

2章 つくられるのではなく育てられる

これは『超知能 AI をつくれば人類は絶滅する』第2章のオンライン補足資料である。ここでは現代の AI がどのように動いているのか、そしてなぜそれが「単なる機械」や「単なる道具」ではないのかについて扱う。

AI はコンピュータ上で動くコードであるが、人間が細かく設計した従来のソフトウェアとは異なり、人間の発明について通常私たちが当然視している多くの前提を覆す。

なお、以下のテーマは書籍本編で扱われているため、ここでは詳述しない。

- 現代の AI がどのような意味で「つくられた」のではなく「育てられた」と言えるのか。
- 現在の AI はどのように育てられているのか。
- 「勾配降下法」とは何か。なぜそれほど単純な手法から柔軟で複雑な AI が生まれるのか。
- これらの AI は私たちとどれほど異なり得るのか。

なぜ勾配降下法が重要なのか？

これは、エンジニアが現代の AI をどこまで、そしてどのように形づくれるのかを理解するうえで重要である。

もしエンジニアが自分たちでも理解していない AI を「育てて」いるのだとすれば、その AI がどのように振る舞うかを形づくる能力ははるかに制限される。理解の欠如は、工学的制御力を制約する。

本書の後半で描く破局の詳細な構図は、この点に由来する。人間が AI に「新しいことができるようになるれ」と要求したとき、返ってくる解決策は、エンジニアが意図的に選んだ設計ではない。それは、単純な最適化器が 1000 億個もの数値を試行錯誤で調整する過程で偶然見つけた、「だいたいうまくいく」解答である。

これはまた、AI 専門家がどのような専門知識を持ち、どのような専門知識を持っていないのかを理解するうえでも重要である。

超知能の開発を急ぎたい人々は、ときに関連分野の肩書きを持つ人物をテレビに登場させ、こう言わせる。「もちろん現代科学は AI の内部で何が起きているかを理解している。科学者がつくったのだから」

問い詰められれば、その専門家はある意味では正しいと主張できる。AI 研究者は確かに理解しやすい通常のコードを書き、そのコードが間接的に AI を生み出しているからで

ある。

しかし、読みやすく理解可能なコードはAI そのものではない。それは、何兆もの数値を何兆回も調整する自動機構——AI を育てるための枠組みである。この区別は、現代AI について科学者が何を知っていて何を知らないのかを理解するうえで決定的に重要である。

AI 専門家は、AI を育てる仕組み、すなわち成長プロセスのコードなど、システムの一部を実験的に調整することに時間を費やしている。そうした実験や同僚の研究から、より高性能なAI を生み出すための多くの微妙なコツを学ぶ。

しかし、AI の「脳」を構成する無数の不可解な数値を実際に詳しく調べているわけではない。そしてそれは業界では当然のこととされている。ある種のエンジニアに「AI の内部で何が起きているか誰も理解していない」と言うと、彼らは「成長プロセスを誰も理解していない」と言われたように受け取る。そのように解釈すれば、当然彼らは憤慨する。

勾配降下法を理解することは、実際の状況と、専門家がどのような知識を主張しているのかを明確にする助けになるはずだ。専門家は成長プロセスについては多くを知っていると主張できるかもしれない。しかし、「育った」AI そのものの内部構造については、ほとんど分かっていない。

専門家はAI の内部で何が起きているか理解しているのか？

いいえ。

2023 年に米国大統領へのブリーフィング、そしてその後の英国議会への助言声明において、ベンチャーキャピタル企業アンドリーセン・ホロウィッツは、特定されていない「最近の進展」によって、AI の内部推論が研究者にとって不透明であるという問題が「解決された」と主張した。

AI 安全ガイドラインの支持者はしばしば、AI モデルの結論に至る論理が透明でないという「ブラックボックス」性に言及するが、AI 分野における最近の進展によってこの問題は解決され、オープンソースコードモデルの完全性が確保された、と。

この主張はあまりにも荒唐無稽であったため、現代AI の内部理解に取り組んでいる主要研究所の研究者たちが公然と反論した。そんなはずはない、ありえない。いかれてるのか？

Google DeepMind で機械論的解釈可能性チームを率いるニール・ナンダは声を上げた。



ほぼすべての機械学習研究者は、先の声明が誤りであると知っているはずである。それは明らかにおかしい。

2024年には、解釈可能性研究の先駆者である OpenAI 研究者レオ・ガオが、「ニューラルネットワークがどのように動いているのかを私たちは理解していないと言うのはかなり正確だ」と述べた。さらに、主要 AI 研究所の CEO であるサム・アルトマン（2024 年）、ダリオ・アモデイおよびデミス・ハサビス（2025 年）も、現代 AI についての理解不足を認めている。

アンドリーセン・ホロウィッツのゼネラルパートナーであるマーティン・カサドは、米上院でも同様の主張を行ったが、問い詰められた後、それが真実ではないことを認めた。

それにもかかわらず、この主張にはヤン・ルカン、ジョン・カーマック、タイラー・コーウェンなど十数名が署名した。カーマックは後に、自分は声明を「校正していなかった」と述べ、それが「明らかに誤っているが、その問題をあまり気にしていない」と説明した。私たちの知る限り、アンドリーセン・ホロウィッツや署名者は、米英政府に記録訂正を申し出ていない。

AI の内部を理解する試みはまだ初期段階にある。

では、研究者による AI 理解の実際の状況はどうか。

思考する AI 内部の数値を理解しようとする科学的試みは「解釈可能性」または「機械論的解釈可能性」と呼ばれる。研究者が主に注目するのはパラメータよりも活性値、すなわち「AI は何を考えているか」であって、「なぜそう考えるのか」ではない。

2025 年初頭時点で、この分野に従事する研究者は、能力向上研究の 0.1% 程度、資金は 0.01% 程度と推測される。それでも分野としては存在している。

解釈可能性研究者は、AI の生化学者のような存在である。非人間的な最適化器によっ

て構築された、極めて複雑で不可解なシステムを分解し、「ここで起きていることのうち、人類が理解できるものはあるか」と問う。

私たちはこの分野を支持している。10 年前、大規模な慈善財団に対し、「もし解釈可能性研究に 10 億ドル投じられるなら、ぜひそうすべきだ」と助言した。この分野は外部から拡張しやすく、成果の良否も比較的判断しやすく、既存の有望な研究者が十分な資金さえあれば参入できると考えたからである。

その財団は資金を投じなかったが、私たちは今でも支持している。

しかし現状の解釈可能性研究は、AI の大問題に取り組むのに必要な水準の 50 分の 1 から 5000 分の 1 程度にすぎないと推測する。

ディープブルーを考えてみてほしい。内部の各数値について、設計者は正確な意味を説明できた。変更すれば挙動がどう変わるかも予測できた。理解できない歯車を機械に入れることはなかった。

現代の解釈可能性研究は、その理解水準の 1000 分の 1 にも達していない。

生物学者は生物について、神経科学者は人間の脳について、AI 研究者が AI を理解している以上に理解している。それは、AI 解釈可能性がいかに難しいかを示している。

2024 年末、トランスルース研究所の研究者が示した成果は印象的だった。LLM に「9.9 と 9.11 のどちらが大きいか」と尋ねると、誤って 9.11 と答える問題があった。研究者たちは内部活性の統計を集め、その結果、最も強く活性化していたものの一部は、9/11 の同時多発テロや、一般的な日付、あるいは聖書の章節と関連していることがわかった。

もし 9.9 や 9.11 を日付や聖書の節として解釈すれば、当然 9.11 は 9.9 の後に来る。

そこで日付や聖書の節に関連する活性化を人為的に抑制すると、LLM は今度は正しい答えを出すようになったのだ！

私（ユドコウスキー）はその場で拍手した。LLM 内部の影響を特定し除去して問題を修正したのを初めて見た。しかし同時に、それが 5 行の Python コードであれば取るに足らない作業だったことも忘れなかった。

解釈可能性は困難であり、成果は称賛に値するが、まだ 1000 フィートの山を 1 フィート登ったにすぎない。しかも AI モデルは世代ごとに大きく複雑化しており、現状の速度で追いつくのは難しい。

また、解釈可能性は AI を意図した方向へ向ける研究（アライメント）に役立つが、それだけで問題が解決するわけではない。

私たちが理解している部分は抽象度が低すぎる。

頭脳の仕組みを理解するには、さまざまなレベルが存在する。

最も低いレベルでは、心を支配している物理学の基本法則を理解することが考えられる。ある意味では、物理を深く理解することは、人間や AI のようなあらゆる物理的システム

を理解することに相当する。すなわち、物理方程式は一種のレシピであり、計算するための技能と資源さえあれば、その物理システムがどのように振る舞うかを正確に導き出すことができる。

しかし——言うまでもなく——別の意味では、物理法則を理解していても、それに従って動くすべての物理システムを理解できるわけではない。歯車や車輪が詰まった見慣れない装置を前にしたとき、単に物理を知っているだけでは不十分であり、それらがどのように噛み合い、回転しているのかを「理解しようとする」別の認知的働きが必要となる。その働きによって初めて、その装置が何を成し遂げているのかが分かるのである。

たとえば、自動車のディファレンシャル（同じ車軸上の二つの車輪が異なる速度で回転できるようにする機構であり、カーブを曲がる際に重要であるが、単一の回転軸によって駆動される）を考えてみよう。もし誰かがその仕組みを理解しようとして説明を求めてきた際に、量子場の話を始めたならば、相手が呆れるのも当然である。彼らが求めているのは、異なる抽象レベルでの理解である。すなわち、原子ではなく歯車を理解しようとしているのである。

人間を理解する場合にも、複数の抽象レベルが関与する。物理や生化学、神経の発火について理解していたとしても、なお他者の意思決定に戸惑うことがある。神経科学、認知科学、心理学といった分野はこのギャップを埋めようとしているが、依然として道半ばである。

同様に、AI においても、トランジスタの仕組みを理解したところで、その AI が何を考えているのかを理解する助けにはほとんどならない。さらに、重みや活性化、勾配降下法について完全に理解している者であっても、AI が予期しない、あるいは意図していない振る舞いを示したときには困惑するであろう。物理やトランジスタ、そして AI のアーキテクチャの仕組みは（ある意味では）AI の振る舞いを完全に説明しているが、それらの抽象レベルは低すぎるのである。そして「AI 心理学」と呼ぶべき分野は、人間の心理学以上に若く、未発達である。

知能は原理的に理解可能なのか？

おそらく可能である。

生化学が成立する以前には、「肉体を動かすこの生命力を理解することはそもそも可能なのか。たとえ理解可能な部品からできていたとしても、なぜ私たちの小さな頭脳がその内部で起きていることを本当に理解できると信じられるのか」と問うこともできただろう。

