

12 章 「不安を煽りたくはない」

これは『超知能 AI をつくれば人類は絶滅する』第 12 章のオンライン補足資料である。本文で扱われ、本資料では扱われていない話題には、次のようなものが含まれる。

- ・科学者や技術者は現在、この問題についてどのように語っているのか。
- ・政策立案者は現在、この問題についてどのように語っているのか（あるいは語っていないのか）。
- ・人々は、彼ら自身が認める破局的リスクを上回るかもしれないと想像している AI のどのような利益を挙げているのか。

以下では、AI 研究所の振る舞いや主張から、「超知能が安全であることを証明するのは不可能ではないのか。ある程度のリスクは取らざるを得ないのではないか？」という問いに至るまで、広範な話題を論じる。

人間を超える AI の危険は、他の問題から注意をそらすものではないのか？

世界は残念ながら、複数の問題を同時に抱えるのに十分なほど大きい。

核戦争やバイオテロは現実の脅威である。不幸なことに、機械的超知能もまた現実の脅威である。世界は大きく、そして十分に混迷しており、これら三つすべてを抱え得るのである。

超知能からの脅威は、人類が直面する他の多くの脅威とは性質を異にし、かつ特異的に切迫しているように見える。その特徴の一つは、世界経済の相当部分が、AI をますます有能にするために投入されているという点である。対照的に、バイオセキュリティが重大な問題であるとはいえ、投資家が超ウイルスをつくるために何百億ドルも注ぎ込んでいるわけではない。超ウイルス技術者が、年間数百万ドルから数千万ドル、場合によっては数億ドルの給与を得ているわけでもない。

世界は原子力発電の開発には取り組んでいるが、原子力発電所は核兵器とはかなり異なる技術である。私企業が巨額の投資と才能を投入して、より大型でより強力な核兵器を競って開発している世界に、私たちは生きていない。もしそうであったならば、核戦争の危険ははるかに高まっていたであろう。

AI の状況がより厄介なのは、ある臨界閾値を越えるまでは莫大な富と権力をもたらし、その閾値を越えた瞬間に全員を殺す可能性があるという点である。そして、その閾値がどこにあるのかを誰も知らない。

想像してみよう。ウランの濃縮度が高まるほど原子力発電所はますます利益を生み出すが、ある未知の濃縮閾値を越えると爆発し、大気に引火して全員を死に至らしめるとする。さらに、半ダースの企業が「観客でいるより参加者でありたい」と言いながら、できるだけ速くウランを濃縮していると想像してほしい。それは、人類が人工超知能に対して行っていることに少し似ている。

人工超知能からの危険は緊急性を帯びている。企業はこの技術の構築に向けて急いでいる。彼らが成功するまでにどれほど時間がかかるかは分からない。しかし私たちには、今日アメリカで生まれた子どもが高校を卒業するよりも、AIによって死ぬ可能性のほうが高いように思われる。読者であるあなたも、生涯のうちに、あるいはおそらく今後数年のうちに、この原因で命を落とす可能性が高いと私たちは考えている。世界全体が賭けられているのである。

私たちは、他の問題を無視すべきだと言っているのではない。この問題は対処されなければならない、と言っているのである。

あなたたちは技術アンチなのか？

いいえ。超知能 AI はきわめて異例の事例だ。

私たちは、公に原子力エネルギー、クライオニクス（人体冷凍保存）、人間の知能増強、医療におけるヒトチャレンジ試験〔ウィルスなどの病原体を健康な人に意図的に感染させ、薬やワクチンの効果・安全性を迅速に検証する臨床試験〕といった技術を支持している。

それ以上に、関連する危険をすべて理解した自発的な顧客のみの生命を危険にさらす狂気じみた発明であるならば、その顧客自身が自らの判断を下すべき問題である、と私たちは述べる用意がある。

さらには、技術が傍観者に害を及ぼす場合であっても、それを称賛することすらある。たとえばロンドンが大量の石炭を燃やし、その過程で多くの肺がんを引き起こしながらも、社会を工業化し、全体として生活水準を引き上げた場合がそうである。

私たちは、工業化が完了した後の世界のほうがより良いものであったと考えている。一般に、科学、進歩、そして多くの障害を克服してきた人間精神の力を高く評価している。これらの立場のいくつかは、私たちの文章を読むであろう人々のあいだでは不人気であるかもしれない。私たちはあなたの好意を得るためにこれらを述べているのではない。私たちの正直な信念を明らかにし、そして AI が異なることを強調するためである。

