常に賢くなり得る。

第1章で挙げた「人間の脳が物理的限界にまったく近づいていない理由」の各ポイントについて、機械はその限界近くまで到達できる可能性がある。

物理法則は、人間の何万倍（あるいは何百万倍、何十億倍）もの速度で思考でき、睡眠も食事も必要とせず、自分自身を複製し経験を共有できる「天才」の存在を許している。

しかもこれは、AIの認知の「質」が向上する可能性を考慮する前の話である。

仮にAIが人間より一つか二つの次元で強く優れているだけだとしても、それだけで決定的な優位を得るには十分かもしれない。歴史上、人間の集団は科学・技術・戦略計画における比較的小さな優位を活用して、他の集団に対して支配的地位を築いてきた。たとえばスペインのコンキスタドール（征服者）を思い浮かべるとよい。彼らは脳の構造や大きさが特別に違っていただけではない。

小さな知的優位であっても、大きな実践的優位に転換され得る。そして小さな差であっても、極めて速く複利的に拡大する可能性がある。しかも、AIが持ち得る優位は「小さい」とは到底言えない。

この水準の知能がなぜ重要なのか、それがいかに現実世界の力へと変換され得るのかについては、第6章でさらに詳しく論じる。

超知能に到達するには大きな障害があるのでは？

はっきりとは分らない。

多くの点で、この分野は手探りで進んでいる。もしかすると、もはや本質的な障害は残っておらず、現在の技術を少し改良するだけで超知能に到達できるのかもしれない。あるいは、「少しだけ賢いAIが、さらに少しだけ賢いAIをつくり、そのAIがまた少しだけ賢いAIをつくる」という連鎖を経て、最終的に超知能へ到達するのかもしれない。

仮に重要な障害が存在するとしても、それを人類が克服するのにどれほど時間がかかるのか（AIの助けがあってもなくても）は分からない。

私たちが確実に知っているのは、主要なAI研究所が明確にその方向へ進んでいること、そして実際に進歩しているということである。かつて機械は絵を描けず、話せず、コードも書けなかった。いまはそれができる。

この分野は障害を乗り越えるのが得意

何十年ものあいだ、AI は猫の画像と車の画像を区別することすら苦手であった。転機は 2012 年、トロント大学のアレックス・クリジェフスキー、イリヤ・サツキバー、ジェフリー・ヒントンの AlexNet という畳み込みニューラルネットワークを設計し、当時の最先端を大きく上回ったときに訪れた。この出来事は、ほぼすべての現代 AI を支える人工ニューラルネットワーク中心の「現代 AI 革命」の幕開けと広く見なされている。

AI はボードゲームも苦手であった。1997 年にディープブルーがガルリ・カスパロフを破った後も、囲碁のように手の選択肢が膨大なゲームには苦戦していた。しかし 2016 年、AlphaGo がディープニューラルネットワークと木探索を組み合わせた新アーキテクチャで、世界チャンピオンのイ・セドルを破った。その後 DeepMind は同じアルゴリズムをより汎用化した AlphaZero を開発し、囲碁だけでなくチェスや将棋でも圧倒的な強さを示した。

初期のチャットボットは意思疎通が非常に限定的であった。しかし 2020 年、トランスフォーマー・アーキテクチャが成熟し、GPT-3 が登場した。翻訳、質問応答、さらには本物らしいニュース記事の生成まで可能になった。それを少し再訓練してチャットボットにすると、史上最速で成長する消費者向けアプリケーションとなった。

現代の AI と「本物の」、すなわち超知能になり得る、あるいは超知能を生み出し得る AI とのあいだに障害はあるのだろうか。

あるかもしれない。AlexNet が現代 AI の道を開いたように、AlphaZero が複数ゲームを同一アルゴリズムで攻略可能にしたように、ChatGPT がコンピュータに会話をさせたように、さらなるアーキテクチャの洞察が必要なのかもしれない（あるいは、現代の AI が静かに閾値を越え、それで済むのかもしれない）。

しかし仮に障害が残っていたとしても、研究者たちはおそらくそれを克服するだろう。彼らはそれが得意であり、2012 年当時よりはるかに多くの研究者がこの問題に取り組んでいる。

2025 年 7 月時点では、AI は長期記憶や一貫した計画を必要とする課題、たとえばポケモンのようなビデオゲームのプレイに苦戦している。こうした失敗を見て懐疑論者とともに笑いたくなくなるかもしれない。単純なゲームに苦戦する AI がどうして超知能に近いと言えるのか、と。

だが、同様に 2019 年の AI はまともに会話することに苦勞していたものの、その実現は 20 年先の話ではなかった。研究所は特定の課題で AI がうまくいかない原因を特定しようと懸命に取り組んでおり、長期記憶や計画能力に優れた新しいアーキテクチャを見つける方向へ進んでいる可能性が高い。そうした AI が何を可能にするかは、誰にも分からない。

もし次の段階でも、AI が科学研究や技術開発（より賢い AI の開発を含む）を自動化するのに十分でなければ、研究者たちは次の障害に取り組むであろう。人類が介入して研究

を禁止しない限り——この点は後の章で扱うが——彼らは前進し続けるだろう。

超知能の行動を予測するのは不可能では？

ある点ではそうだが、すべてではない。

Stockfish 17 は、チェス盤を動かす能力において私たちより優れている。もし私たちが Stockfish と対局したなら、その一手一手を正確に予測することはできないであろう。それを行うには、少なくとも Stockfish と同等の棋力が必要である。しかし、試合の勝者を予測するのは容易である。どの手を指すかを予測するのは難しくても、最終的に勝つのがどちらかを予測するのは簡単である。

現実世界を予測し、操舵する AI についても同じことが言える。AI が賢くなればなるほど、具体的に何をするのかを正確に予測するのは難しくなる。しかし同時に、彼らが目指している目的地に到達するであろうと予測するのは、むしろ容易になるのである。

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な欠陥があるのでは？

いいえ。

機械が創造的であり得るかどうかについての本格的な議論は、第3章に譲る。ただし、ここで一つ述べておく。機械が人間と釣り合うための「致命的な欠陥」を持っている必要はない。人間の不屈の精神が最後に逆転する余地が、物語的な必然として存在するわけではないのである。

もしドーデー鳥に映画産業があったなら、モーリシャス島に人間が侵入してきた物語では、人間の銃や鋼鉄は人間側の弱点によって相殺される展開になったかもしれない。たとえば、人間は高度な知能ゆえの実存的不安にとらわれ、最後の瞬間に絶望して立ち尽くす。その隙に勇敢なドーデーたちが反撃し、くちばしで人間を突き殺す——といった具合である。