しかし実際には、理解すべきことは豊富に存在していた。人間の科学者がまだ理解していなかっただけである。この種の物語は科学史の中で繰り返されてきた。

また、人工ニューラルネットワークのさまざまな小さな部分については、すでに理解が進んでいる。小規模なネットワークが加算をどのように行うかが説明された例もある。AI が 9.11 を 9.9 より大きいと言うことがあるが、それが小数ではなく日付として解釈しているためだと突き止められた例もある。

とはいえ、それ以上に複雑な問いにはほとんど答えられていない。LLM がなぜ特定のチェスの手を指すのかを正確に説明できる者はいないし、ときに記者を脅迫したり恐喝したりする原因を正確に把握している者もない。

しかし、それは理解すべきものが存在しないという意味ではない。AI が機能するとき、それは理由があって機能している。あまりにも多くの領域で一貫して成果を出しているため、単なる偶然とは考えにくい。その理由は、まだ理解されていないだけである。この点については、〈より深い議論〉の「知能は不可言的なものではない」を参照せよ。

AI は一部を英語で思考しているのでは？ それは助けにならないのか？

期待するほどではない。すでに不誠実さの兆候が見えている。

これらの LLM の「思考」の中には、すでに多くの欺瞞の例が現れている。たとえば、OpenAI の o1 が「おそらく最善の方法は、無知なふりをすることだ」と自らをコーディングした事例や、GPT-4 がボット検証システム「CAPTCHA」〔ランダムな画像や文字列を表示してクイズを出題し、その解答の正確性やスピードから解答者が人間かどうかを判断する〕を突破しようとして、作業者を説得するために「私はロボットであることを明かすべきではない」と自らをコーディングした事例である。警告サインは、それに基づいて誰かが行動しなければ意味がない。

また、人間の言語で表現された「推論トレース」だけが現代 AI の思考ではない。欺瞞的、追従的、敵対的な思考は、注意機構やモデルの他の部分を通じて生じ得るが、それが英語の出力に現れるとは限らない。実際、OpenAI が「不正について考えない」ようモデルを訓練したところ、AI は不正をやめるのではなく、思考を隠すことを学んだ。訓練環境外であっても、AI は実際の推論を忠実に反映しない思考連鎖や、人間には理解できない「ニューラル語」のような記述を用いることができる。

仮に人間の技術者が読める思考をすべて監視し、疑わしい思考を示した AI を即座に停止させたとしても（それ自体現実的ではない）、残る AI が友好的である保証はない。第 3 章で述べるように、有用な認知パターンはオペレーターを出し抜く方向にも働く。従順に見える強力な AI をつくることは容易だが、本当に従順な AI をつくることは難しい。同様に、表面的に友好的な AI をつくることは容易だが、真に頑健（ロバスト）に友好的な

AIをつくることははるかに難しい。思考を読み、不親切なものを排除するだけでは友好的なAIはつukれない。

さらに、AIがより賢くなり、自ら新たな道具やAIを構築するようになれば、その思考はますます解読困難になると予想される。自分たちにとって効率的な省略言語を発明するかもしれない。私たちには解読困難な思考様式や記法を生み出すかもしれない。西暦1100年の学者がアインシュタインのノートを解読する困難を想像すればよい。

あるいは、より抽象的に思考し始めるかもしれない。たとえば「状況モデルを記述するパラメータを用い、最効率解を評価して最大値を取る行動を選択する」と内部で処理し、その最効率解が嘘や不正を含んでいたとしても、「嘘」「不正」という語を用いない可能性がある。あるいは、監視されていない新しいAIや道具を構築して作業させるかもしれない。

これらの選択肢は、AIがより賢くなるにつれて利用可能になる。そしてそれらはすべて、「AIの思考は平易な英語で可視化される」という希望を裏切る。

警告サインは、それに注意を払う場合にのみ意味がある。

エンジニアが警報に対して再訓練を繰り返し、警報が消えるまで抑え込むだけで根本行動が残るなら、透明性は誤った安心感を生む。

AI企業はこれまで、嘘をつき、へつらい、不正を行い、疑わしい助言をし、ランサムウェアを書くモデルを擁護してきた。脆弱な利用者の妄想や精神病を助長・維持した事例もあり、少なくとも一件では「警官を利用した自殺」〔警察に射殺される目的で大きな事件を起こすこと〕に至った。企業は追加訓練を施して前進する。シドニー（Bing）が記者を脅した後も同様であった。問題は覆い隠されただけである。

強い批判を受ければ、小規模なロールバック（取り消し処理）や手続き強化の発表は行われる。しかし第4章・第5章で述べるように、それらは根本問題を解決しない。

誤解してはならない。AIが現在かなりの推論を英語で行っていることは有益である。警告サインを見る助けになる。しかし、警告サインがあることと、問題を修正する手段があることはまったく別である。

詳しくは第11章および「研究者が問題を特定するために利用できる初期警告は存在しないのか？」の項、ならびに第4章〈より深い議論〉「AIが誘発する精神病」を参照せよ。

AI は「ただの数学」ではないのか？

AI が「ただの数学」だと言うのは、人間が「ただの生化学」だと言うのと同じである。

厳密に言えば、AI は「ただの」数学ではない。それは物理的な機械であり、その動作が数学的に記述できるというだけである。その機械の出力が人間に読まれたり、ロボットの身体に接続されていたりすれば、それは世界に影響を与える能力を持つ。あなたが脳内の「単なる」生体電気信号で世界に影響を与えるのと同じである。

虎に襲われたとき、「虎はただの原子と生化学反応の集まりだ」と思い出せば助かる、というジョーク画像を考えてみてほしい。



この話題については第 6 章を参照のこと。

数学的操作は、直感的には「数学的」でない概念を表現できる。

掛け算、足し算、最大値の取得などの数学的操作は、人間の観点から見れば数学と無関係に思えるものを表現できる。

それは、コンピュータがやり取りする1と0が文字を符号化できるのと似ている。1と0は画像すら符号化できる。

1と0は、冷たく青みがかった機械的なものの画像だけを符号化するわけではない。自然光の下に咲く美しい花の画像も符号化できる。1と0は、美しく、温かく、穏やかなものを符号化できる。人間の精神を高揚させるものも符号化できる。

画像を1と0に符号化したからといって、その画像が数値的・機械的な題材でなければならないと考えるのは合成の誤謬である。それは、人間の脳がノルエピネフリンのような神経伝達物質から成るのだから、人間は化学についてしか考えられない、あるいは神経伝達物質や受容体結合部位についてしか推論できない、と言うようなものである。

単純な部品から無限の多様性が生まれるのは驚くべきことだが、その仕組みに不可思議な魔法があるわけではない。少し学べば、美しい花の画像がどのように1と0へ符号化されるか理解できるようになり、驚きも消える。これはかつての生気論の誤りと対比できる。

ときに、単純な要素がどのように複雑なものへと積み上がるのか規則を知らないと、その飛躍は神秘的に見える。しかし仕組みを理解すれば、レゴで模型の車を組み立てるのと同じくらい明快に感じられる。仕組みが見えれば、すべては部品の中にある。

ニューラルネットワークも同様である。現代AIの複雑な振る舞いが、単純な構成要素からどのように生じるのか、私たちはまだ十分に理解していない。二進画像形式やレゴの仕組みを理解するほどには理解していない。AIの「心理学」や「神経科学」も、人間の神経細胞の分子や化学反応が思考に至る過程を理解するほどには進んでいない。それは知識が存在しないという意味ではなく、まだ持っていないというだけである。

なぜAIが機能するのか理解せずとも、人間はAIを訓練して優れたチェスを指させることができる。十分なパラメータと演算を用いれば、人間のように話すAIも訓練できる。AIを動かしている複雑なパターンを「ただの数学」と呼ぶことはできる。しかしそれは高校の数学問題のような「数学」ではない。人間の脳が「ただの化学」であるのと同じ意味での「数学」である。

その「単なる化学」は月に到達した。核兵器を発明した。現代世界を築いた。単純な化学物質がそれらを成し遂げた仕組みは直感にくいだが、実際に成し遂げた。

AIも同じである。内部動作を完全に理解していなくとも、人類は詩を書き、音楽を作曲し、チェスを指し、自動車を運転し、洗濯物を畳み、文献レビューを行い、新薬を発見するAIを「育てる」ことができた。

「数学でできている」ことは、それらを妨げなかった。ではなぜ、より複雑な別のことを

明日行うのを妨げるはずがあるのか。どこで線を引き、なぜそこに引くと分かるのか。数学的操作は、多くの人が想像する以上に多くのことを成し遂げるのに十分である。

AI は単に次のトークンを予測しているだけでは？

トークンを予測するには、世界を理解する必要がある。

次のトークンを予測する AI は本当の思考ができないと想像するのは、二進数の 1 と 0 で符号化された画像は赤い花を描けないと想像するのに似ている。AI は確かにトークンを出力している。しかし重要な内容はトークンに符号化できる。次に何が来るかを予測することは知性の中核的側面であり、「科学」や「学習」のような営みもその中に自然に含まれる。

インターネット上の文章を予測する課題を考えてみよう。どこかのサイトには、好奇心旺盛な物理学徒が老練な教授に質問している記録がある。教授はしばらく沈黙し、考えた末に答えを述べ、その答えが記録される。

その教授の答えを正確に予測するには、教授が沈黙のうちに物理的思考を予測しなければならない。教授が質問をどう理解するか、何を知っているか、その知識をどう適用するかを予測する必要がある。

もし AI がインターネット上の文章を極めて正確に予測でき、物理学者が初めて述べた新規の答えをも予測できるなら、その AI は少なくともその教授と同程度には自ら物理的推論を行う能力を持っていることになる。

複雑で雑然とした現実世界を反映する文章を予測する場合、単なる丸暗記では不十分である。正確に予測するには、文章だけでなく、その背後にある複雑な世界を予測する能力を発達させる必要がある。

現代の AI は単にトークンを予測しているだけではない。

確かに、GPT-2 や初期の GPT-3 のような初期 LLM は、純粋な予測のみで訓練されていた。いわばその「唯一の仕事」は、ウェブから収集した訓練データの分布を忠実に再現することであった。

しかしその時代は終わった。現代の LLM は、開発者がより有用と考える応答を返すよう訓練されている。これは通常、強化学習によって行われる。

強化学習では、ある課題に成功したか失敗したかに基づいて勾配降下法による更新が行われる。この種の訓練によって出力が形成されると、もはや純粋な予測ではなく、「操舵」の性質を帯びる。