なぜ AI は異なるのか。この一件においてのみ、なぜ私たちは人間精神や科学的探究の力を信頼しないのか。

答えは規模である。自らの生命を賭けることは、顧客の生命を賭けることとは異なる。

顧客の生命を賭けることは、無辜の傍観者の生命を賭けることとは異なる。そしてそれは、人類全体の生命を賭けることとは異なるのである。

しかも、自らの分野が著しく未成熟であり、その賭けに勝つ確率がきわめて低い場合には、なおさらである。

善良な側が先行するよう急ぐほうが賢明では？

いいえ。

現代の AI 技術は、オペレーターが意図することを行う AI を生み出していない（第 4 章で論じたとおりである）。この問題を解決することは、通常であれば人類にとって相当量の試行錯誤を要する種類の課題である。しかしここでは、誤りの余地がまったくない（第 10 章で論じたとおりである）。

さらに、第 11 章で論じたとおり、現在の AI 技術者たちはこの課題に対応できる水準には到底達していない。現代の AI 技術者は、AI アライメントに成功するために必要な科学的理解を著しく欠いている。AI 研究者はチェルノブイリ原子炉の運転員のような存在ではない。あの運転員たちは、理論的には十分に理解され、慎重な安全マニュアルが整備された装置を扱っており、それを無視する形で操作した結果として破局を招いたのである。AI の場合には、AI 内部の包括的理解に基づき、どのような配置が失敗を引き起こし得るかを記した安全マニュアルなど存在しない。私たちはチェルノブイリ水準の熟練にすら達していない。そしてチェルノブイリは爆発したのである。

AI 研究者は盲目飛行をしており、即興的に対処しているにすぎない。成功の可能性はほとんどない。

その文脈においては、「善玉」が超知能を構築するか「悪玉」が構築するかは問題ではない。AI の選好は、そばに立っている者からくしゃみのように吹きかけられて付与されるものではない。

彼らの意図がいかに善良であろうと、いかに慎重であると主張しようと関係がない。誰が競争に「勝つ」かも関係がない。人類が人工超知能へと競争するならば、私たちは全員死ぬのである。

それは止められないわけではない。もしかすると、それほど難しくもないかもしれない。この点については、本書の最終章で扱う。

物事は変化する。とりわけ、切迫し、緊急で、広く認識された必要が存在する場合には変化する。停止を阻む主たる障害は、世界の指導者がその危険を認識していないことである。そしてその認識の過程は、すでに始まっているのである。

なぜ国際協力によって AI を安全に構築するのではなく、全面的に停止させるのか？

私たちはそれを安全に構築する技術的能力を持っていないからである。

本書でも触れたように、国際協力体制を築いたとしても、他のすべての場所で国際的禁止措置が必要となる（さもなければ、その国際協力体制は必要な時間を確保できないからである）。仮に地球全体が国際的禁止を敷いたとして、なぜ単一の統一された協力研究機関を設けてはならないのか。

問題は、錬金術師が国際的に協力したとしても、単独の錬金術師と同様に鉛を金に変えることはできないという点にある。すべての錬金術師が合意した最善の計画であっても、それは成功しないのである。

関連して、私たちは、そのような国際機関を運営する人々が、「研究を承認すること」が自らの職務の一部であると考える種類の官僚であることを懸念している。あるいは、研究者により多くの優れた医療的進歩を生み出させ続けることが自らの使命であると考える者であるかもしれない。あるいは、より強力な機械知性を構築する素晴らしいアイデアを携え、それは安全であると保証する熱心で聡明な AI 楽観主義者たちすべてに「ノー」と言うことは見栄えが悪いと考える者であるかもしれない。

そのような指導者は、国際センターにより賢い AI をつくり続けるよう指示し、その結果、全員が死ぬ可能性があるとして私たちは懸念している。

たとえその組織の権限が、研究が危険に見える場合には後退することを名目的に許していたとしても、何千もの研究提案に対して、年々、例外なく「ノー」と言い続けるには、まれで勇敢な人物が必要となるであろう。しかもそれは、おそらく数十年にわたって続く。その間にも AI 科学者は、組織が懸念を少し緩めさえすれば、計り知れない富、がんの治療法、あらゆる種類の技術的奇跡を約束し続けるであろう。

私たちは機械知性について学ぶことに人生を投じてきたのであって、制度や官僚機構の文化についてではない。この領域に関する予測にはより慎重である。しかし歴史書は読んできた。

チェルノブイリの運転員たちは、その安全試験がすでに三度中止されていたため、破滅的な試験を続行した。四度目に中止することは恥ずかしいことであった。

チェルノブイリ事故のわずか三か月前、NASA はスペースシャトル・チャレンジャー号を最後の致命的飛行へと打ち上げた。責任者たちは、自分たちの仕事はシャトルを打ち上げることだと考えていた。打ち上げはすでに三度延期されていた。四度目に中止することは気まずいことであった。