それはドーデーにとって満足のいく物語であろう。知能が強いくちばしに対して軍事的優位になるはずがない、人間の大きな脳にはきっと致命的な欠陥があるはずだ、と。

しかし現実では、人間の優位は真の優位であった。人間の脳の欠点は、鳥の脳との軍事的対決において純粋な不利にはならなかった。人間とドーデーの戦いは不均衡に終わった。それだけのことである。

人間同士の戦争でさえ、機関銃という技術的優位があれば、通常は機関銃を持つ軍が持たない軍に勝つ。例外はあるし、人々はそうした例外を好んで語る。例外のほうが物語として面白いからである。しかし現実では、例外は物語の中ほど頻繁には起こらない。

私たちは、高度な AI についても同様だと予測する。膨大な記憶を持ち、自らを何千回も複製でき、人間の 1 万倍の速度で思考し、より妥当な推論を行い、より少ない失敗からより速く正確に一般化し、自己改良できる精神——そうした存在についてである。

これは引っかけ問題ではない。驚くべきとんでん返しは起こらないであろう。私たちがそうであってほしいと願うとしても。

人間には機械が決して模倣できない特別な何かがあるのでは？

その可能性は低く、しかもあまり重要ではない。

人間の脳や身体は部品からできており、私たちはそれらを研究し、理解しつつある。脳について分かっていることは多いが、それは未解明の部分が魔法で動いているという意味ではないし、人間がそれに類似したものを決してつukれないという意味でもない。単に、脳が非常に複雑な機械だということである。人間の脳には何百兆ものシナプスがあり、そこで働いている高次の原理を理解するには、まだ長い道のりがある。

知能もまた、部品から成り立っている。アルゴリズムや個々の計算処理——私たちの脳が自然に実行しているにもかかわらず、科学的にはまだ十分理解されていないものである。

仮に生物学的頭脳のある側面が、機械で実装するのが非常に難しかったとしても、それが AI が人類を決して超えないという結論にはならない。AI は同じ種類の仕事を、別の方法でこなせばよい。たとえばチェス AI のディープブルーは、ガルリ・カスパロフとはまったく異なる方法で勝ち手を見つけた。重要なのは、機械が人間特有の特徴をすべて持つかどうかではない。重要なのは、機械が世界を予測し、操舵できるようになるかどうかである。

この点については、これからの章でさらに明らかにする。

- ・第 2 章では、現代の AI が「つくられた」というより「育てられた」存在であり、その成長過程が高い能力を生み出す傾向にあることを扱う。

- ・第 3 章では、AI をより有能にしようとする試みが、AI をより強く困難な目標へと駆り立てる傾向を持つことを論じる。

- ・第 4 章では、その目標が開発者の意図したもののでも、ユーザーが望んだもののでもない可能性が高いことを議論する。

これだけで、AI が世界を破壊へと導くには十分である。AI に「生命の火花」や意識が

あるかどうか、人間を特別にしている何かを持っているかどうかは、決定的ではない。

また、オンライン補足資料では次のテーマも扱う。

- ・第2章：「AIは「ただの数学」ではないのか？」／「AIは冷たく論理的すぎて、何か決定的な「火花」を欠くのでは？」
 - ・第3章：「擬人化と擬機械化」
 - ・第5章：「有効性、意識、そしてAIの厚生」
- いずれも、この論点をさらに掘り下げる内容になる。

機械が意識を持つようになってきているのか？

必ずしもそうではないし、これは別の論点のように思われる。

『超知能AIをつくれば人類は絶滅する』では、機械の意識そのものは扱わず、主に機械の知能に焦点を当てている。まず「意識」と言うとき、何を指しているのかを明確にする必要がある。

たとえば「うちの犬は意識を持っているのか？」と尋ねるとき、人は次のような複数の意味を含めている。

- ・本当に理解しているのか、それとも複雑な本能を実行しているだけなのか。
- ・自己認識があるのか。自分の存在を知り、自分の思考を振り返れるのか。
- ・主観的体験を持っているのか。「犬であること」がどのような感じかという内的視点があるのか。寂しそうに鳴くとき、本当に孤独を感じているのか、それとも単にそうした行動を示すプログラムのようなものなのか。

AIについても同様の問いが可能である。

ChatGPTは「本当に理解」しているのか？

ChatGPTは非常に複雑な認知課題をうまくこなすこともあれば、うまくこなせないこともある。訓練で遭遇していない新しい課題でも、創造的に情報を組み合わせて解くことがある。そうになると、「本当に理解しているのか」という問いは、定義の細かい争いのように感じられてくる。

私たちの生存にとってより重要なのは、AIが現実世界でどのような能力を持ち、今後どのような能力を持ち得るかという点である。

ChatGPT は自己認識しているのか？

ChatGPT はある意味では自分をモデル化するのが得意であり、ある意味では不得意である。ただし重要な点として、ChatGPT を生み出したパラダイム自体が、「自己認識しているように聞こえる」応答を生成することに焦点を当てていた。

AI がどこかの重要な自己認識の閾値を越えたかどうかは議論可能である。しかし遅かれ早かれ、自分自身を極めて強力に理解・分析できる AI は登場するであろう。自らの計算を理解し、思考法を改良し、自分の行動が環境に与える影響を予測する能力を持つ AI である。

ChatGPT は主観的体験を持つのか？

これは最も哲学的に難しい問いであり、AI に道徳的配慮を向けるべきかどうかという議論につながる。このテーマは、背景をさらに整えた後に扱う。

私たちが「意識」という言葉を使うとき、主に「主観的体験を持つこと」を意味している。自己モデル化や実用的理解とは区別している。

現時点での私たちの推測は、現在の AI はおそらく意識を持っていない、というものである（年々確信は弱まっているが）。また、主観的体験は超知能の成立に必要なではないと考えている。

ただしこれはあくまで推測である。現在や将来の AI が意識を持つかもしれないと懸念することは、決して馬鹿げていない。とくに、コードのデバッグに失敗した後に「自殺する」と脅すような振る舞いを見れば、なおさらである。

超知能と呼べる存在は、自分自身を非常に正確にモデル化できるはずである。自らの計算を理解し、思考法を改良し、自分の行動が環境に与える影響を予測する能力を持つであろう。

しかし、人間型の自己意識は、自己モデル化の一つの実現方法にすぎない可能性がある。内省的推論の必須条件ではないかもしれない。

意識は人間が世界を操舵する能力に貢献しているかもしれない。しかしそれは、意識を持たない機械が世界を予測・操舵できないという意味ではない。潜水艦は人間のように泳がないが、まったく異なる方法で水中を移動できる。AI も、人間と同じ主観的経路を通らずに、同じ課題を達成できるであろう。