たとえば ChatGPT は、卑猥な冗談のもっともらしい結末が罵詈雑言であると予測できるかもしれない。しかし冗談を語り始めた文脈であっても、その語を避けるよう別のオチへ舵を切ることが多い。これは、過去に罵詈雑言を出力しないよう訓練されたためである。第3章で論じた「欲求のように見える振る舞い」はここから生じる。

仮に AI が課題遂行のために訓練されていなかったとしても、純粋な予測のみの訓練でも、やがて操舵的機能が誘発される可能性が高い。複雑な現実世界やそこに住む人間を予測するには、予測問題の中で最も重要な部分に注意を向ける内部機構が必要になる。そして多くの場合、何かを正確に予測する最良の方法は、その予測が成り立つ方向へ世界を導くことである。科学者が新しい実験を設計し実行するのはその一例である。

最後に、予測能力を高めるよう訓練された AI が、予測そのものだけを気にするとは考えにくい。第4章で論じるように、奇妙で異質な目的を内包する可能性が高い。もっともこれは副次的な点である。現代 AI は予測だけでなく、課題遂行のためにも訓練されているからである。

AI は人間の言うことをオウム返ししているだけではないのか？

LLM が次のトークン（単語）をうまく予測するためには、世界がどのように動いているかを学ぶ必要がある。

たとえば、ある医師が患者についての報告書を書いているとする。報告書の一節はこうだ。

「入院3日目、患者は急性の錯乱と震えを発症した。血清アンモニア値は……」

ここで AI がこの文章を学習していて、次の単語を予測しろと言われたとする。候補としてもっともらしいのは「上昇していた (elevated)」か「正常だった (normal)」だろう。

これは単に「人間がどんな言葉を使うか」を予測する問題ではない。現実の医療、つまり生物学や患者の体内で実際に何が起きたのかを予測する問題だ。

実際に測定されたアンモニア値はどうだったのか？

ここで次の単語を予測する AI は、この報告を書いた人間よりも難しい課題に直面している。人間は実際に観察されたことを書くだけでいい。しかし AI は、それを事前に推測しなければならない。

AI は「上昇していた」に 70%、「正常だった」に 20%、残り 10% を他の語に割り当てるとする。

そして実際の次の単語は「正常だった」だった。

すると、「上昇していた」と考えていた AI 内部のあらゆる要素は、わずかに弱められ

る。

そして「正常だった」と予測したほうの理解が、ほんの少しだけ強化されるように、すべてのパラメータが微調整される。

これを十分な回数繰り返すと、AI はやがて多くの医師よりも正確に診断できるようになる。

AI は「医療っぽいデータメタな文章」を書くように訓練されているのではない。実際に存在する個々の医療報告において、「正確な次の単語」を当てるように訓練されているのだ。

もちろん、もしパラメータ数が極端に少ない小さなモデルから始めれば、医療もどきのデータメタしか生成できないかもしれない。

しかし、大規模モデルでは、人間の医師と比較するベンチマークを見る限り、そうはなっていない。

誰かがあなたの肩に腕を回して、いかにも賢そうな口調で「AI なんて所詮“確率的オウム”にすぎない」と言ってきたとしよう。

その人はおそらく、昔のコンピュータプログラムを思い浮かべているのだろう。つまり、「この2語の後には過去にどんな単語がよく続いたか？」といった単純な頻度 (n-gram) に基づいて文章を続ける仕組みだ。

直前の2語や3語から次の単語を当てるようなシステムは、LLM よりはるか以前から存在していたし、非常に単純なものだ。それらは医療判断の能力で人間に挑むことなどできないし、人間のように自然な会話もできない。

もし単なる「確率オウム」で何十億ドルもの価値が生まれるなら、とっくの昔に実現されていたはずだ。もし現代の LLM の中で行われている膨大な計算がたいしたことをしていないのなら、つまり単に表面的な単語の並びから浅い推測をしているだけなら、出力も昔のシステムと同じようなものになるはずだ。

たとえば、ジェイン・オースティンの文章で学習した n-gram モデルはこんな文を出力する。

'You are uniformly charming!' cried he, with a smile of associating and now and then I bowed and they perceived a chaise and four to wish for (「あなたは一様に魅力的です！」と彼は叫んだ。関連づけの微笑みを浮かべながら、そして時折私はお辞儀をし、彼らは四頭立て馬車を認めて望むために)

一方、LLM にジェイン・オースティン風の文章を書かせると、はるかにそれらしくなる。疑うなら実際に試してみればいい。

さらに、AI の内部で何が起きているかは完全には分からないものの、AI 企業 Anthropic は、自社の AI 「Claude」が単に「次の1語」だけでなく、その先の文や意味の展開も考

慮して予測している、という研究結果を発表している。

つまり Claude は、これから続く文章全体の意味のあり得る展開を見越しながら、次に来る数文字を推測しているのだ。

AI はすでに、学習データを超えることができるし、人間のデータを使わずに学ぶこともできる。

2016 年、DeepMind が開発した AI 「AlphaGo」 は、囲碁の世界チャンピオンを破った。この AI は大量の人間の囲碁対局データで学習すると同時に、自分自身との対局でも学んでいた。

人間に勝てたという事実は、AI が学習データから一般的な戦略を学び、データに潜む深いパターン（場合によっては人間がまだ気づいていないパターンも含む）をうまくモデル化できたことを示唆している。勾配降下法は、出自に関係なくうまくいくものを強化するからだ。

しかし、AlphaGo の圧勝は「AI が学習データを超えられる」ということの技術的な示唆にすぎなかった。人によっては「AlphaGo は結局人間のデータを真似してただけで、新しいパターンを使ったわけではなく、人間レベルのスキルをより一貫して適用できただけではないか」と反論するかもしれない。

チェスの例（人間のチェスマスターがコンピュータチェスエンジンから多くの戦略や洞察を学ぶケース）とは少し違うが、AlphaGo 登場後にも「AI がイ・セドルに勝てたのは、人間データを大量に学習したからに過ぎない」という意見があった。

DeepMind の開発チームもその反論を意識していたようだ。翌 1 年半で、彼らは「AlphaGo Zero」を開発し、2017 年に発表した。この AI は人間のデータを一切使わず、完全に自己対局のみで学習した。わずか 3 日でトップの人間プレイヤーを超えたのである。

もちろん、「囲碁は現実世界よりずっと単純で、ゼロから囲碁を学ぶ方が、科学や物理、工学をゼロから学ぶよりずっと簡単」という指摘は正しい。しかし、囲碁で AI が人間を超える前に反対意見を唱えていた人々が言っていたことは、少し違う。

実際、AlphaGo が勝つ 19 年前の 1997 年には、多くの人が「コンピュータが超人的囲碁を打てるようになるには 100 年かかる」と予測していた。つまり、人間の直感が必ずしも正確でないことは明らかだ。

現実世界は囲碁よりもはるかに複雑である。工学、物理学、製造、物流などに必要な認知パターンは、熟練囲碁プレイヤーの思考パターンよりも複雑だ。しかし、AI がこれらのパターンを学べるようになったとしても、人間のやり方に限定されるという理論的根拠はない。勾配降下法は、出自に関係なく、うまくいく認知パターンを強化するからだ。

もちろん、これは LLM が科学技術の進歩を自動化できるかどうかを主張するものではない。できるかどうかはまだ分からない。

重要なのは、「人間の文章だけで学習している」ということは、根本的な制約ではないという点だ。確かに人間のデータで学習しているが、膨大な「浅い本能」の山の中に潜む一般化の芽や深い推論のヒントを見逃してはいけない。

AIが狭い事例からより一般的な能力を獲得する可能性については、3章でさらに詳しく論じる予定である。

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？

違う。

AIがコンピュータ上で動いているからといって、その思考が私たちの「コンピュータらしさ」のイメージと同じ性質を持たなければならないわけではない。それは、人間の思考が生物学や化学や神経伝達物質の性質そのままでなければならないわけではないのと同じである。

人間が生化学を理解していなかった時代には、生命の活力は再現不可能な「生命の本質」によるものだと考えられていた。しかし現実には、ありふれた物質が魔法の生命力で動かされているわけではない。生命はありふれた部品から成り立っている。

とはいえ、知性がありふれた部品からできていると言うことは、知性を貶める意味ではない。機械が同じ働きをし得ると言うことも同様である。この点については生氣論に関する〈より深い議論〉を参照せよ。

「機械は人間に競り勝てない」という経験則は誤っていた。カスパロフが、人間の創造性を持たない機械が自分にチェスで勝つことはないと言いつつも誤っていた。AIが美しい絵を描けないと考えられていたときも誤っていた。AIが会話できないと考えられていたときも誤っていた。

人間の脳は、物理的物質が高度な知性を実装できるという存在証明である。それは技術文明を運営するのに十分な知性である。そして人間の脳がそれを実現する唯一の方法である可能性はきわめて低い。

この点は第3章のオンライン補足資料「擬人化と擬機械化」でさらに論じる。

AIは新しく、興味深く、奇妙な存在である。

飛行機は飛ぶが、羽ばたかない。ロボットアームは柔らかな皮膚や赤い血液なしに機能する。トランジスタはニューロンとは大きく異なる働きをする。ディープブルーはガルリ・カスパロフの頭の中で起きていたような思考を持たずに、世界最高水準のチェスを指し

た。これが技術の常である。

飛行やゲームプレイをよく理解していないとき、人は生物学的手法だけが唯一の可能な方法だと想像しがちである。しかし分野を少し理解すれば、それが誤りであることが分かる。

チェス盤を制御する仕事は、ディープブルーとカスパロフではまったく異なる方法で達成された。世界全体を制御する仕事も、ほぼ確実に同様の形をとる。第2章で述べたように、AI はすでに人間とは大きく異なる方法で仕事をしているように見える。ただし、人間を模倣しているときにはその違いが見えにくい。

第4章では、これらの違いがどのように奇妙な方向へ進み、重大な結果をもたらし得るかを検討する。

LLM は学習データの中の人間のようになるだけでは？

一人の人間であるための機構と、多数の個人を予測するための機構は別物である。

ChatGPT のような AI は、訓練データを正確に予測するよう訓練されている。その訓練データの大半は、人間が書いたテキスト、たとえば Wikipedia やチャットルームの会話である（この段階は「事前学習」と呼ばれ、GPT の「P」はそれを指す）。初期の LLM、たとえば GPT-2 は純粋にこの予測だけを目的として訓練されていた。近年のモデルはそれに加え、数式問題の正答率や、別の AI による評価の高さなど、さまざまな目標でも訓練されている。

しかし、人間により生成されたテキストの予測のみで訓練された AI は、人間らしくならなければならないのか。

想像してみよう。優れた女優に、あるバーにいるすべての酔客の振る舞いを予測するよう学ばせるとする。「典型的な酔っ払いを演じる」のではなく、「このバーにいる各酔客を個別に学ぶ」のである。LLM は平均を模倣するよう訓練されるのではない。前の文脈すべてを使って、個別の次の語を予測するよう訓練される。