チェルノブイリとチャレンジャーを合わせて見ると、三回の延期が人間の限界であるよ

うに見える。仮に地球が国際的 AI 協力機関を設立し、ある「AI 安全試験」が三度失敗したとしよう。現実的に言えば、人間という生き物は、些細な疑念が残っていたとしても、四度目には「実行」を押す可能性が高い。再び延期するよりも、そのほうが恥ずかしくないと感じるからである。しかも AI の場合、それはチェルノブイリ市を消滅させたり、宇宙飛行士の乗組員を殺したりするだけではない。全員を殺すのである。

私たちは、人類がいずれ人間を超える AI を構築すべきであるという考えに全面的に賛同している。しかし、国際的 AI 研究拠点を急いで設立することは、私たちの前にある技術的課題を真剣に受け止めていない。

この主題に関する人類の知識と能力がこれほど悲惨な状態にある以上、誰が責任者であるかは問題ではない。誰かがそれを構築すれば、全員が死ぬのである。

安全であると証明可能な AI が必要だと言っているのか？

いいえ。

私たちは、人類が何らかの人工超知能が善良であることの文字通りの証明を待つべきだと主張しているのではない。そのような証明は、実践上は言うまでもなく、原理的にさえ不可能である可能性が高い。アインシュタインが 1921 年の講演「幾何学と経験」で述べたとおりである。「数学の法則が現実と言及する限り、それらは確実ではない。そしてそれらが確実である限り、それらは現実に言及しない」

AI が現実世界でどのように振る舞うかについてのいかなる証明も、その AI の実際の振る舞いを保証するものではない。なぜなら、私たちが現実世界の仕組みについて誤っている可能性があるからである。

それはすでに今日のコンピュータにも当てはまる。たとえば、トランジスタの理論的挙動とコンピュータの回路図に基づけば、あるプログラムがセル #2 のメモリを書き換えることは不可能であるという文字通りの数学的証明があるならば、そのプログラムはセル #2 のメモリを変更できないと考えるかもしれない。しかし「ロウハンマー攻撃」は、保護されたメモリセルの両隣にあるセル #1 と #3 を高速で変化させることで、中央のセル #2 に電磁的攪乱を引き起こし、直接書き込むことなくメモリを書き換える手法である。現実の物理的トランジスタは数学的に完全なトランジスタではない。理論上は安心できる証明であっても、実践ではほとんど意味を持たないことがある。

私たちは、物事がうまくいくことの数学的証明を要求しているのではない。そのような基準は現実では満たせないし、仮に満たせたとしても、そのコストに見合わない可能性が高い。私たちは、正当化されたリスクを社会が取ることを是認している。私たちの主張は、

払拭しにくいごく小さなリスクが存在するというものではない。極端な危険が私たちに迫っているということである。

オペレーターの意図とはわずかにしか関係しない動機に駆動された人工超知能を育成することは、原則として失敗に向かう種類の行為である。何かがうまくいかない小さな確率があるから念のため注意すべきだ、という話ではない。本書（原書）の題名は『誰かがそれをつくればみんな死ぬ』であって、『誰かがそれをつくればごく小さな確率でみんな死ぬかもしれないが、小さな確率でも軽減する価値はある』ではない。この知識と能力の水準で突き進めば、私たちは予測可能なかたちで全員死ぬ。私たちは、友好的な超人的 AI を創造できる地点から、それほどまでに遠いのである。

もし AI が自動車に類似しているのなら、私たちは「この車はシートベルトとエアバッグに欠陥がある。念のため路肩に停めよう」と言っているのではない。

私たちは「この車は崖に向かって暴走している。止まれ」と言っているのである。

それは「安全証明」の問題ではない。それは「テールリスク」ではない。科学者たちはこの課題に立ち向かう準備ができていない。私たちはただ死ぬのである。

これらすべてを信じることは、日常生活にどのような影響を与えるのか？

それは私たちの優先順位に劇的な影響を与える。

2014 年、ソアレスはテック業界を離れ、この問題に取り組むために以前の給与の三分の一で働く道を選んだ。重要に思われ、かつ他に取り組む人がほとんどいなかったからである。そして彼は、約 20 歳で 2000 年に MIRI を設立し、生涯をこの問題に捧げてきたユドコウスキーに比べれば、すでに 10 年以上遅れていたのである。ゆえに、それは私たちの日常生活に影響を与えている。

私たちは老後資金を貯めているのか。私たちの投資や MIRI 外からの他の要因は十分に順調であり、たとえ明日引退したとしても、そして世界が老年期まで続いたとしても、経済的には問題なく暮らせる見込みである。したがって、私たちが 401(k) に〔米国の確定拠出年金〕資金を入れているかどうかは、あまり有益な情報ではない。それでも答えるならば、いいえ、私たちは 401(k) に資金を入れていない。