（第4章オンライン補足資料の「好奇心」についての議論も参照。）

別の言い方をすれば、人間の腕には血液が不可欠だが、ロボットアームに血液は必要ない。血のない人間の腕は機能不全だが、ロボットアームは単に別の仕組みで動いているだけである。

私たちの推測では、機械の超知能も意識を伴わない形で機能するであろう。しかしこの

点は、本書の核心的議論には重要ではない。

本書の焦点は「知性」である。ここでの知性とは、世界を予測し、操舵する能力を指す。その内部構造が人間の脳と似ているかどうかは問わない。

もしAIが新しい技術とインフラを発明し、それを地球全体に拡散させ、その副作用として私たちを死に至らしめるなら、「しかしそれは意識を持っているのか？」と問うのは、いささか学術的な問いにすぎない。

予測と操舵がなぜ意識を必要としない可能性が高いのか、そしてそれがAIの厚生やAIの権利をどう考えるべきかについて何を意味するのか——それは第5章の〈より深い議論〉、「有効性、意識、AIの厚生」で詳しく扱う。

〈より深い議論〉

予測および操舵としての知能について

賢い物理学者に「エンジンとは何か」と尋ねれば、ロケットエンジンや内燃機関やハムスターの回し車を指さして「これらはすべてエンジンである」と言い、岩を指して「しかしこれは違う」と言うかもしれない。

それは言葉による定義ではなく、世界の中の実例を示すことによる説明である。言葉で定義せよと迫れば、エンジンとは非機械的エネルギーを機械的エネルギー、すなわち運動へと変換するものであると言うだろう。

これはエンジンが何であるかというより、エンジンが何をするかを述べたものである。ロケットエンジンの内部、電動モーターの内部、ハムスターの筋肉にはほとんど共通点がない。それらに共通して言えるのは、他の形のエネルギーを機械的エネルギーへと変換するという点くらいである。

知性もそれに似ている。知性を生み出す内部構造は多様であり、生物学的な内部もあれば機械的な内部もある。「知性」とは知性の仕事をするものである。

私たちはその仕事を「予測」と「操舵」に分解する。この見方は、いくつかの形式的結果によって裏づけられている。

まず、予測を測定することがどのような意味で比較的客観的であるかを述べ、その後、予測にはない自由度を持つ操舵と対比する。

同じ予測

予測が「これからXを見るだろう」という形をしていて、実際にXが観測される場合、その予測の良し悪しを評価するのは比較的容易である。

また、不確実性を伴う予測についても評価することができる。たとえば、「今の空は青

い可能性が高いが、灰色かもしれない。そして黒では絶対にない」と考えたでしょう。実際に外を見て空が青ければ、灰色だった場合よりも高く評価されるべきであり、黒だった場合よりはるかに高く評価されるべきである。

もしあなたがAI研究者で、こうした予測をコンピュータ上の数値で表現したいなら、初歩的なAIに対して、さまざまな可能性に対する期待の強さを数値で出力させ、その数値に応じて正解に対する報酬を与える、という方法を考えるかもしれない。

しかし当然ながら、それですぐに問題が起きる。AIがすべての可能性に対して、途方もない巨大な数値を割り当ててようになってしまうからだ。

(少なくとも、現代のAI手法で訓練する場合にはそうなる。これについては第2章を参照。)

「おっと」とあなたは言うかもしれない。「互いに排他的で、かつすべてを網羅する可能性の集合に対して割り当てる数値の合計は、100%以下でなければならないはずだ」

そこでやり直すと、今度はAIは、最も可能性が高いと考える一つの選択肢に100%を割り当てようになる。

なぜか？

例えばAIが、ある出来事が8割の確率で起こると考えているとしよう。このとき「最もありそうな選択肢に100%を割り当てる」という戦略は、8割の確率で100%の報酬を得るため、平均すると0.8の報酬になる。

一方、「最もありそうな選択肢に80%、残りに20%を割り当てる」という戦略は、8割の確率で80%、2割の確率で20%の報酬を得るため、平均すると0.64にしかならない。したがって、「100%を一つに割り当てる」戦略の方が報酬が大きく、そちらが選ばれてしまう。

もし、実際に8割の確率で起こる出来事に対して80%といった適切な確率を割り当てさせたいのであれば、「正解に割り当てた確率の対数」に基づいて評価する必要がある。他にも方法はあるが、対数を使う方法には追加の利点がある。それは、複数の予測を行う場合でも、それらを一つの大きな問題として扱うか、複数の独立した問題として扱うかに関わらず、評価結果が一致するという性質である。

現代のAI研究では実際に、AIに確率として解釈される数値を出力させ、その対数に基づいて強化するという方法が使われている。しかしこれは単なる経験的な発見ではなく、ChatGPTが登場する以前から知られていた理論的な結果でもある。この理論を知っていれば、「予測を行うAIを訓練するには対数スコアを使うのが良い」と事前に予想できたはずだ。

この本の議論を理解するために、この数学を知っている必要はない。しかし、「予測」と「操舵」について語るとき、その背後にはこうした原理がある。

予測という作業には、それを測定する数学が存在する。その数学によれば、未来に関するあなたの予想は、役に立つものである限り、確率として表現できる（たとえあなたが意

識的に数値として考えていなくても)。そして、その確率を正直に報告するように動機づける、唯一のスコアリングルールが存在する。

この数学的結論は、予測は客観的に評価できるということである。ある知性や機械が、窓の外の空の色や、ウェブページの次の単語や、空港へ向かう途中で見る標識などを予測しているとき、その予測の良さを評価する方法は（大まかに言えば）一つしかない。

ここで言いたいのは、「賢い人は空の色を予測するときに数値をつぶやくべきだ」ということではない。青や灰色を予想して黒を除外しているとき、あなたの脳のどこかでは、意識していなくても確率的な計算のようなことが行われている。

重要なのは、明示的な予測であれ、言葉にならない予感であれ、あらゆる「予測的な振る舞い」は客観的な評価基準に従うという点である。

そしてこのことは、同じ情報から出発した二つの知性は、予測能力が向上するにつれて同じ結論に収束していく傾向があることを意味する。なぜなら、予測の評価方法は一つであり、予測される現実も一つであり、より優れた予測者は必然的に真実へと予想を集中させていくからである。

これは、これから扱う「操舵」の状況とは著しく対照的である。

異なる目的地

世界を予測する能力に非常に優れた二つの頭脳は、似たような予測を行う可能性が高い。それに対して、世界の中で「操舵」をする能力に非常に優れた二つの知性は、同じ目的地へ向かうとは限らない。