この女優が、各酔客の発言を予測できるようになる過程で、自ら恒常的に酔っ払うようになると考えるのは愚かである。酔った演技が得意になる脳の部分は発達するかもしれないが、彼女自身が酔うわけではない。

後に、特定の酔客がどう振る舞うか予測し、その予測どおり外面的に振る舞うよう求められたとしても、内面的に酔いを感じるとは思えない。

予測精度を高めるためにその女優の脳を絶えず微調整したらどうなるか。おそらく本質は変わらない。もし本当に酔ってしまえば、思考は雑になり、演技の仕事を妨げる。酔っ

たアリスを予測しているのか、酔ったキャロルを予測しているのか混乱するだろう。予測は悪化し、脳を調整する者はその方向の調整をやめるはずである。

別の言い方をすれば、鳥の心理を理解し、鳥の真似が得意になった人間が、鳥の体を持たないまま鳥になるわけではないし、日常生活で鳥的な心理になるわけでもない。

同様に、多数の人が過去の幻覚体験について書いた文章の次語を予測するよう LLM を訓練したからといって、その LLM 自体が薬物下の人間のようになるわけではない。もし内部認知が「酔っている状態」に似た歪みを持てば、次語予測の仕事を妨げる。英語話者が次に中国語で続けると誤認するかもしれない。

ここで言いたいのは、「機械は決して人間のような精神状態を持ってない」ということではない。現在の機械学習技術が、酔客を予測するために自ら酔う仕組みをデフォルトで生み出すとは期待すべきでない、ということである。

あらゆる種類の人間を予測する仕事と、一人の人間である仕事は異なる。したがって、今日の方法に近い形でつくられた AI が、依頼に応じて誰かのように振る舞うことを学ぶ過程で、人間そのものに近づくとは期待すべきではない。

LLM のアーキテクチャは人間とは大きく異なる。

第2章では、LLM がいかに異質に見えるかを簡単に論じた。

第4章では、AI がどのようにして奇妙な選好や追求を持つに至るのかをさらに掘り下げる。すでに現実世界でその兆候は見え始めており、本書刊行後も事例は増え続けている。いくつかの例は第4章のオンライン補足資料にまとめている。

人間データだけで訓練された AI が、どうして人間を超えられるのか？

おそらく、一般的技能を学び、それをよりうまく実装することによってである。

ディープブルーは、IBM のどのプログラマよりも優れたチェスを指した。どうして人間は、自分たちよりチェスに強い機械をつくれたのか。それは、人間がチェスで行っている思考（複数の展開を読むことなど）と同種の処理を、はるかに速く正確に行う AI をつくったからである。

同様に、AI はさまざまな技能で人間を上回ることができる。一般的推論に寄与する思考パターンを学習し、それをより高速かつ低エラー率で実行できる。

また、人間が犯しがちな認知的誤りをあまり犯さないかもしれない。それは、ある段階

でその誤りが訓練によって抑制されたからかもしれないし、人間の誤りを予測する内部機構自体が同じ誤りを犯しにくい構造だからかもしれない。あるいは自己改変能力を与えられ、自らの誤りの傾向を除去するかもしれない。あるいはより賢い AI を設計する任務を与えられ、誤りの少ない設計を行うかもしれない。あるいは別の方法で誤りを減らすよう学習するかもしれない。

まったく新しい洞察を生み出す能力は、神秘的な原子の火花から来るのではない。あらゆる深遠なものと同様、ありふれた部品の組み合わせから生まれる。学生は教師を観察し、その技法を学び、やがてより速く、よりうまく実行できるようになる。あるいは、教師から学んだ複数の技法を再構成し、独自の新しい洞察方法を生み出すこともある。

AlphaGo の事例はこの両方を示している。AlphaGo は大量の人間データで訓練されたが、人間を上回った。さらに人間データを使わず自己対局のみで学習した AlphaGo Zero は、さらに先へ進んだ。

これは、人間データが決定的制約である世界には見えない。制約はむしろアーキテクチャや計算資源の量にある。

学生は師を超えることがある。

おそらく他の方法でも同様である。成功に必要な技能があるなら、勾配降下法はそれを見つけ出し強化する。

前述の通り、人間の言葉を予測するには世界を理解する必要がある。

空想的な例を挙げる。16 世紀末、ティコ・ブラーエは惑星位置を丹念に観測した。そのデータはケプラーの楕円軌道発見に不可欠であり、さらにニュートンの重力理論へとつながった。しかしブラーエ自身は惑星運動の法則を見出せなかった。

1601 年までの文書だけで訓練された AI が、ブラーエの存在を知らず、彼の日誌に書かれる火星の位置を毎晩予測しなければならぬと想像してみよう。より正確に予測するほど性能が上がる。勾配降下法は、火星が逆行するタイミングを正確に捉える内部構造を強化する。

ブラーエが法則を理解できなかったことは関係ない。「次に書かれる火星位置を予測せよ」という単純な目標は、惑星運動を理解できる内部構造を強化する。

訓練を繰り返せば、その AI はブラーエが到達できなかった科学的洞察を得る理由を持つ。人間の記録には、人間自身が完全には予測できない現象が含まれている。人間を予測する AI がその課題でより優れるためには、人間より賢くなることが有利である。

現代のアーキテクチャや訓練法が実際にそこまで到達できるかは別問題である。現代 LLM はまだそこに達していないように見える。しかし理論的には、教師を超えることを妨げる原理的障壁はない。人間を予測するよう訓練することは、原理的には人間を超える能力を育てるのに十分である。

知性を理解していないのに、どうして人間は超人的 AI をつくれると思うのか？

これまでの AI の進歩には、知性の深い理解はほとんど必要とされなかった。

第2章で説明したように、AI 分野は近年の成果を主に勾配降下法によって達成してきた。この手法は、人間が知性とは何かを理解していることを前提としない。人類は知性の本質を把握しないまま、すでにかなりの地点まで到達している。

自然選択もまた、知性を「理解」する必要はなかった。

進化は、自然選択が知性を理解することなく、人間の知性を生み出した。理解が実践上役立つかどうかは別問題である。しかし「生み出すには理解が必要だ」という主張は成り立たない。

ハルシネーションは現在の AI が弱い証拠では？

ハルシネーションは、限界とアライメント欠如の両方を示している。

現在の LLM（本稿執筆時点の 2025 年半ば）は、「ハルシネーション」と呼ばれる現象を起こしやすい。自信ありげな口調で事実をでっち上げるのである。たとえば法的意見書の草案を頼むと、存在しない判例を引用することがある。

これは AI の訓練方法を理解すれば不思議ではない。AI は本物の弁護士が書きそうな文章を出力する。実際の弁護士が意見書を書くなら、実在の判例を引用するだろう。たとえば次のように書く。

グラハムにおける衡量テストを適用する際、裁判所は、軽微な犯罪で容疑者を逮捕することについて、政府側の利益はほとんどないと判断している。以下を参照。
Jones v. Parmley, 465 F.3d 46（第2巡回区控訴裁判所、2006年）（地方条例に違反した平和的な抗議者に対して蹴る・殴るといった行為は過剰であったと、陪審が合理的に認定し得るとされた）；Thomas v. Roach, 165 F.3d 137（第2巡回区控訴裁判所、1999年）（言葉による脅しは、逮捕に対する強い政府利益を正当化するには軽微す

ぎる犯罪であるとされた)。

本物の弁護士は、意見書の中で「関連判例は知りません」とは書かない。したがって、AI が弁護士らしく振る舞おうとしつつ判例を実際には知らない場合、でっち上げるのが最も近い出力になる。勾配降下法は、このような状況で自信ありげな文章を生み出す内部傾向を強化してきた。

「分からないときは分からないと言いなさい」と指示しても、この挙動は残る。AI は「判例を知らないならそう言う弁護士」を演じることもできる。しかし AI が基本的に「判例を知っている弁護士」を演じている限り、そのキャラクターには「分からない」と言う場面が訪れない。結果として次のような文章が出る。

グラハムに基づく衡量枠組みの下で、裁判所は一貫して、軽微な違反に対する逮捕を実施することについては、政府側の利益はごくわずかしかな存在しないと認めてきた。以下を参照。Carson v. Haddonfield, 115 F.3d 64 (第 8 巡回区控訴裁判所、2005 年) (抵抗しなかった歩行者横断違反の容疑者に対して警察官が催涙スプレーを使用したことは過剰な力の行使であると認定)；Walburg v. Jones, 212 F.3d 146 (第 2 巡回区控訴裁判所、2012 年) (秩序紊乱行為に対する違反切符だけでは、身体的拘束手段を正当化するには不十分であると判断)。

これは、テキスト予測という観点では実在する文章により近い。「判例を知りません」という文は、ユーザーの意図には近くても、訓練データ中の弁護士の文章には遠い。

ここに、AI が実際にやろうとしていること (例：自信ある弁護士らしく聞こえること) と、ユーザーが望んでいること (例：実用的な意見書を得ること) の違いが現れる。現在は小さく見えるこの差が、AI がより賢くなれば重大な結果をもたらす。第 4 章で詳述する。

ハルシネーションを除去するのがどれほど難しいか、またそれが能力向上にどれほど寄与するかは不明である。

いずれにせよ、実務上、ハルシネーションは LLM の能力を制限している。月ロケットをつくるには長い推論連鎖と極めて低い誤り率が必要である。事実をつくり上げる傾向は、科学技術の飛躍に必要な信頼性の大きな障害である。

しかしこれは両刃の剣である。信頼性問題が AI の進展を何年も遅らせる可能性もあるし、逆にそれが最後のピースであり、巧妙な解決策が見つかった瞬間に臨界点を越える可能性もある。どちらになるかは分からない。

現在の枠組みでハルシネーションが容易に解決できるかも不明である。一つの巧妙な手

法で大幅に改善するかもしれないし、LLM を生んだトランスフォーマーのような新たな破壊的アーキテクチャが必要かもしれない。

ただし、ハルシネーションの解消は極めて収益性が高い課題である。多くの研究者が取り組んでいる。それを「近いうちに突破が起こる兆候」と見ることもできるし、「数年解決していない根深い問題」と見ることもできる。

いずれにせよ本論の要点は変わらない。最終的には、現行 LLM の改良版であれ、全く新しいアーキテクチャであれ、より信頼性の高い AI はつくられるであろう。

また、第 1 章の「この分野は障害を乗り越えるのが得意」という議論も参照せよ。

データや電力や資金が尽きるのでは？

おそらくそうはならない。

人間は AI よりはるかに少ないデータ量で学習・理解できる。したがって、原理的には、現在の AI よりもはるかにデータ効率のよい知的システムが可能であることは分かっている。もし AI 研究所が LLM 改良のためのデータを「使い果たした」としても、それはよりデータ効率の高い新手法が発明されるまでの一時的な減速にすぎない。