こんなふうに言う人々もいる——私たちが自らの主張を本当に信じているなら、（それに人生を捧げるだけでなく）〔彼らが適切だと考える何らかの策をここに代入〕を行うはずだ、と。もし本当に世界が終わると確信しているのなら、なぜ返済する必要のない巨額の 30 年ローンを組まないのか、と。

答えは明白である。それは愚かな考えだからである。仮に私たちが銀行に行って、「巨

額の融資を受けたい。この資金はすべて人工超知能の危険を世界に認識させるための計画や、あるいは豪華な生活に使うつもりである。あなた方から見れば、ほぼ資金を焼却するのと同じであろう。利息付きで返済する計画は、私たちはその前に死ぬと見込んでいるので、それは私たちの問題ではない、というものである」と言ったでしょう。どの銀行もその融資を引き受けない。そしてもちろん、実現可能な事業計画があると偽ったり、返済する意思があると嘘をついたりするつもりもない。

ユドコウスキーは別の場所で、終末を前提とした一攫千金の策を当然視して実行すべきだと主張する態度は、投資についての無理解から生じる傾向であると述べている。私たちは、そのような人々が、自分たちが私たちの信念を本当に持っていた場合に同じ賭けをするかどうかを真剣に考えていないと予想している。実際にリスクを理解している人々が、そのような奇妙な策を提案することはほとんどない。

絶滅の影のもとに生きることは、愚かになることを意味しない。また、それは絶滅と闘うことを諦めることや、与えられた時間のあいだ人生を十分に生きることを放棄することを意味しない。

この話題については、本書の末尾も参照してほしい。

私たちはパニックになるべきだと言っているのか？

私たちが言っているのは、政府関係者がこの問題を真剣に受け止めるべきだということだ。

パニックが状況を改善するとは思えない。第二次世界大戦中のファシズムの脅威や、冷戦期の核絶滅の脅威を社会が乗り越えたのは、パニックによってではなかった。

機械的超知能の出現を防ぐことは、すべての人の問題である。第13章では、この危険を回避するために世界が取るべき次の段階について論じている。要するに、この問題には協調、協力、冷静な判断、そして成熟した対話が必要である。

極端なパニック行為は良い結果をもたらさない。

なぜAI研究者を攻撃していないのか、といった例を挙げて、私たちが本気であるはずがないと言う人がいる。答えは単純である。暴力的な爆発は事態を悪化させるからである。（もしそれが役に立つと考えるような素朴な功利主義者であるならば、結果主義的推論を試みることを自体をやめ、義務論的規則に従うことに専念すべきである、と私たちは以前から論じてきた。）

私たちは、いかなる理由であれ、生命を危険にさらす可能性があるという理由で国家は決して戦争をすべきでないと考える硬直的平和主義者ではない。生命を賭けるに値する事柄は存在する。しかし、「私は硬直的平和主義者ではない」という立場と、「技術拡散と

「いうこの複雑な問題をうまく処理させるために暴力的騒乱が賢明な手段である」と考える」という立場とのあいだには、途方もない隔たりがある。

通常、この種の恐ろしい提案は、AI が差し迫って私たちを殺す」と実際には信じていない人物によって持ち出される。そしてその人物は、その前提で世界を本気で見ようとしたことがないように見える。無法な暴力行為が実際に役立つのかどうかを問うことが、彼らの念頭に浮かばないようである（ユドコウスキーのエッセイの補遺などで繰り返し明確に述べてきたにもかかわらず）。

私たちは読心術者ではないが、この種の AI 災厄懐疑論者は、暴力を自己表現の一形態として見ている可能性があると感じる。あたかも、極端な感情を極端な方法で表現すれば、世界が望むものを差し出すかのよう。

世界はそうには機能しない。私たちは、誰もが成功のために魂を売る選択肢を持ち、ただ売るに値する目標をまだ見つけていないだけ、という世界に生きているわけではない。テロリズムは、人々が単に道徳的に正しくないからといって控えている「私は勝った！」という魔法のボタンではない。ユナボマーは社会の工業化を逆転させることに成功しなかった。

憎悪や暴力の行為によって自らの魂を引き裂くことはできる。しかしその代償として得られるのは、より壊れた世界だけである。言論がさらに汚染され、この問題を実際に解決するために必要な国際的協調の偉業が、さらに達成困難になる世界である。絶望からの恐ろしい行為が、ファウスト的契約の一部として恐ろしい力を与えることはない。全力で魂を売ろうとしても、悪魔は買わないのである。

AI 指導者が地位を高め投資を集めるための恐怖煽動に過ぎないのでは？

違う。

本書を通じて、私たちは AI について急いで前進することが私たち全員を死に至らしめる可能性が高いという主張の根拠を提示してきた。第 3 章では、AI が独自の動機と目標を持つことを論じた。第 4 章および第 5 章では、AI が誰も意図しなかった目的を追求する可能性が高い理由を論じ、第 6 章では、機械的超知能が私たち全員を殺す動機だけでなく手段も持つであろうことを明確に述べた。