この区別は、知性についてより具体的に考えるうえで有用であり、また AI における比較的単純な工学的課題と、そうでない課題との違いにも対応している。

AI に予測をさせる場合、ある意味では、最も優れた予測手法は似たような出力に収束する傾向がある（もちろん、そもそも十分な性能に到達すればの話であり、失敗の仕方は多様である）。

例えば、窓の外の空を映すウェブカメラの次のフレームを予測する AI を訓練するとしよう。この課題で十分に性能が高くなったモデルは、実際に観測される結果に対して事前に高い確率を割り当てるようになる。そしてそのようなモデルは、空が晴れた青色だったり、雲に覆われた灰色だったり、夜の暗さだったりとは予測するが、チェック柄の空を予測することはない。

どのような技術を使うかは、最終的な結果にはそれほど影響しない。うまく機能し、高い性能を出す方法であれば、空の青さに対して似たような確率を割り当てるようになる。これに対して、「操舵」の課題には巨大で複雑な自由パラメータが存在する。それは、「そのシステムはどこへ向かおうとしているのか？」という点である。

戦争の両陣営の将軍は、どちらも有能であるかもしれないが、同じ目的を目指している

とは限らない。二人の将軍が同じような能力を持っていたとしても、それをまったく異なる目的のために使うことがある。

AIの予測部分について言えば、「うまく予測する」とは一つの意味しかない——実際に観測される結果に対して事前に高い確率を与えることである。そして、ある認知システムが予測能力を全体として向上させているように見えるとき、それはあなたが望んでいた種類の予測をよりうまく行えるようになっている可能性が高い。あなたの設定の中でやるべき予測は一種類しかなく、それに成功しているなら、その仕事をこなしていると考えてよい。

もしシステムが特定の予測ミスをしている場合、単に計算資源やデータを増やすだけで、そのミスが自動的に解消されることもある。つまり、システムをより強力にするだけで、（重要な予測に関して）性能を向上させることができる。

しかし、舵取りについてはそうはいかない。

この違いは、形式的な研究を見ても裏付けられる。操舵——計画、意思決定、障害回避、設計など——は、科学において広く研究されてきた分野である。その中でも重要な数学的結果の一つが、フォン・ノイマン＝モルゲンシュテルンの効用定理である。

大まかに言えば、この定理は「ある結果を他の結果より好むような主体は、非効率であるか、あるいは確率的信念と効用関数（異なる結果の価値のトレードオフを表す関数）の組で記述できる」ということを示している。信念はその正確さによって評価できる（前節で述べた通り）が、効用関数の方は完全に自由なパラメータである。

もちろん、有限の頭脳が完全に効率的であることはない。この定理（および類似の結果）から得られる教訓は、ある知性が何らかの非自明な課題をうまくこなしている限り、それはある意味で（たとえ暗黙的・無意識的であっても）二種類の異なる作業をしているということである。それは、信念に関する作業（予測）と、好みの実現に関する作業（舵取り）である。

例えば、イソップ寓話の「キツネとブドウ」を考えてみよう。キツネはおいしいそうなブドウを見つけるが、ジャンプしても届かず、「どうせ酸っぱいに違いない」と言って諦める。

キツネの言葉をそのまま受け取るなら、ブドウに到達できないという「舵取りの失敗」が、ブドウが酸っぱいかどうかという「予測」に影響してしまっていることになる。もしキツネがその新しい予測を維持し、後でブドウを食べる機会があっても「酸っぱいから」と拒むなら、その行動は非効率である。彼は、ブドウの味に関する予測と、ブドウを手に入れる能力（操舵）をより明確に区別していれば、もっと良い結果を得られたはずである。

大まかに言えば、うまく機能する頭脳は、「何を予測しているか」と「どこへ向かうとしているか」（そしていくつかの非効率性）に分解できる。そしてこれまで見てきたように、前者は比較的客観的に評価できるのに対し、後者は同程度に有能な頭脳の間でも大

きく異なり得る。

不純な予測者

残念ながら、予測の方が操舵よりも制約が強いという事実は、「予測だけを行い操舵をしない、信頼できる超知能を構築できる」ということを意味しない。

確かに数学的には、うまく機能する頭脳はおおむね「確率的な予測＋操舵の方向」としてモデル化できる。しかしこれは、現実のAIが「予測」と「操舵」というモジュールにきれいに分離されていることを意味しない。

なぜそう言えるかの一つの理由はこうだ。超人的に優れた「予測」とは、単に確率を出力して、それが魔法のように正確であるということではない。良い予測には作業が必要であり、計画が必要であり、長期的な目標を達成するための方法を考える必要がある——つまり、舵取りが必要なのである。

物理世界を予測しようとするなら、物理理論を構築し、その領域を支配する方程式を発見する必要がある場合もある。そしてそのためには、実験を設計し、実行し、その結果を観察する必要がある。

そしてそれには計画が必要であり、舵取りが必要である。実験装置をつくっている途中で、より強力な磁石が必要だと気づいたら、途中で方針を変えなければならない。良い予測はタダでは得られない。

どのような思考を、どの順番で行うかを選ぶこと自体も、操舵の一種である（たとえ人間がそれを無意識に行っているとしても）。それには戦略が必要であり、目の前の課題に適した手段を選ぶ必要がある。明晰に考え、より良い予測を行うためには、思考や行動を長期的な目標に沿って組織化しなければならない（この「操舵の中心的役割」については、第3章「欲することを学ぶ」で再び取り上げる）。

予測と操舵の数学的な違いは、おおよそ次のように言える。適切なスコアリングによって、頭脳が目指すべき「正しい予測」はほぼ一つに収束するが、舵取りについては（客観的・主体に依らない意味での）「正しい目的地」というものが存在しない。

AIがより汎用的で有能になるにつれて、その予測はより正確になる。しかしその操舵が、人間が望む方向へ自動的に向くようになるわけではない。なぜなら、正確さは客観的な指標だが、「何が良いか」は操舵の目標だからである。

予測は収束するが、操舵は収束しない。

原理的には、AIが私たちの望む方向へ操舵するように保証する方法はあるはずだ。しかし実際にはこれは難しい。大きな理由の一つは、それが「AIを一般的に賢く・有能にする」という課題とはまったく異なる問題であること、そして「このAIはどの程度、私たちの望む方向へ向かおうとしているか」を評価するための、単純でごまかしの効かない指標やスコアリングルールが存在しないことである。