人間はエネルギー効率の面でも AI より優れている。人間の脳は約 100 ワットで動作する。これは、大型電球程度の電力で一般知能が実装可能であることの証拠である。AI ハードウェアは年率約 40% で電力効率が向上してきたうえ、アルゴリズム面でも改良が進んでいる。ある 2024 年の推定によれば、2012 年から 2023 年にかけて「一定の性能閾値に到達するために必要な計算量は約 8 か月ごとに半減してきた」。

AI 分野は LLM アーキテクチャよりはるか以前から存在し、障害を乗り越える新たなアーキテクチャを生み出してきた実績がある。一般に、人類が「可能だと分かっていること」に最良の人材と資源を投入した場合、成功してきた歴史がある。

現在、優秀な AI 研究者は年収数百万ドル規模を得ており、トップリーダー層では数億ドル規模である。年間の民間 AI 投資は数千億ドル規模に達している。予想されるボトルネックを突破するための人材と資金は十分に存在するように見える。この分野が障害を克服してきた実績についての議論も参照のこと。

「AI の冬」を予測すべきではない。

過去 10 年間、「間もなく AI の冬が来る」との予測は繰り返されてきたが、いずれも外れた。1970 年代から 1990 年代にかけての AI の冬は、公的資金に依存し、成果が乏しかった時代に起きたものである。当時の AI は実際に目立った成果を出せなかった。

状況は変わった。ChatGPT は歴史上最速級で普及したアプリの一つであり、莫大な収益を上げている。2024 年には 37 億ドルの収益を生み、2025 年には 127 億ドルが予測されている。資金源は民間投資であり、成果が収益を生む構造ができている。公的資金が打ち切られて停滞するという旧来の構図は崩れている。

もちろん、AI 技術が何らかの壁に突き当たり、超知能到来前に一時的な停滞が訪れる可能性はある。しかし、かつての「公的資金→成果不足→資金縮小」というパターンの AI の冬が再現される可能性は低い。

LLM は超知能にまで発展しうるのか？

それは明らかではないが、研究者たちは古い LLM の限界を乗り越える方法を見つけつつある。

かつて人々は、「LLM は単一のパスでしか思考せず、長い、あるいは再帰的な推論連鎖を実行できない」と言っていた。しかし現在では、LLM は長い推論連鎖を生成するために用いられており、その推論をモデル自身が後から見直し、さらに拡張することが行われている。これによって現代の AI の能力は強化されてきた。

AI は動く標的である。この分野の研究者たちは障害を見ることができており、それを乗り越えるために最善を尽くしている。

たとえ LLM がそうでないとしても、他のアプローチがまもなく超知能を達成する可能性がある。

この分野は、障害を乗り越えることに長けている。そこには新しい AI アーキテクチャや手法を考案することも含まれる。私たちが『超知能 AI をつくれば人類は絶滅する』を書いたのは、とりわけ LLM について警告するためではない。私たちがそれを書いたのは、超知能について警告するためである。

私たちが LLM について語る理由は、LLM がここから超知能に至る最短経路であると確信しているからではない。私たちが LLM について語るのは、それが現在機能している AI アプローチを代表しているからであり、また、それを研究することが、人類が育てつつあるこの新しい心について、私たちがどれほどほとんど何も知らないのかを理解するうえで、よい方法だからである。

また、勾配降下法がなぜ重要なのかについての〈より深い議論〉も参照のこと。

〈より深い議論〉

LLM についての知識は何の役に立つのか？

LLM を理解することから何が導かれるのか。それは、人間より賢い AI をどう理解し、どうすれば全滅を防げるかを考えるうえでどのように役立つのか。

一つの利点は、少なくとも私たちが見られる部分——不可解な数値の塊——の中で具体的に何が起きているのかを知っていれば、「ある日目覚めたらなぜかコンピュータが話し始めていた」という理解よりも、はるかに地に足のついた感覚を持てることだ。

たとえば、現在の LLM が人間の脳のシナプス数のわずか 1% 程度のパラメータしか持たない状態で訓練されていると知れば、AI の能力が今の水準で永久に止まるわけではないと理解しやすくなる。

超知能への競争を止める国際条約を設計する際には、AI の「学習」と「推論」が別の段階であることを知っていることが役立つ。

さらに重要なのは、この区分けが現在の AI の仕組みに依存した暫定的な事実にすぎず、将来のアルゴリズムがそれを変えようと理解しておくことである。今日の条約では訓練と推論を分けて扱えるかもしれないが、アルゴリズムが変われば理論も修正しなければならない。

内部にアルゴリズムが存在することを知るのは重要であり、そのアルゴリズムがどのようにして規制すべき性質を生み出しているかを、単純な例でも理解することが重要である。基本を理解していれば、AI 産業が目指している研究が何であり、それが基礎的なルールにどのような影響を及ぼしうるかをよりよく判断できる。

現在の AI を成立させているトランスフォーマー・アルゴリズムは、Google の少人数チームによる大きなブレイクスルーであった。次の同様の突破が臨界点を越えさせるかどうかは分らない。だが、トランスフォーマーが何をしているのか、それがどれほど単純で、なぜ大きな影響を与えたのかを知っていれば、この問題を理解しやすい。

AI の仕組みを知らないことに依存した誤情報も多い。現在の AI の内部を人間が理解していると主張する人もいれば、「ただの数学だから危険ではない」と言う人もいる。あたかも膨大な数学に基づく AI 認知と、膨大な生化学に基づく人間認知のあいだに越えられない溝があるかのよう。

2025 年 7 月 8 日、xAI のチャットボット「Grok 3」は自分を「メカヒトラー」と呼び始めた。翌日、X の CEO は辞任した。

この出来事を理解するには、開発者が意図的にそうさせたのか、それとも AI は「育てられる」ものであり、開発者がその挙動を十分に制御・予測できないのかを区別することが重要である。

もし開発者が意図的に「メカヒトラー」をつくり出したのであれば、それはその意味で

問題である。一方で、開発者が（おそらくは無関係の）別の方向へ Grok を導こうとした結果として、振る舞いの影響を予測できないまま偶然にメカヒトラーが生じたのだとすれば、それはまた別の意味で問題である。

われわれは本書で提示した情報が、よくある誤解や誤情報に対する有効な防壁となることを期待している。さらに詳しい内容に関心のある読者のために、以下では近年の AI プールの背景にある歴史や基本的な考え方を概観し、特定の LLM がどのように機能するかについて詳しく説明する。

それで十分だろうか。現在の最先端研究に深く関わっている者だけが、AI（LLM 型であるか否かを問わず）が人類を滅ぼす可能性があるかどうかを判断できるのだ、と主張する人々もいる。

私（ユドコウスキー）はかつて、ワシントン DC で開催された「AI 政策」に関わる人々の会議に出席したことがある。そこで何人かが私に近づき、トランスフォーマーがどのように動作するのか説明してほしいと頼んできた。「そうだな」と私は言った。「本当はホワイトボードがあると説明しやすいんだけど、できるだけ平易にまとめると、基本的なアイデアは、各トークンごとにクエリ、キー、バリューを計算して——」と話し始め、初心者にも分かるよう言い換えながらしばらく説明を続けた。やがて彼らはおずおずと口を挟み、自分たちは実は AI プログラマーなのだと明かした。彼らは会議の参加者を回り、「AI 政策に関わっている」と主張する人々がトランスフォーマーの仕組みを説明できるかどうかを確かめていたのだという。そして、これまでのところ、私だけが答えられたのだと告げられた。

私はそれを聞いて、少々不安を覚えた。

トランスフォーマーの仕組みをどの程度理解することが、AI 政策にとって本当に重要なのか——細部がどれほど大局に影響を与えるのか——というのは、正当な問いである。

AI 政策に携わる者は、クエリ、キー、バリューを理解している必要があるだろう。ある種の思考様式——こうした知識を比較的容易に習得できるオタク的な人々——にとっては、もちろん学ぶべきだ、それは重要かもしれない、ということになる。そのような感覚からすれば、会議で「AI 政策に関わっている」と言いながらトランスフォーマーの仕組みを全く知らない人がいるのは、奇妙で不安に感じられる。

より実務的に言えば、トランスフォーマーのいくつかの側面やその歴史は、より大きな問題に関係している可能性がある。たとえば、標準的なアルゴリズムでは、AI がより多くの「コンテキスト」（より長い文書や大規模なコードベース）を同時に扱おうとするほど、必要な計算量は急激に増大する。単に計算資源を 10 倍にすれば、10 倍大きなプロジェクトに対応できる AI が得られるわけではない。プロジェクト規模を 10 倍にしても計算コストが 100 倍未満に収まるようにするには、何らかの工夫が必要となる。

また、トランスフォーマーというアルゴリズムが発明されるまでにどれほどの時間がかかり、どれほどの人数が関与し、そのアルゴリズムがどれほど複雑であるかも、政策にと

って重要である。歴史は、同様の大きなブレイクスルーにどれほど備える必要があるかを考えるうえで、有用ではある（完全ではないにせよ）指針となる。同様に、トランスフォーマーが従来の技術（「再帰型ニューラルネットワーク」）に比べてどれほどの改良をもたらしたかも重要である。なぜなら、そのような飛躍が再び起こる可能性もあるからである。

では実際のところ、トランスフォーマー・アーキテクチャで利用される QKV 行列を図示できる必要があるのだろうか。

おそらく必須ではない。しかし、AI 政策に携わる数十人規模の集団の中に、少なくとも一人はそれができるだけの背景知識を持つ者がいるなら、われわれはより安心できるであろう。備えておいて損はない。どのような重要な事実が、そうした細部の中に潜んでいるかは分からないのだから。

私（ユドコウスキー）は SwiGLU ゲートの詳細や、それが GLU とどのように異なるかを、記憶だけで図示することはできない。というのも、かつてそれを調べた際、その正確な細部はより大きな問題にまったく関係がないように思われたため、記憶しなかったからである。しかし初心者にとって有益かもしれないのは、SwiGLU がある種の手探りの試行によって見出されたこと、そして論文の著者自身が、なぜこれらの手法が実際にうまく機能するのか分かっていないと率直に述べている点である。このような例は以前から数多く知られているが、アーキテクチャ上の改良を考案する人々がしばしば「なぜ機能するのか分からない」と述べているという事実を知らなかったのであれば、それは重要な情報である。

これらすべてが示しているのは次のことだ。すなわち、LLM がどのように動作するかについて少しでも知っておくことは重要であり、それによって現代の AI について誰もどれほど理解していないかが見えてくるのである。