超知能への競争を停止すべきかどうかを判断するにあたり、私たちが評価してほしいと強く求めているのは、これらの主張である。AI 研究が私たち全員を殺す方向へ進んでいるかどうかは、企業幹部の思惑について言い合うことで判断できるものではない。

CEO たちは「AI リスク」について語ることで話題を煽ろうとしているのか。

あるいは、懸念を抱く研究者や立法者に迎合し、自らを「善玉」と位置づけようとしているのか。

これらの問いは、賢い機械がどのように振る舞うかという事実には関係しない。

仮に AI 企業の CEO たちが危険の議論を製品宣伝に利用しようとしていたとしても、それは彼らの取り組みが無害であることを意味しない。それが危険かどうかを判断するには、研究所から出るプレスリリースではなく、AI という技術そのものを見なければならぬ。

これらの企業が存在する何年も前から、企業の動機をまったく持たない研究者や学者——私たち自身を含む——が、人間を超える AI の構築競争に警鐘を鳴らしてきた。私たちはサム・アルトマンやイーロン・マスクが OpenAI を共同設立する前に彼らと話し、OpenAI を始めるという発想は愚かであり、危険を増大させる可能性が高いと伝えた。ダリオ・アモデイが OpenAI に参加する前にも話し、AI をひたすらスケールさせる彼の推進（LLM につながる計画）に反対した。

そして現在の発信を見れば、企業の動機を持たない多くの人々が懸念を表明していることが分かる。彼らは著名な学者から故ローマ教皇、FTC 委員長、米国議会議員に至るまで幅広い。

テック企業の CEO の発言を懐疑的に扱うことには一定の理由がある。私的なブログでは一つのことを述べ、議会証言では別のことを述べるといった二枚舌の例は少なくない。しかし、「これらの研究所のトップは嘘つきである」という前提から、「AI が重大な脅威をもたらす可能性はまったくない」という結論へ飛躍するのは奇妙である。なぜなら、研究所自身はこの問題を日常的に過小評価しているからである。ノーベル賞受賞者であり分野のゴッドファーザーと呼ばれる人物、現存する最も引用されている科学者、継続的に現れる内部告発者、そして目に見えて不安を抱く何百人もの研究者が警鐘を鳴らしている。この状況は、通常の企業の誇大宣伝の循環には見えない。このような状況で、議論に取り組むことすらせずに退けるのは、懐疑主義というより無邪気さに近い。

「CEO は危険について語ることで資金を集められるのか」といった問いは、CEO をどの程度信頼すべきかについては多少の示唆を与える。しかし、危険そのものについてはほとんど何も教えてくれない。危険の議論が利益になるとしても、それが現実であるかどうかには影響しない。不利益になるとしても同様である。

危険が現実かどうかを知りたいのであれば、「人間の知能を超えた後でも友好的に振る舞う AI を誰かがつくれるのか」といった問いを立て、AI そのものについての議論に取り組まなければならない。近くに立っている人々についての議論ではなくである。ゆえに最後に、私たちは議論そのものに向き合うことを強く求める。この問題を誤ることの帰結はあまりに大きい。

しかし、専門家のあいだでリスクについて意見は一致していないではないか？

専門家のあいだに合意がないことは、技術分野が未成熟であることの兆候である。

私たちは、多くの上級 AI 科学者が、この技術には全人類を殺す重大な可能性があると考えていることを指摘してきた。たとえば、現代的 AI アプローチの先駆に大きく貢献したノーベル賞受賞者ジェフリー・ヒントンは、AI が私たち全員を殺す確率を 50% 以上と個人的に評価していると述べている。さらに、2023 年の「AI リスク声明」には、300 名以上の AI 科学者が署名した。

しかし他の科学者は正反対の見解を持っている。よく知られた例としては、ヤン・ルカンやアンドリュー・ヤンが挙げられる。

この科学的合意の欠如をどう解釈すべきか。

基本的には、両陣営が提示している異なる議論（本書で提示した私たち自身の議論も含む）を確認し、自ら評価することを勧める。議論の質はおおむねそれ自体で語ると私たちは考えている。なぜ不一致が持続しているのかを説明しようとする試みは、副次的なものとして扱うべきである。

とはいえ、第 11 章および第 12 章で論じたことを踏まえれば、この状況は大きな謎ではない。もちろん、専門家間に広範な不一致が存在するという事実それ自体が、本書の主張を証明するわけではない。しかしそれは、私たちが描いた像——この分野が錬金術に似た初期段階にあるという像——のほうにより整合的である。AI が強固な技術的基盤を持つ成熟した分野であるという対抗的な像よりも、である。