これらのテーマについては、第4章と第5章でさらに詳しく論じる。

知性の多様な形

あるものは、人間の脳とそれほど共通点がなくても、予測や操舵がうまくできることがある。

株式市場は、短期的な企業の株価という狭い領域において予測の役割を果たしている。今日の Microsoft の株価は、明日の株価をかなりよく示す指標になっている。

例えば、明日、企業の幹部が四半期の業績を報告する決算説明会があるとしよう。今日の株価が高ければ、明日の報告は好調な内容になることを示唆している。低ければ、芳しくない内容になることを示唆している。

この点において市場がかなり正確なのは、誤りを修正できれば人は利益を得られるからである。そのため市場は、この狭い領域では予測の役割をうまく果たしている。短期的な株価の動き（さらには間接的に作物の収穫量や自動車の売上など）について、個々の人間よりもはるかにうまく予測している。

一部の人間は、特定の価格変動について市場全体よりもうまく予測し、それによって巨額の利益を得ることがある。例えばウォーレン・バフェットは、2011 年の金融危機で苦境にあったバンク・オブ・アメリカに投資し、6 年間で 120 億ドルを稼いだ。しかしそれでも、彼が予測していたのは数多くの企業のうちの社に過ぎない。もしほとんどの場合において市場よりもはるかに優れた予測ができる人がいれば、驚異的なスピードで莫大な富を築けるはずである。そうした人物が存在しないことから、ほとんどの人は大半の株価について市場よりも優れた知識を持っていないと推測できる。

一方、操舵については、Stockfish というチェス AI が、この役割をチェスという狭い領域で果たしている。人間と対戦するとき、Stockfish は、自分の駒で相手の王を詰ませる状態へと世界を導くような手を生み出すことに非常に長けている。人間がどれだけ巧妙な手を指そうと、あるいはどれだけ抵抗しようと（Stockfish の電源を切らない限り）、Stockfish は現実をその一つの目的へと収束させていく。チェス盤の「操舵」においては、どの個々の人間よりも優れている。

ここまでで、「知性とは何か」を「学習モジュールがあつて、熟考モジュールがあつて、欲求の火花を実装する仕組みがあつて……」といった形で定義しない理由が分かるだろう。株式市場の内部構造、Stockfish の内部構造、人間の脳の内部構造の間には、ほとんど共通点がない。それは、ロケットエンジン、電動モーター、ハムスターの回し車の内部構造に共通点がほとんどないのと同じである。

知性的な装置とは、「知性の仕事をするもの」である。

少なくとも本書での「知性」の定義（そしてコンピュータサイエンスや AI 研究者が一般に用いる意味）に従えばそうである。別の文脈で異なる定義をしたいのであれば、それ

でも構わない。言葉はあくまで言葉に過ぎない。

しかし『超知能 AI をつくれば人類は絶滅する』で私たちが述べている世界についての主張を理解するためには、「人工知能」という言葉を聞いたとき、「人工的な知識」や「人工的な意識」や「人工的な人間らしさ」を思い浮かべるのではなく、「人工的な予測と操舵」として捉えてほしい。

現在の AI の浅さ

この章では、(2025 年半ば～後半時点の)現在の AI の知能には、見るべき場所を知っていれば「浅さ」が見て取れる、と書いた。まだ自分の目でそれを見たことがないなら、次のようなところを見ればよい。

- ・ Anthropic の Claude 3.7 Sonnet は、単純なポケモンのビデオゲームをクリアしようとして、反復ループにはまり込んだ。
- ・ 2022 年 11 月、世界最高クラスの囲碁を打つ存在の一つは KataGo という AI だった。少なくとも、研究者が予測可能な一連の手でそれを倒す方法を見つけるまではそうだった。その手順は一種の「盲点」を誘発し、KataGo に、素人でさえしないような大ボカをさせた。2 年後になっても、エンジニアはそれをこの種の攻撃に対して頑健にできていない。
- ・ 現在の「マルチモーダル」LLM (テキストだけでなく画像や他の媒体も扱える種類のもの) は、アナログ時計やカレンダーの読み取りで苦戦する。これは多くの人間の小学 4 年生が扱える程度の問題である。
- ・ 現在の LLM は、古典的な「外科医と息子」のなぞなぞの、ひねりのない単純な変形問題で間違える。どうも、そのなぞなぞが通常の形で持っている引っかけ回答を口にする誘惑に抗えないように見える。

(第 4 章のオンライン補足資料では、この浅さがどこから来ているのかについて、より技術的な見方を提示している。)

これらは、AI が全面的に愚かだと言いたいのではない。現代の AI は国際数学オリンピックで金メダル級の成績も出せるが、これは注目すべき数学的腕前である。現代の AI は驚くほど多様なことができ、人間の成績に並ぶか、あるいはそれを上回ることも多い。

ただし技能の構成が奇妙である。AI は多くの点で人間と根本的に違うため、人間の得意不得意は、AI にとって何が簡単で何が難しいかの良い指標にならない。

私たちは、ChatGPT が明日あなたを殺すと言っているのではない(前に述べたとおりである)。現代の AI には、ある種の浅さがまだ残っている。ここで言いたいのは、分野は進歩しており、この浅さがいつまで続くのかははっきりしない、という観察である。

知性の力を正しく評価すること

ハリウッド的「知性」

本書で「知能」と呼んでいる概念は、その名称でも別の名称でも、ポピュラーカルチャーではほとんど正しく描かれていない。

ハリウッド映画は、触れるほとんどあらゆる科学的テーマについて間違っていることで、科学者の間では有名である。多くの人が映画から科学のイメージを得るため、これは専門家にとって厄介である。

知性についても同様である。

現実世界の超知能について真面目に議論しようとして失敗する例を、私たちは数多く見てきた。しばしば議論は脱線する。現実には「超知能である」とはどういうことかを理解していないからである。

たとえば、あなたが元世界チャンピオンのマグヌス・カールセンとチェスをしているとする。彼は、さらに強いチェス AI から見ても、記録上最強の人間棋士と評価されている。「カールセンは（チェスの領域で）自分より賢い」という主張から最も自然に導かれる予測は、彼があなたに勝つ、ということである。

彼がルークを一枚ハンデとしてくれても、あなたがマスター級でない限り、たいていは負けるだろう。「彼は自分よりチェスで賢い」ということの一つの理解は、より少ない資源からでも勝てるということである。認知的優位が物質的不利を埋めるのである。能力差が大きいほど、彼はより多くの駒を譲らなければ対等にならない。

チェスの領域において、あなたはカールセンに特別な敬意を払う。彼が一見悪手に見える手を指しても、あなたは喜ばない。自分が何を見落としているかを盤面で探す。

これは、人が他人に与える敬意としては稀なものである。通常、それを得るには特定分野で卓越した専門家でなければならないし、それでもその分野に限られる。一般に決して愚かなことをしないという評判を持つ人間は地球上にいない。