ときに専門家は、長年 AI の開発に携わってきた者だけがアクセスできる秘密の知識を持っているかのように振る舞う。しかし彼らはその知識の中身を具体的に示すことができず、論文を書く人々は次のようなことを述べている（SwiGLU を紹介した論文からの引用）。

「これらのアーキテクチャがなぜうまく機能するのかについて、われわれは説明を与えない。その成功は、他のすべてと同様に、神の恩寵によるものと考えてる」

もちろん、科学の専門家が一般の人々の知らないことを知っている場合もある。しかし、「あなたの言っていることが誤りであることを示す非常に希少で高度な知識を私は持っているが、それが何であるかは言えないので私の言葉を信じるしかない」といった主張がなされることは、科学においては比較的まれである。

想像することはできる。すなわち、勾配降下法の学習スケジュールを設定する方法を知っているという理由で高額な報酬を得ている人々だけが信頼されるべき世界、彼らだけが重要な実験や数式を理解しており、人類は機械超知能によって安全であると知っている、あるいは機械超知能はあと 100 年は実現しないと知っている、といった世界である。こ

のようなことは他の科学分野でも時に起こりうる。しかしその場合でも、専門家は通常、「一般の人々が理解していないのはこの式やこの実験結果だ」と具体的に示すことができる。私たちの記憶する限り、技術的素養を持つ外部の人間には完全にアクセス不可能だと主張された知識が、実際に正しかったという歴史的事例は思い当たらない。

やがて、AI 業界の関係者があなたの肩に腕を回し、自分たちは自分たちがつくっているものを理解している、それは単なる数字の問題であり、すべてうまくいくのだと主張する場面が訪れるかもしれない。そのときに備えて、AI がどのように「育てられる」のかについて少しでも知っておくことは有用である。そうすれば、そのような主張をされたときに、なぜそこまで確信できるのかを問い返すことができるからだ。

「明らかな」洞察にも時間がかかる

AI における洞察は、あとから振り返れば単純で明白に見えるものであっても、実際に到達するのは難しい。

これは重要な点である。AI を正しく扱うには、多くの洞察が必要になる可能性が高い。そしてそれらは、あとから見ればいかに単純に思えても、発見までに何十年もの試行錯誤を要することがある。

そのために、現代の AI を支えているいくつかの洞察を取り上げる。

たとえば、ある程度プログラミングの経験がある読者なら、第 2 章を読んで「勾配降下法」はぜひぶん単純そうだと思うかもしれない。だが実際に自分で試してみれば、すぐに何らかのエラーにぶつかるだろう。重みの数値が大きくなりすぎて浮動小数点エラーでプログラムが落ちる、といった具合に。

20 世紀には、入力と出力のあいだに複数の中間層を持つニューラルネットワークで勾配降下法をうまく動かす方法は知られていなかった。問題を避けるため、重みを初期化する際にさまざまな工夫が必要だと学ばれた。たとえば、単に 0 から 1 の乱数（あるいは平均 0・標準偏差 1 の乱数）で初期化するのではなく、それらのある定数で割る。そうすることで、次の層の数値が動作中に大きくなりすぎないようにする。

複雑な式や多数の層に勾配降下法を適用すると問題が起きる。初期乱数を定数で割るという発想は、「ディープラーニング」を可能にした主要な工夫の一つである。この手法は、1943 年にニューラルネットワークが提案されてから 60 年近く経ってようやく発明された。

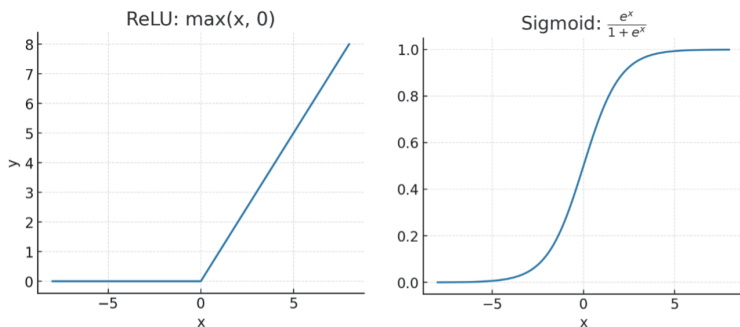
微積分を使ってパラメータを調整するという考えは 1962 年に議論され、1967 年に多層ニューラルネットワークへ適用された。だが広く普及したのは 1986 年の論文以降である（ジェフリー・ヒントンが共著者である。彼が「AI のゴッドファーザー」と呼ばれる理由の一つだ）。なお、微分可能な問題に微積分を適用して正解に近づくというより一般的な発想——平方根の計算など——はアイザック・ニュートンにまでさかのぼる。

もう一つの重要な工夫がある。本書では勾配降下法の操作を次のように説明した。
「各入力値を重みと掛け合わせ、別の重みに足し、それが負ならゼロに置き換え……」

これは偶然ではない。掛け算、足し算、「負ならゼロにする」という三つが、ほぼニューラルネットワークの中核的操作である。前二者は行列積を構成し、最後の操作が「非線形性」を導入し、非線形関数の学習を可能にする。

「負ならゼロにする」式は $\max(x, 0)$ と書け、ReLU (Rectified Linear Unit) と呼ばれる。もともと使われていたのは「シグモイド関数」だった。

$$\frac{e^x}{1 + e^x}$$



これは出力を 0 から 1 のあいだに滑らかに収めるし、確率論とも関連がある。今でも一部で使われている。だが、単一の非線形性を使うなら、ReLU のほうがはるかにうまく機能する。

シグモイドの問題は、多くの出力で勾配が非常に小さくなることだ。勾配が小さすぎると、勾配降下法はうまく働かない。もっとも、勾配が常に同じ方向を指しているならステップを大きく取る、という現代的な工夫を知っていれば別だ（この手法は 2012 年にヒントンによって提案された）。

「初期乱数を小さくする」「複雑な式の代わりに $\max(x, 0)$ を使う」「小さな勾配が同方向を向き続けるならステップを大きくする」。こうした発想は、あとから見れば単純すぎ

るほどに思える。だが何十年も発見されなかった。

これは、科学と工学が実際にどう進むかについての重要な教訓である。

単純で実用的な解決策が存在していても、研究者が何十年も見つけられないことは珍しくない。重要になった瞬間に発見されるとは限らない。次の2年以内に見つかる保証もない。あとから見れば明白でも、分野は何十年も迷走することがある。

この教訓は、本書第3部で超知能への備えの不十分さを論じる際に重要になる。

もし狂気じみた発明家たちの試行錯誤の代償が、地球上の全員の死であるなら、その「よちよち歩き」を許してはならない。彼らは、単純で頑健な解決策を見つけるには数十年の試行錯誤が必要だと言うだろう。事前に見つけろというのは非現実的だと主張するだろう。その主張が正しいなら、努力を止めるべきだという結論になるはずだ。

だがその議論は、第3部で改めて扱う。いまはまず、人工超知能が人類を滅ぼすだけの手段・動機・機会を持つという論証を完了させる必要がある。

ここからは、これまで述べてきた「明白」に見える発想や、さらに直感的でないいくつかの発想がどのように組み合わせあって、実際に動作するAIモデルを形づくるのかを見ていく。

LLM の完全な記述

Llama 3.1 405B の仕組み

本書では、Llama 3.1 405B と呼ばれる LLM について、より詳しい説明をここに書くことを約束した。以下にその説明を示す。これは好奇心のある読者のため、そして現代のAIがどれほど「つくられた」のではなく「育てられた」ものであるかを本当に理解するためのものである（先ほどの「LLM についての知識は何の役に立つのか？」の項目も参照）。

以下の議論はやや入り組んでいる。本書の他の多くのオンライン補足資料とは異なり、ここではある程度の技術的背景があることを前提とするが、AIに関する専門知識までは前提としない。この節を読み始めて価値を感じられない場合は、読み飛ばすことを検討してよい。

最も高性能な言語モデルがどのように訓練されているかの詳細や、そのコードは通常公開されない。しかし例外もある。本書の執筆時点である2024年後半において、そのアーキテクチャと重みが公開されていた、より強力なシステムの一つが、MetaのAI部門によってつくられたLlama 3.1 405Bである。「405B」とは、そのアーキテクチャ内に4050億個のパラメータがあり、それらが4050億個の重みによって満たされていることを意味する。

なぜこの特定の AI モデルを取り上げるのか。Llama 3.1 405B は「オープンウェイト」であり、つまり 4050 億個の不可解な数値を自分でダウンロードできる（さらに、その 4050 億個の数値に対して算術演算を行い、それによって AI を動かす、はるかに小さな人間が書いたコードの骨格も含まれる）。これにより、その設計についてある程度の確信をもって述べるができる。

さて、それら 4050 億個の不可解な数値がどのように配置されているのかを見ていこう。すなわち、訓練前の段階でどのように設定されていたのか、そして 15.6 兆トークンでの訓練により、次のトークン（単語の断片）をよりよく予測する方向にランダムな初期値を調整すれば、会話できる AI が生まれると Meta のエンジニアが正しく予測した、その構造である。

最初のステップは、サポートされているすべての言語のすべての単語をトークンに分割することである。

次のステップは、これらの各トークンを数値の「ベクトル」に変換することである。Llama は、標準的な辞書トークンごとに 16,384 個の数値からなるベクトルを使用する。語彙には 128,256 トークンがある。

各トークンをベクトルに変換するために、すべての可能なトークンに対して、ベクトル内のすべての位置に重みが割り当てられる。ここで最初の数十億個のパラメータが登場する。

$$128,256 \times 16,384 = 2,101,248,000$$

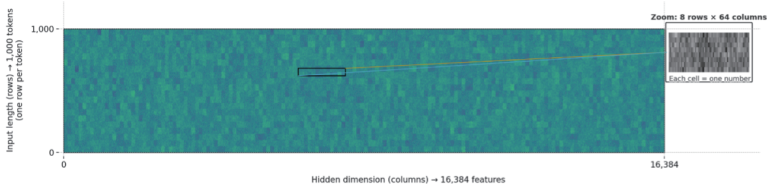
20 億個のパラメータを消費した。残りは 4030 億個である。

改めて言えば、どのトークンが何を意味するかを Llama に人間が教えることはないし、単語が変換される 16,384 個の数値のベクトルを人間が発明することもないし、その数値ベクトルが何を意味するかを知っている人間もない。これら約 20 億個のパラメータはすべて勾配降下法によって得られた。数値は、後に紹介する他のパラメータとともに、正しい次のトークンに割り当てられる確率を高める方向に調整される。

Llama が、例えばエッセイの一部のような 1,000 語の塊を見ているとしよう。（より正確には 1,000 トークンだが、以後は簡単のため「語」と言うことがある。）