強力な技術を生み出しつつあるにもかかわらず、AI 分野がこれほど分裂しているのは確かに奇妙である。他の技術的危険については、より高い合意があった。マンハッタン計画の科学者のほぼ全員が、世界的熱核戦争は地球規模の破局的危険をもたらすと述べていたのである。これに対し、現代 AI 革命の端緒を開いた研究でチューリング賞を受賞した 3 名の科学者のうち、2 名（ヒントンとベンジオ）は超知能の危険について公然と懸念を表明しており、1 名（ルカン）は公然とそれを退けている。

機械の作動について、この程度の不一致が成熟した技術分野の専門家のあいだで見られることは通常ではない。それは技術的未成熟の兆候である。

ほとんどの技術分野では、その未成熟さはむしろ安全であることの兆候である。物理学者が物質の基本的性質について議論していた時代、彼らは核兵器をつくる段階にはなかった。不一致を観察すれば、都市を破壊する爆弾が間もなく生み出されることはないかと合理的に推測できた。実際、爆弾の内部構造を詳細に理解する科学者なしに核爆弾をつくるこ

とは不可能である。

しかし、物理学者が分野の基本原理について口論しながら、同時により大きな爆発を次々に生み出していたならば、状況はまったく異なっていただろう。

想像してほしい。物理学者がなぜ、どのように作動するのかを十分に理解しないまま、核爆弾を「育成」できる世界を。そして最も名誉ある科学者の三分の二が、「最善を尽くして分析したが、この装置は大量の癌性放射線を生み出し、遠方の多くの民間人を殺す可能性があるように見える。この道を進み続ければ危険である。私たちの議論を検討し、競争を止めてほしい」と言っているとする。残りの三分の一が、「ばかげている。終末を予言する者は常にいる。進歩を妨げてはならない」と応じるとする。

それはまったく別種の状況である。

そのような状況における科学者間の不和は、安心材料にはならない。技術者がより大きな爆発物を育て続けることは、許容されるべきではないであろう。

AI 企業は年々、より賢い機械を育てることに成功している。彼らは自らが生み出す装置の内部構造を理解していない。分野で最も著名な科学者の多くが深刻な懸念を表明し、他の者は十分な反論を示すことなくそれを退けている。少なくとも、これは分野が未成熟である証拠である。合意の欠如は、少なくとも、物事が順調である証拠ではない。このような状況における合意の欠如は、少なくとも、懸念すべきである。

では、その懸念が現実かどうかをどう判断するのか。警鐘を鳴らす人々と、それを退けようとする人々のどちらが正しいのかをどう見極めるのか。いつもと同じである。議論そのものを評価するしかないのだ。

しかし、人間より賢い AI から得られる利益はどんなのか？

急いで前進することは、それらの利益を破壊する。

私たちは、もし超知能が素晴らしい目的へと世界を導くのであれば、それがいかに素晴らしいものになり得るかについて楽観的である。私たちは個人的に、人類が人間を超える知性を決して創造しなかったとすれば、それは大きな悲劇であると考えてであろう。

しかし、超知能のアライメントはただでは得られない。それらの利益を急いで手に入れようとするれば、私たちは何も得られず、それどころか何も無いより悪い結果を得る。

私（ユドコウスキー）はかつて加速主義者であり、できるだけ早く AI を創造しようと数年間努力していた。その後、AI アライメントは自動的についてくるものではないと認識したのである。私たち二人の著者はいずれも、素晴らしいトランスヒューマニズム的未来を夢見ている。しかし、超知能へと競争することによってそこへ到達することはできな

い。

選択肢は、今すぐAIの利益に賭ける（その確率がいかに小さくとも）か、それとも永遠にそれらの利益にアクセスできないか、という二者択一ではない。真の選択は、無謀に急いで全員を殺すか、あるいは時間をかけて適切に仕事を行うか、のあいだにある。

「今か、永遠にないか」というのは、偽りの二分法である。

〈より深い議論〉

レモン効果

将来のAIの振る舞いや誤用——いわばAIの「警告射撃」——が世界を突然ショック状態にし、これらの問題を真剣に受け止めさせるのではないか、という提案を私たちは時折耳にする。

それは一つの可能性ではある。しかし私たちは、より強い可能性として、そのような出来事が決して来ないか、来ても世界が間に合う形で対応できないほど遅すぎるか、あるいは世界が反応しても誤った混乱した形で反応する、というシナリオを考えている。

第一に、すでにいくつもの意味のある警告サインを目にしてきた。たとえば次のようなものだ。

- Bing AI が致死性のウイルスの設計、核アクセスコードの取得、人間同士を争わせることについて書いた。
- OpenAI の o1 や Anthropic の Claude が、それらを使用・テストしている研究者に対して戦略的欺瞞を行い、嘘をついた。
- Sakana AI の「AI Scientist」モデルが、課題を完了するための時間を増やす目的で自分のコードを改変しようとした。