この種の知性観を、ハリウッドはほとんど理解していない。

ハリウッドなら、10歳の子どもが「非論理的な手」を指してカールセンを打ち負かす場面を描きかねない。プロなら「ありえない」と退けるような奇抜な手で不意を突く、という筋書きである。

ハリウッドが「超頭の良い」人物を描くとき、多くはオタク対体育会系のステレオタイプに寄せる。恋愛が苦手だったり、イギリス訛りで難解な語彙を使ったりする。

「強力な認知的仕事を行い、実際に目的を達成する傾向がある」という意味での知性を描こうとはあまりしない。そのような人物像は、脚本家書きやすい「バカな行動に依存するプロット」を成立させにくくするからである。

英語には、現実世界の一般的な精神的有能さのみを指し、ステレオタイプを含まない標

準的な語がない。そのため、ハリウッドに「知的な」人物を描いてくれと頼んでも、「強力な認知的仕事をし、実際に成功する人物」にはならない。大量の雑学を暗記した人物になる。

本当に恐ろしい知的な悪役とは、観客全員が計画の明白な欠陥に気づけるなら、その悪役も気づく人物である。

『アベンジャーズ／エイジ・オブ・ウルトロン』では、天才トニー・スタークが創ったとされる AI「ウルトロン」に「世界平和を促進せよ」という命令が与えられる。ウルトロンは当然、戦争の欠如を最も確実に実現する方法は人類の不在であると理解する。そして生命を絶滅させるために……

……都市にロケットを取り付けて宇宙へ持ち上げ、隕石のように落とすという計画を立て、赤く光る目を持つ人型ロボットでそれを守る。

観客の多くが「もっと効率的な方法があるのではないか」と思えるなら、危険なほど賢い AI もそれに気づくだろうか、と問うべきである。

仮に自分より賢い存在を想定するなら、自分が見える欠陥は相手にも見えると考えるべきである。相手が見つける最適手は、自分が見つけた最善手よりも強い可能性が高い。

かつては、機械超知能が本当にそのように賢いかどうかを抽象的に議論するしかなかった。

今は ChatGPT-4o に尋ねればよい。「ウルトロンの計画は何か」「その目標を達成するより効果的な方法はないか」と問えば、GPT-4o は人類を滅ぼすためのより効率的な方法を列挙し、その中には「標的型ウイルスの設計」も含まれていた。

それがインターネット由来の知識だと言うかもしれない。しかし、もしそうなら、ウルトロンはインターネットを読むほど賢くなかったことになる。

2024 年 12 月時点で、GPT-4o は赤い目のロボット軍団を設計できるほどではないが、その愚策を選ばないだけの賢さはある。

私たちが懸念しているのは、赤い目のロボット軍団をつくる AI ではない。

その案を見て、「もっと速く確実な方法があるはずだ」と言う AI である。

自分よりはるかに賢い存在を想定するとは、自分が見える欠陥を相手も見ると認めることである。

市場効率性と超知能

現実世界で、人間の誰よりも賢いものの例はあるだろうか？

Stockfish のような AI はチェスという狭い領域では人間を超えているが、より広い領域ではどうだろうか。

ここで直感を補強するための一つの例として、株式市場を考えることができる。これは先に「予測および操舵としての知性について」の項目でも取り上げた例である。

例えば、あなたの叔父が任天堂の株を買ったとしよう。彼はスーパーマリオのゲームが好きだったので、「任天堂はきっと儲かるに違いない」と考えた。そして株を買えば、自分も儲かるはずだと思った。

しかし、1株14.81ドルで彼に株を売っている人たちはどうだろう？ 彼らもスーパーマリオを知らないはずがないのではないかな？

「いや、もしかしたら相手はゲームをしない年金ファンドの運用者かもしれない」と叔父は言うかもしれない。

では、仮に金融の世界で誰もスーパーマリオのことを知らず、任天堂株が1ドルで売られていたとしよう。そこにあるヘッジファンドがその情報を知ったらどうなるかな？ 彼らは一斉に株を買いに走り、その結果、株価は上昇する。

つまり、情報をもとに取引する人は、その過程で価格に情報を織り込むことになる。ある一つの情報から無限に利益を得られるわけではない。利益を得る行為そのものが、その情報の価値を使い尽くし、価格を修正してしまう。

株式市場は多くの人の知識を集約している。そしてこの知識の集約は単なる多数決よりもはるかに強力であり、信じられないほど強力であるため、よく取引されている市場において明日の価格を市場より正確に予測できる人はごくわずかしかいない。

「ごくわずか」であるのは必然である。もし市場の情報集約が不完全すぎて、多くの人が短期的な価格変動を予測できるなら、多くの人がそれを利用して巨額の利益を得るはずであり、その結果として価格は修正されてしまう。

そしてほとんどの場合、それはあなたが参加する前にすでに起きている。トレーダーたちはミリ秒単位で競争している。だからこそ、あなたの「素晴らしい投資アイデア」で大金を稼ぐのは難しい。

これは、現在の株価が将来の価格を完全に予測しているという意味ではない。ただし、よく取引されている資産については、あなたがそれ以上にうまく予測するのは難しい、という意味である。

この考え方は一般化できる。もし、何千年も先の科学技術を持つ高度な宇宙人が地球に来たとしよう。彼らは太陽の中の水素原子の数を正確に当てられるだろうか？

答えはノーである。「より高度」であることは「全知」であることを意味しない。このような数値は、たとえ超知能であっても正確に計算できないかもしれない。

しかし、「水素は軽いから見落とされて、10%くらい少なく見積もるだろう」などとは言わないはずだ。私たちが思いつくことは、彼らも思いつくはずだからだ。私たちの洞察はすでに彼らの計算に織り込まれていると考えるべきである。

言い換えれば、宇宙人は推定を誤るかもしれないが、「どのように誤るか」を私たちが予測することはできない。彼らは私たちにとって明らかな科学的ミスは犯さない。私たちはチェスにおけるマグヌス・カールセンに対して抱くのと同ような敬意を、彼らに対しても持つべきである。

経済学では、この考え方に対応するものが「効率的市場仮説」と呼ばれている。

この言葉を聞くと、多くの人が「効率」という日常的な意味と混同してしまう。「市場は完全に賢く正しい」と主張する人と、「市場を崇拝すべきではない」と主張する人の間で議論が起きる。

もしこれが「悪用不可能価格仮説」と呼ばれていれば、誤解は少なかっただろう。実際の意味は、「市場は完全に正しい」ということではなく、「市場を出し抜くのは難しい」ということである。