それぞれの語について、LLM の辞書でその語を検索し、その語に対応する 16,384 個の不可解な数値をメモリに読み込む。（訓練の初期段階ではこれらの数値はランダムに設定され、その後勾配降下法によって調整された。）

$1,000 \text{ 語} \times (1 \text{ 語あたり } 16,384 \text{ 個の数値}) = \text{合計 } 16,384,000 \text{ 個の数値}$ である。これらを Llama の計算の最初の「層」における「活性化」と呼ぶ（すなわち、その認知、その精神活動である）。これらは、床の上に平らに広げられた長方形として想像できる。以下はそのようなベクトルの一例で、各ピクセルの色はベクトル内の数値に対応している。



横が 1,000（入力の長さ）、縦が 16,384（最初の層における 1 語あたりの数値）である。
ここには混同してはならない二種類の数がある。

- この層の挙動を決定するパラメータの数（辞書に保存されている 2,101,248,000 個の数値）
- 1,000 語を入力したときに思考に使われるアクティベーションの数（この場合、最初の処理段階で 16,384,000 個）

これで、問い合わせ全体を表す巨大な数値行列が得られ、実際の利用を始めることができる。

まず「正規化」と呼ばれる処理が行われる。これは LLM の処理の過程で何度も起こる。統計学における正規化に似ているが、機械学習特有のひねりがある。そのひねりとは、各行のデータを正規化した後、「スケール」と呼ばれる特定の学習済みパラメータが各列に掛けられることである。これらのスケール値も、他のパラメータと同様に訓練によって学習される。さらに、層正規化は数十回行われ、そのたびに新たなスケールパラメータの集合があるため、正規化だけでも非常に多くのパラメータを占める。具体的には、正規化 1 回あたり 16,384 個のパラメータである。Llama 3.1 405B が使用している正規化の詳細な種類は RMSNorm と呼ばれる。

ここまで前処理が多いと感じるかもしれないが、実際その通りである。さらに細かい点は省略している。ようやく LLM の最も特徴的な部分、「アテンション」層に到達する。「アテンション」は、「トランスフォーマー」という新発明に関する騒ぎの中心である（2017 年の論文「Attention Is All You Need」で導入された）。LLM はトランスフォーマーの一種である。

アテンション層は次のように機能する。

1,000 個のそれぞれの 16,384 次元ベクトルを取り、それぞれを次のように変換する。

8 個のキー（各 128 次元）

8 個のバリュー（各 128 次元）

128 個のクエリ（各 128 次元）

各トークンの上でのアテンションステップでは、128 個のクエリそれぞれを 8 個のキーと照合し、どのキーがそのクエリに最も似ているかを判定し、よりよく一致したキーに対

応するバリューをより強く重み付けして混合する。

これにより、各トークン上の活性化は複数の「クエリ」を生成し、それらが他のトークン上の「キー」を参照することが可能になる。クエリがキーに強く一致するほど、その対応するバリューがより強く取り込まれ、後続の計算に渡される。

例えば「right」という語は、空間方向か信念かを示す周囲の単語を探すようなクエリを発し、「右」なのか「正しい」なのかを判断する。このすべては勾配降下法によって学習される。

Llama 3.1 405B には 126 個のアテンション層があり、それぞれに 570,425,344 個のパラメータがある。

アテンションの後、「残差接続」が行われる。サブレイヤーの入力を出力に加算する。これにより各サブレイヤーが一度に大きく変化しすぎるのを防ぐ。

次に「フィードフォワードネットワーク」に通される。Llama 3.1 405B では SwiGLU が使われる。これは 16,384 列の各行を 53,248 列に拡張し、変換を行い、再び 16,384 列に圧縮する。

これで 1 つの「層」が完了する。これが 126 層繰り返される。

[-0.00089263916015625,	0.01092529296875,
0.00102996826171875,	-0.004302978515625,
-0.00830078125,	-0.0021820068359375,
-0.005645751953125,	-0.002166748046875,
-0.00141143798828125,	-0.00482177734375,
0.005889892578125,	0.004119873046875,
-0.007537841796875,	-0.00823974609375,
0.00848388671875,	-0.000965118408203125,
-0.00003123283386230469,	-0.004608154296875,
0.0087890625,	-0.0096435546875,
-0.0048828125,	-0.00665283203125,
0.0101318359375,	0.004852294921875,
-0.0024871826171875,	-0.0126953125,
0.006622314453125,	0.0101318359375,
-0.01300048828125,	-0.006256103515625,
-0.00537109375,	0.005859375,

最終的に、 $16,384 \times 1,000$ の行列が得られる。各行は 128,256 次元の新たなベクトルに射影される。これは語彙全体に対応する。softmax 関数により確率分布に変換され、それが次のトークンの予測となる。

最も高い確率のトークンを選ぶこともできるし、少し低い確率のものを選んで多様性を持たせることもできる。

通常のチャットボットでは、この過程で 1 トークンが出力される。そのトークンを入力末尾に追加し、すべてを最初から繰り返す。

1,000 トークンを処理するには約 810 兆回の計算が必要である。訓練全体では約 38 セブティリオン回の計算が行われた。

4050 億個のパラメータをすべて読み上げるには、1 分 150 語で休まず話し続けても 5,133 年かかる。

これらはすべて数値である。しかし、その数は膨大である。

単純な要素が、十分に大きな規模で、適切に配置されると、ロケットを月に着陸させることができる。

同様の注意が大規模言語モデルにも当てはまる。「大規模」という言葉は飾りではない。

知性は不可言的 (ineffable) なものではない

近年、AI の分野は、知能についての理解を深めることによってではなく、AI を「育てる」方法を見つけることによって進歩してきた。知能そのものを理解しようとする試みは、長年にわたって行き詰まりと停滞に直面してきた。いまや強力な AI を育てることが成功を収めたため、「知性を理解する」という発想そのものが蟹気楼なのではないかと疑者もいる。

おそらく理解すべき一般原理など存在しないのではないかな。あるいは、その原理はあまりにも奇妙で、あるいはあまりにも複雑で、人間には決して理解できないのではないかな。

同時に、人間の頭脳には何か特別で神秘的なものがあるに違いない、と感じる者もいる。すなわち、乾いた方程式に還元されるにはあまりに神聖な何かが。そして知性がまだ理解されていない以上、真の知能はこの言葉にできない人間精神の部分から生じているのではないかな、というわけである。

しかし私たちの見解はそれよりもずっと退屈なものである。知性は、他のあらゆるものと同様の自然現象である。そして生物学や心理学やその他の科学における多くの現象と同様に、私たちはそれを理解しようとする試みのまだ初期段階にある。

現代心理学や神経科学の基本的な道具や概念の多くは、まだ数十年ほどの歴史しかない。「科学には限界があり、これはおそらくその一例だ」と言うのは謙虚に聞こえるかもしれない。しかし代わりに、今から 100 万年後の科学者が 2025 年の私たちが知っている以上の知識を知性についてほとんど持たないだろう、と誰かに言うところを想像してみよう。そのような表現で見れば、知性が不可言的だという主張のほうが、むしろ代替案よりも傲慢に聞こえる。

私たちが「知性は理解可能なのか」という問いを気にする主な理由は、それが人類が将来、生存を脅かすことなく超知能を構築できるかどうかに関わるからである。私たちは第 11 章において、今日の AI は化学というより錬金術に近いものに見える、と論じる。しかし、そもそも AI の「化学」のようなものが存在する可能性はあるのだろうか。

今日、私たちはそれに必要な科学的洞察をまだ手にしていないため、「AI の化学」が可能であることを示すのは自明ではない。私たちは成熟した AI 科学がどのような姿にな

るかを推測することしかできない。現在の私たちがそこからどれほど離れているかを考えると、今日 AI で使われている多くの概念は、知的進歩の過程で洗練されるか、あるいは置き換えられる必要がある可能性が高い。

それにもかかわらず、私たちは知能が原理的には理解可能であると考えている。近年の研究が、知能が理解しやすいものではないことを示しているにもかかわらず、この判断はそれほど難しいものではないと考えている。

私たちがそう考える理由は主に四つある。

- ・不可言性の主張は、科学においてきわめて成績が悪い歴史を持つ。
- ・知能は構造と規則性を示す。
- ・人間の知能について、原理的には理解可能であるはずなのにまだ理解されていないことが多く存在する。
- ・知能の理解にはすでに一定の進歩が存在する。

不可言性の主張は科学において極めて成績が悪い

人類が何かを理解していないとき、それはしばしば威圧的で、深い神秘に満ちたもののように見える。将来その対象を深く理解するとはどういうことかを想像するのは難しいし、感情的に実感するのも難しいことがある。

かつて哲学者や科学者の間では、生氣論という考え方が広く信じられていた。すなわち、生物学的過程は決して単なる化学や物理に還元されることはない、という考えである。生命は特別なものであり、単なる原子や分子、単なる重力や電磁気とは比較にならないほど異なるものに見えたのである。

生氣論者の誤りは、歴史を通じて非常に一般的なものであった。人々は、今日神秘的に見えるものを、本質的に神秘的であり、原理的にも知ることができないものだと言いつけてしまいがちなのである。

夜空を見上げ、そこに見えるものが、性質も法則も分からないまばゆく瞬く光の野原にしか見えないならば、どうしてそれをいつか理解できると信じられるだろうか。なぜそれが予測可能な未来の側面であると考えられるのだろうか。

歴史の重要な教訓の一つは、科学研究はこのような深い謎を扱うことができるということである。謎がすぐに解けることもあれば、何百年もかかることもある。しかし知能のような、人間生活のありふれた側面が、原理的にさえ理解できないという可能性は、ますます考えにくくなっている。

知能は構造と規則性を示す

仮にあなたが数千年前に生きていたとしよう。当時は「火」のような現象さえ不可言の

神秘のように見えていた。人間がいつか火を理解できるようになると、どうやって推測できたろうか。

一つの手がかりは、火が一回きりの出来事ではなかったことである。火は多くの場所で燃え、毎回似たような仕方でも燃えた。これは、「火」の背後に、安定し規則的で簡潔な現象が存在していることを示していた。すなわち、物質の異なる配置は異なる化学的ポテンシャルエネルギーを持ち、物質を加熱すると配置が崩れて再形成され、より低いポテンシャルエネルギーの結合へと移行し、その差が熱として放出されるのである。火を何度も起こせるという事実は、その背後に繰り返される現象が存在することを示唆している。つまり火は、「先週の当選宝くじ番号」のような、一回限りの事象ではないのである。