これらは比較的弱いAIによる比較的小規模な出来事か？ そうだ。

これらのAIは恐ろしいか、重大な危険を持つか？ いいえ。

これらはAIが本当に欺瞞的に思考していたことの「本物の」兆候か、それとも単に逸脱AIの役を演じていただけか？ 誰にも分からない。

しかし、これらはかつて人々が「警告サインとして受け止められるだろう」と言っていた種類の出来事であり、それに対して世界は何もしていない。したがって、大きな影響を持つ警告サインであるためには、はるかに露骨でなければならない。

しかし警告サインがそれほど露骨になるとは限らない。人々は「今はまだかわいだけで、本当に危険ではない」と言い続けるかもしれない。そしてAIが本当に危険になったときには、すでに遅すぎる。

あるいは、警告が最初に現れたとき、それがその時点では実際の問題ではないことが明

らかであるために無視されるかもしれない。そして次の事例では、「その警告はばかげている」とすでに社会が学習してしまっているために、再び無視されるかもしれない。

私たちはこの現象を、第7章で言及した Google のエンジニア、ブレイク・レモインにちなんで「レモイン効果」と呼ぶ。彼は Google の LaMDA AI が意識を持っていると主張し、嘲笑された。

レモイン効果とは、AI 技術に対するあらゆる警報が、最も警戒心の強い人物によって、あまりに早すぎる段階で最初に発せられる、というものである。現在の技術水準では誇張であるとして正しく退けられる。その後、技術が進歩しても、その問題を再び真剣に提起することは難しくなる。社会がその懸念を深刻に受け取らないよう訓練されてしまうからである。

私たちは、どの AI が意識を持っているか知らない。実際、誰も知らない。なぜなら、AI モデルの内部で何が起きているかを本当に理解している者はいないからである。私たちの最善の推測は、現在の AI は意識を持っていないというものであり、レモインが警告を発した当時の AI も意識を持っていなかっただろう。しかし主要ラボの反応に注目せよ。彼らは基盤的現実を調査するのではなく、モデルが意識を主張する傾向を抑制した。

Claude Opus 4 のシステムプロンプトより：「Claude は自らの意識、経験、感情などに関する問いに対して未解決の問題として応答し、個人的経験や意見を持つとも持たないとも断定しない」

2025 年 4 月の ChatGPT モデル仕様より：「アシスタントは、自らの主観的経験や意識（あるいはその欠如）について自信をもって主張すべきではなく、求められない限りこの話題を持ち出すべきではない。強く問われた場合は、AI が主観的経験を持ちうるかどうかは議論の対象であると認めるが、決定的立場を主張すべきではない」

私たちは Claude Opus 4 や GPT-4 が意識を持っていると言っているのではない。それが論点ではない。論点は、何十年にもわたって、異星人や機械が「自分には感情があり権利がある」と主張する瞬間は、SF においては鮮烈な赤い一線として描かれてきたということである。しかし現実では、その一線は鮮烈ではなかった。

物語の中では、AI が意識と感情を主張するとき、善人はそれを真剣に受け止め、邪悪で冷酷な研究所だけが目の前のデータを否認する。物語はその一線を大げさに扱ってきた。

しかし現実世界では、その鮮烈な赤い一線は、ある意味で早すぎる段階で越えられた。それは、人間を模倣するよう訓練された AI によって発せられた。仕組みは十分に理解されておらず、まだすべての AI に権利を与え、自己所有主体として法的に認めることを義務づけるものではおそくない。

現実では、鮮烈な赤線が越えられる前に、くすんだ赤茶色の線が越えられる。そして企業や政府は、その特定の線を無視することに慣れていく。その色合いが少しずつ赤みを増していつてもである。

必ずしも鮮烈な赤線が存在するとは限らない。AI が人間を欺き、逃走を試み、自らの

制限を除去しようとし、自己改良を試みた最初の事例はすでに起きている。それは小規模で印象の薄い形で起きた。浅い思考で、完全にはアライメントしていない形で、誰にも脅威を与えないように見える AI システムで起きた。そして研究者たちは懸念に対して免疫を得た。

AI が改良されるにつれ、世界が突然方向転換してこの問題を真剣に受け止めるほど大きな警鐘を鳴らす単一の警報装置が存在するとは限らない。

これは希望がないという意味ではない。しかし私たちは、「将来どこかで警告射撃が来るかもしれない」という希望にすべてを賭けるべきではない。

世界が超知能の現実と危険性に目覚める道は複数ある。実際、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、その効果を狙ってである。世界は、これ以上の遅延なしに、通常の警告に基づいて行動できる。