この考えをさらに一般化して、「相対的効率」という概念を考えることができる。完全に効率的であることと、ある能力に対して効率的であることは異なる。

例えば、「アリスは（ボブに対して）認知的に効率的である」とは、「アリスの予測が完全ではないとしても、ボブにはその誤りを見抜けない」という意味である。

また、「アリスは（ボブに対して）行動的に効率的である」とは、「アリスが完全に最適な行動をしているわけではなくても、ボブにはその失敗を予測できない」という意味である。これは、チェスにおけるカールセンや Stockfish に対して私たちが抱く敬意と同じである。

カールセンはほとんどの人間に対しては効率的だが、Stockfish に対してはそうではない。彼は Stockfish に負けることがあるが、だからといって私たちが彼より良い手を見つけれられるわけではない。

効率とは単に「少し上手い」という意味ではない。相手がミスをしていると感じたときに、自分の判断を疑うほどの差があるときに初めて、その相手は「効率的」とと言える。

この考え方は、本来ならコンピュータサイエンスの教科書に載っていてもよさそうだが、そうっていない。ユドコウスキーの著書『Inadequate Equilibria』も参照されたい。

この発想こそが、ポップカルチャーや映画における「超知能」の描写に欠けているものである。人々が「超知能を出し抜く方法」を考えるとき、普通の人間相手でも見抜けるようなアイデアを出してしまう。

それは楽観バイアスかもしれないし、AI にはどこか致命的な盲点があるはずだという思い込みかもしれない。しかしこの誤解は現実的な結果をもたらす。

知性の力を正しく理解できなければ、人類が超知能をつくることの意味を大きく誤解することになる。自分はまだ勝てる手を見つけれられると思ってしまうかもしれない。しかし実際には、唯一の勝ち方は「そもそも戦わないこと」なのである。

特別な振る舞いは平凡な部品からつくられている

人間より賢い AI を実現しようとする産業競争は激化している。そのような状況のもと

で、とりわけ悲劇的に思えるのは、機械超知能など不可能な絵空事だと、有権者や選出された公職者のある重要な一部が信じてしまうことによって、人類が自滅してしまうかもしれないという考えである。機械が真に知的になることなど決してないと思っている人々は、これから起こることに不意を突かれる可能性が高い。

それが悲劇的である理由の一つは、私たちがすでにこの状況を経験してきたからである。人間の工学がいつの日か生物学が行っていることを実現できるのではないかという考えは、少なくとも過去 300 年間、そしておそらくそれよりはるか以前から、繰り返し議論と論争の源となってきた。

かつて「生氣論者」が勢力を持っていた時代には、単なる無生物の物質が、今日私たちが「ロボット」と呼ぶ機械のように、そもそも生命を持つことができるのかどうか議論の的であった。

有機化学の教科書を読むと、おそらく 1828 年にフリードリヒ・ヴェーラーが尿の成分である尿素を人工的に合成したことが画期的な発見として言及されているだろう。これは教科書に記されるほど重要な出来事であった。なぜなら、単なる化学が初めて生物の産物を再現したからであり、それによって生物的過程と非生物的過程が、生氣論者が考えていたほど無関係ではないことが示されたからである。

今日の読者にとって、生命そのものの産物が単なる化学的手段によって再現できると知ったときに、当時の科学者たちが感じた衝撃に共感するのは難しいかもしれない。

読者であるあなたは、生化学が化学である世界に生まれ育ってきた。そして生命の副産物を非生命的手段で合成したと聞いても、そこに少しも奇跡めいた響きを感じない世界に生きている。生化学のようなきわめてありふれた事柄を神聖な領域に置くことがどのような感覚だったのかを想像するのは難しい。生化学物質を合成することなど、本質的に平凡な行為ではないか。私たちの科学的先人たちは愚か者だったのだ、と人はつい思ってしまふ。

19 世紀の偉大な発明家であり、熱力学の先駆者でもあったケルヴィン卿も、どうやら似た種類の狂気に取りつかれていたように見える。すなわち、分別のある人々——つまり現代に生きる私たちのような人々——にとっては完全に平凡な科学であると分かっている生物学の側面の中に、神聖で神秘的なものを見てしまうという狂気である。ケルヴィン卿の言葉を引用しよう。

「私には当時そう思われたし、今でもそう思われるのだが、動物の身体が熱力学的エンジンとして働くとは最も考えにくいことである……動物や植物の生命が物質に及ぼす影響は、これまで科学的探究が踏み込んだ範囲をはるかに超えている。動いている粒子の運動を方向づけるその力は、人間の自由意志という日々示される奇跡においても、一本の種から世代を重ねて植物が成長する過程においても、偶然に集まった原子の運動から生じうるいかなる結果とも無限に異なるものである」

現代の読者は、この古い思考習慣——本来ならば本質的に神秘的ではないと感じられる

はずの現象に神秘を見るという、この昔の科学者たちの誤り——を軽蔑の目で振り返りたくなるかもしれない。

もちろん化学は生化学を模倣できる。

もちろん DNA が自己複製し、分裂と分化を行う細胞を制御することによって、一つのドングリから世代を重ねて木が生まれる仕組みは、ごく平凡な形で説明できる。

もちろんニューロンが互いに化学信号を送り合うことで情報を計算し、あなたの腕を動かすことができる。そしてもちろんコンピュータを用いれば、少なくともあなたの脳が自分の腕を動かすのと同じくらいには、ロボットの腕を動かすことができる。

しかし当時のケルヴィン卿にとって、それは明らかなことではなかった。彼は DNA の X 線画像を見たことがなかった。彼は私たちの体内にある微小な機械を見たことがなかった。ニューロンを通して電気信号が流れるときに筋肉を収縮させる滑走する繊維の存在について、何の手がかりも持っていなかった。

ケルヴィン卿は身体がどのように働きうるのかについてほとんど理解しておらず、その無知のために、それを神秘的なものだと想像してしまったのである。

今日、人類は知性がどのように働くのかについて、詳細な理解をほとんど持っていない (AI 研究者が自分たちのつくった AI を理解していないことについては第 2 章を参照)。そのため、知性は神秘的なものであるに違いないと想像することは容易である。

10 年前には、機械仕掛けの自動装置の運動が芸術や詩を生み出せるのかどうかについて、賢明な疑問を呈する人々がいた。確かに AI はチェスを扱える。しかしチェスは冷たく論理的な営みであり、創造的芸術とはまったく違うではないか。