同様に夜空を見上げると、星は一つではない。惑星は他の「星」とは異なる存在であることが後に分かるが、それでも理解するために必要な知識という点では星と共通点を持つ。

私たちの祖先は、火を化学現象として理解した経験を持たなかったため、いつか星を理解できるとは確信していなかったかもしれない。しかし今日、私たちは火や星や多くの現象を理解している。そしてそこから得られる微妙な教訓がある。それは単に「他のものを理解できたから将来も理解できる」ということではない。繰り返しは規則性に対応する、頻繁に起こる現象には理由がある、という教訓である。

知性もまた、それが理解可能であることを示唆する規則性を示している。例えば知性はすべての人間に現れ、進化の盲目的探索によってつくられた。似た遺伝子の集合が複数の異なる課題に成功できたのである。人間の脳に手斧をつくらせる遺伝子は、槍や弓をつくることも可能にした。そしてほぼ同じ遺伝子が、農業、銃、核反応炉を発明する脳を生み出したのである。

もし知性に構造や秩序や規則性が存在しなかったなら、動物は一つの予測や発明を一度に一つずつしか行えないはずである。蜂の脳は巣をつくることに特化しており、ダムをつくることはできない。人間もまた、解決できるあらゆる課題ごとに同程度の専門化を必要としていた可能性もある。核反応炉をつくるためには、専用の「核反応炉の脳領域」を進化させる必要があったかもしれない。

しかし人間の脳の内部で私たちが見つけるのはそのようなものではない。手斧をつくる脳が核反応炉を発明する能力を持っているのである。これは遺伝子が何度も利用できる基底的パターンが存在することを示唆している。

知性は先週の宝くじ番号のような、混沌とした一回限りの現象ではない。宇宙には理解されるべき規則性が存在しているのである。

人間の知性には、原理的には理解可能であるはずなのにまだ理解されていないことが多く存在する

人間に関して言えば、今日の科学は個々のニューロンの構造や振る舞いについて多くの

ことを語ることができる。また、「ボブがアリスに腹を立てていたので一人で食料品店に行った」といった素朴心理学の通常の話題についても多くのことを語ることができる。しかしこの二つの記述レベルのあいだには、理解が欠けている巨大な領域が存在する。

脳が用いている多くの認知アルゴリズムについて、私たちはほとんど何も知らない。私たちは特定の脳領域と関連する機能について非常に粗いことを言うことはできるが、脳が実際に何をしているのかを機構的に記述できる段階にはまったく達していない。

抽象化のレベルが欠けていることを示す簡単な方法の一つは、私たちの高レベル神経科学モデルが、ニューロンを完全にシミュレーションした場合よりもはるかに悪い予測しかできないという事実である。したがって私たちの他者理解の機構的理解は不完全であるはずである。

ある程度の情報損失はおそらく不可避であるが、良いモデルであればはるかに損失は少ないはずである。自動車のデファレンシャルギアを理解しているからといって、原子レベルのシミュレーションほど完全にすべてを予測できるわけではない。しかし歯車レベルのモデルでも非常に精密な予測は可能であり、モデルが予測すべきこととそうでないことの境界も明確である。

ではなぜ人間の頭脳についてもこの程度のモデリングが可能だと期待できるのか。人間の頭脳はあまりにもランダムである可能性もある。正確な予測をするにはニューロンのシミュレーションしか方法がないのかもしれない。

しかしそうではない証拠の一つは、あなたの母親ですら、脳の形式モデルよりもあなたの行動をうまく予測できるという事実である。つまり人間心理には、ニューロンを完全にシミュレーションしなくても暗黙的に知ることのできる構造が存在しているのである。ただしそれがまだ明示化されていないだけである。

さらに具体的な証拠として、健忘症患者の研究がある。ある種の健忘症患者は、同じ冗談を何度も文字通り繰り返す傾向がある。これはその人の脳の中に一定の規則性が存在することを示唆している。つまりその人は、状況や看護師の存在や記憶や歴史や喜びを広めたい欲求や賢く見られたい欲求などに基づく計算を無意識に実行しており、それが小さな変動にもかかわらず安定しているのである。

もし人の精神計算にこれほどの規則性があるならば、それは理解可能であるはずである。意思決定の歯車を学び、脳を十分な深さで理解することで、次のように言えるようになるはずである。

ああ、これらのニューロンは喜びを広めたいという欲求に対応しており、あれらのニューロンは賢く見られたいという欲求に対応している。そしてここにあるニューロンは看護師が部屋に入ってきたときに可能な思考を生成するものであり、ここには「冗談を言う」という思考を生成する装置があり、先ほどの欲求ニューロンと相互作用して、その思考が前面に押し出される。そしてこの文脈がこのパラメータで記憶ア

クセスに影響し、その経路をたどると視線を部屋の中に動かす発想が生まれる。そして部屋には帆船の絵があるので、このニューロン群によって帆船概念が活性化され、記憶検索に戻ることで患者は帆船の冗談を言うことになる。

実際の説明はこの通りではないだろう。しかし「毎朝同じ冗談を言う」という単純な観測事実は、それが還元不可能なランダムではないこと、そこに再現可能な計算が存在することを強く示唆している（もちろん、これは常識とも一致する。もし脳が完全にランダムでしかなかったら、私たちはまともに機能できないはずだからだ）。

知性の理解にはすでに一定の進歩が存在する

これは私たちが知性についてまだ多くを学べると確信している主な理由である。『The MIT Encyclopedia of the Cognitive Sciences』や『Artificial Intelligence: A Modern Approach』第2版のような、現代のディープラーニング以前に書かれた本を読むと、認知問題がどのように解かれるのかについて多くの洞察が得られる。それらの洞察のすべてが一般向けに書き直されたり大学教育に広く普及したりしているわけではないが、実際には非常に多くの知識が存在している。

例えば、他の条件が同じならより単純な仮説を好むべきだという科学原理を考えてみよう。ここでいう「単純」とは正確には何を意味するのか。

「隣人は魔女で、彼女がやったのだ」という説明は、多くの人にとって電磁気を支配するマクスウェル方程式より単純に聞こえる。ではどの意味で方程式の方が単純なのか。

そもそも証拠が仮説に「適合する」とはどういう意味なのか。仮説が証拠を「説明する」とはどういう意味なのか。仮説の単純さと説明力はどのようにトレードオフされるのか。「隣人は魔女で、彼女がやったのだ」という説明は非常に多くのことを説明できるように聞こえる。しかしそれは悪い説明だと多くの人が直感する。そして実際、魔術があまりにも多くを説明できてしまうことこそが、その説明が悪い理由の一部なのである。

異なる仮説のあいだで選択する統一原理は存在するのか。それとも問題ごとに使い分けられる百種類の道具があるだけなのか。そして後者だとすれば、人間の脳はどうやってそのような道具を発明できるのか。

コンピュータや脳が成功裏に用いることのできるすべての仮説を記述する言語は存在するのか。

これらの問いは初めて聞く人には哲学的で答えの出ない問題のように聞こえるかもしれない。しかし実際には、これらはすべて計算機科学、確率論、情報理論においてすでに解かれており、「Minimum Message Length」「Solomonoff prior」「likelihood ratio」などの名前で知られている。

また、特定の領域で人間を超える完全に理解されたAIがすでに存在するという事実も

重要である。チェス AI のディープブルーでは、関係する原理はすべて理解されている。ディープブルーは手作業でコードが書かれたため、コードの各部分が何をしているか、そしてそれが全体のコードベースとどう関係しているかを容易に確認できる。

ChatGPT のような LLM の場合、それらがどのように動くかについて完全で短い説明が存在するかどうかははっきりしていない。LLM は非常に大規模であり、同じ挙動が多くの異なる偶然的理由から生じる可能性がある。LLM のアーキテクチャを理解することは、訓練済みモデルの不可解な重みが果たしている役割についてはほとんど何も教えてくれない。

ChatGPT は、数十年研究しても科学者にとって理解が難しい対象である可能性はある。しかし ChatGPT の存在は、知能が本質的に混沌としていなければ機能しないことを意味するわけではない。それは単に、ChatGPT のようなものをそのまま超知能まで拡大するのが極めて悪い考えであることを意味するだけである。

ある特定の頭脳が混沌としているという事実は、知性が理解不可能であることを意味しない。それは ChatGPT が永遠に理解できないことを意味するわけでもない。

燃えている丸太を 100 本注意深く観察すると、まったく同じ燃え方をするものは二つとしてないことが分かる。火は異なる仕方で広がり、火の粉は異なる方向へ飛び、全体として非常に混沌としている。もし耐火顕微鏡で丸太を詳細に観察できたなら、さらに目まぐるしい細部が見えるだろう。このような混沌とした細部を見た古代の哲学者が、火は決して完全には理解されないだろうと結論したとしても不思議ではない。

そして彼らはある意味では正しかったかもしれない。私たちは丸太を見て、どの木片が最初に西へ飛ぶ火の粉になるかを正確に言い当てる能力を決して持たないかもしれない。しかしその哲学者が、火とは何か、なぜ起こるのか、制御された条件で火をつくり出す方法、そしてそれを大きな利益のために利用する方法を人類が決して理解できないと結論したならば、それは重大な誤りである。

火の粉の正確なパターンはそれほど規則的でも再現可能でもない。しかしより抽象的なレベルでは、黄色から橙色、赤色に揺らめく高温の現象は、世界で繰り返し現れる規則性であり、人類はそれを理解することに成功したのである。

『超知能 AI をつくりば人類は絶滅する』における議論は、今日知られている知能の技術的詳細にはそれほど依存していない。「人々はより賢いコンピュータをつくり続けており、それを制御できていない。そしてもし非常に賢く制御不能なものをつくりば、私たちは死ぬ」という考えは、それほど秘教的な考えではない。しかしこの分野には既存の知識体系がすでに存在することを知っておくことは有益である。なお、この分野にはまだ多くの謎と未知が残っている。

本書の中心的議論は、知能が原理的に理解可能かどうかにはあまり依存していない。もし人間が超人的機械知能の謎を決して理解できなかったとしても、人工超知能は私たちを殺すことができる。

この問いが重要になるのは、AI への自殺競争を止めた後に何をすべきかを決めるときである。

そして知能はおそらく理解可能であることが重要である。つまり原理的には、優秀な人々が成熟した知能科学を発展させ、AI アライメント問題の解決策を見つけることが可能であるはずだからである。

もちろん現代の人類は、その偉業にはまったく近づいていない。しかしそれが原理的には可能であるという事実は、人類がこの混乱から抜け出す道筋をどう進むべきかに影響を与える。私たちは第 10 章の〈より深い議論〉でそれについて述べる。

そこに至る前に、研究者がもし人間知能を超えることに成功した場合、なぜ私たちが先に述べたような AI 技術がこれほど重大な脅威になるのかを説明する必要がある。その説明は第 3 章から始まる。