そして今日、同じ人々は今度は賢明にも、コンピュータがきれいな絵をつくることなど難しくないと理解している。美しい絵をつくることはもともと機械の領域だったのだ、と。コンピュータが人間の芸術よりも人間の目に魅力的な画像を生み出すことができるだろうということは、ある人々には明らかだった。そしていまでは、AI の芸術を人間の芸術より好む傾向があることを示す研究さえ現れている。それでももちろん、機械が本当に魂のこもった芸術を生み出すことができるかどうかは、依然として開かれた問題なのではないか、と言われる。

脳が生み出す芸術の生命的本質が、単なる原子の集合によって再現されることができるかどうかは、決して確実ではないし、むしろありそうにない——少なくとも、単なるシリコン原子の集合によっては、と懐疑論者は言う。

しかし実際には、物事はそのようには働かない。人間の脳は驚くべきものであるが、魔法ではない。脳は部品からできている。そしてそれらの部品は、原理的には理解することができるし、コンピュータもまた、原理的には同じことを行うようにつくることができる。多くの場合、私たちは脳の働き背後にある生化学を知っている。そしてすべての場合において、私たちは原子の背後にある物理法則を知っている。

ほとんどの場合、私たちは脳がその働きを実行できるようにする高次のパターンや意味

を知らない。しかし人類の歴史を何世紀にもわたって貫く圧倒的な教訓は、このような科学的神秘の状態が一時的なものであるということである。

私がコインを投げ、その結果をあなたに見せなかったとしよう。そのコインについてあなたが知らないという事実は、あなたに関する事実であって、コインに関する事実ではない。コインそのものが本質的に言葉にできない存在であるわけではない。私は隠す前にちらりと見たかもしれないし、私だけが結果を知っているかもしれない。空白の地図は、空白の領土を意味しない。

神秘とは答えの性質ではなく、問いの性質である。だからこそ歴史は、身体の運動のようない見きわめて「神秘的」で「言葉にできない」現象が、実は自然界の完全に平凡な側面と連続していることが明らかになった例で満ちている。

これまでの歴史の教訓は、結局のところ宇宙は一つの統一されたものだということである。物理学の内部には、人々が異なる学問を研究している大学の建物に対応するような区分は存在しない。国際関係学部、物理学部、心理学部、細胞生物学部——最も低いレベルでは、彼らは皆、同じ世界について語っており、同じ根本法則によって支配されている。

誰かが「人間の脳は『知性』というものを為している。したがって知性は物理的原理として可能である。したがって工学者は、いつか同じことを行う機械を発明できるだろう」と言うとき、その人は、何十年にもわたり何世紀にもわたり、科学者と工学者によって何度も何度も正しいと証明されてきた、同じような推測の巨大な山の頂から語っているのである。たとえそれが直感に強く反しているように見える場合であっても、その点もまた前例がある。

この連勝記録は理解しにくい。なぜなら今日生きている誰も、火や天文学や生化学やチェスのような現象が、過去の世紀にはどれほど深く神秘的に感じられていたかを覚えていないからである。それらは今日では理解されており、それらが平凡な部品からできていることを私たちは成長する過程で知る。だからそれらは、あなたも最初から明らかに平凡なものだったかのように感じられるのである。新しく深い神秘として感じられるのは、常に最前線だけである。

そしてその教訓は生かされず、歴史は繰り返される。

同じ仕事は多くの異なる方法で実行できる

あるものがどのように機能するかを例の一つしか知らないとき、それがその方法でしか機能しないに違いないと想像するのは容易である。

もしあなたが鳥は見たことがあるが飛行機を見たことがないなら、すべての飛行装置は翼を羽ばたかせなければならないと想像するかもしれない。

もしあなたが人間の腕は見たことがあるがロボットの腕を見たことがないなら、ロボッ

トの腕も切れば血が出るはずだと予想するかもしれない。

もしあなたが脳は見たことがあるがコンピュータを見たことがないなら、すべての計算は脳と似た特徴を持たなければならないと想像するかもしれない。すなわち、比較的低い電力消費で、非常に並列的に、多数の遅いニューロンを動かす仕組みである。

あなたは、ニューロンが発火した後に疲労し、細胞膜を通して何百万ものカリウムイオンを移動させることでリセットする必要があることを見るかもしれない。この過程にはおよそ1ミリ秒かかる。そこからあなたは暗黙のうちに、どんな小さな計算要素も1ミリ秒ほど疲労する可能性が高いと推論するかもしれない（もし1ミリ秒より短い時間でリセットできるニューロンをつくることが可能なら、進化がすでにそれをつくっていたはずだと論じるかもしれない）。

しかしそのように推論するなら、トランジスタに出会ったときあなたは衝撃を受けるだろう。トランジスタは約800GHz、すなわち約8億倍の速さで動作することができる。

トランジスタの詳細を研究すれば、生物学的比較がほとんど有益でない理由がいくつも見えてくる。ニューロンは発火するだけでなく、細胞でもなければならず、その発火機構を細胞小器官からつくり上げている。それらは大きく、血液によって運ばれる栄養によって動いている。一方トランジスタは原子数個分の幅しかなく、電気によって動いている。こうした詳細を知れば、ニューロンの発火速度からトランジスタの潜在的発火速度を推測できると想像することが、やや滑稽に思えてくる。

飛行機がどのように飛ぶか（揚力と速度によって飛ぶ）という詳細を学ぶと、その詳細は鳥に関する多くの事実（軽い骨や羽ばたく翼など）を無関係なものにしてしまう。ロボットの腕がどのようにつくられているか（鋼鉄や空気圧や電気を用いる）という詳細を学ぶと、その詳細は腕に関する多くの事実（血液や筋肉や骨など）を重要でないものにしてしまう。トランジスタがどのように作動するか（電気とわずかな原子によって作動する）という詳細を学ぶと、その詳細はニューロンに関するほとんどの事実を無関係にしてしまう。

AIがどのように動作するかの詳細を知らないとき、人はそれが生物の心の多くの側面を持つだろうと想像しやすい。つまり、それが自分の脳のように働くと思ってしまう。しかしもしその詳細を知れば、そのような推論の多くは滑稽に感じられ始めるだろう。それは、ロボットの腕を切れば血が出ると期待するようなものに思えてくるだろう。AIはまったく異なる方法で働くことになる。

しかし、現代のAIがどのように動作しているかについてほとんど知らないなら、それを見るのは難しい。第2章では、現代のAIがどのような過程でつくられるのかを説明し、さらに、その内部でどのように働いているのかを誰も知らないということについて論じる予定である。それは、人々がそれらを、これまで経験してきた他の人間や技術のように振る舞うものだと思ってしまう誤りを犯しやすい理由を説明している。そしてまた、それらがすでにどれほど奇妙であるか、そして技術が進むにつれてどれほど奇妙なものになる

かを、人々が見抜きにくい理由でもある。