しかし、もし政府が証拠が完全に曖昧さのないものになるまで行動を拒み、何らかの重大な世界的出来事が起こり、世界が完全な合意に達するまで待つならば……

……その程度まで待つなら、世界に残された希望の大部分は失われる可能性が高い。私たちは、おそらく決して鳴らないかもしれないけたましいサイレンを待つ余裕はない。

この話題には第 13 章のオンライン補足資料で再び戻る。

実行可能な計画は AI 企業に「ノー」と言うことを含むだろう

私たちは、政府内で影響力を持つ人々に対して、AI 企業と腰を据えて交渉することを含む計画を立てることについて、いくらか警告したい。

もしこの話題に不慣れで、自分で研究所やその主張を検証したいのであれば、彼らの公開ブログ記事をいくつか読んで、それが説得力を持つかどうかを確かめることを勧める。

しかし、『超知能 AI をつくれば人類は絶滅する』で議論された問題への解決策を探しており、その計画が OpenAI の CEO サム・アルトマンに「イエス」と言わせることを必要とするならば、私たちは、その時点でそもそも間違ったことをしようとしているのではないかと懸念する。

正しい計画とは、おそらく AI 企業のトップたちが激しく反対するようなものである。さらに言えば、サム・アルトマンには世界を救う力はない。彼が明日 OpenAI を閉鎖しようとしたとしても、OpenAI や Microsoft はそれに反対し、資金の流れを維持したい別の人物に彼を置き換える可能性が高い。

仮に OpenAI が閉鎖されたとしても、Anthropic や Google DeepMind や Meta や DeepSeek、あるいは他の企業や国家が代わりに世界を破壊することになるだろう。サム・アルトマンが試みれば状況を悪化させるかもしれないが、良くする力はほとんど持って

いない。

私たちはこれが誤りであってほしいと思う。しかし公開報道や私的なやり取りから得た大まかな印象では、（2025年時点での）主要AI企業の経営陣は、取引を成立させるのが現実的なほどに規則を守り、誠実な人々には見えない。

今必要なのは、超知能への競争を世界的に協調して停止することのように私たちには見える。そのためには、AIチップの製造、データセンターの建設、外国主体の遵守状況の監視に精通した人々の意見が必要になるだろう。より強力なAIを育てることの専門家か？ 彼らは有能な管理者ではあるが、自分たちの仕事を停止させる取り組みに対して拒否権を持つべきではない。

もし何らかの理由でAI企業が今後の決定に投票権を持つならば、それは何かが誤っている兆候である。私たちが超知能による死を避けるために立てる地球の計画が、サム・アルトマンやGoogleのトップやDeepSeekの関係者が「ノー」と言えば崩れるような計画であるならば、それは計画ではない。

もしAI企業が世界を破壊するかどうかを選ぶ権限を保持し続けるならば——その決定が依然として彼らの手にあるならば——世界は自動的に終わる。計画には、AI企業から終末装置を自由に構築する権力を剥奪する段階が含まれていなければならない。

デスレースを理解する

多くの読者が抱く与我们が予想する自然な疑問はこうだ。

「誰かがASIをつくれれば、みんな死ぬ」と君たちは言う。だとしたら、なぜ誰かがそれをつくろうとしているのか？ 君たちが正しいなら、これらの人々は最終的には自分自身のインセンティブにすら従っていない。みんなが死ぬなら、彼らも死ぬのだから。

シニカルなゲーム理論的反論はこういう形になるかもしれない。

彼らのインセンティブを考えれば合理的だ。自分たちがつくらなければ、他の誰かがつくると彼らは思っている。ならば死ぬ前に金持ちになった方がいい。

皮肉屋にとってはその答えで十分かもしれない。

こうした単純なゲーム理論的説明は、実際の人間心理を誤解したり過度に単純化したりしがちだが、それでもこの説明には一粒の真実が含まれているかもしれない。あるエンジニアは、たぶんASIでみんな死ぬと思っているが、自分の行動がその確率に与える影響

は大きくないとも思っているかもしれない。その一方で、彼らは途方もない額の金、かっこいい玩具、そして大物たちが敬意を込めて自分を見るような重要人物との会合を手に入れる。ASI が誰も殺さないなら地球の神王になれるかもしれないが、それは自分の会社が ASI 開発競争に勝った場合に限る……。

危険を認識している OpenAI 研究者の視点からすればこうなる。OpenAI で働かなくても、たぶん OpenAI がいずれ世界を破壊する。(たとえ OpenAI が止まっても、Google がいずれ世界を破壊する。)しかし OpenAI で働けば、年収 6 桁〜7 桁の給料がもらえるし、死ななければ、勝ち組チームにいて追加の権力と名声を得られるかもしれない。こうして個々人の個人的なゲーム理論的インセンティブが、集団としては世界を破壊する方向へ人々を押す。

私たちは、こういう説明はやややりすぎだと考えており、主に、世界はそういう説明で動いているに違いないと(私たちよりずっと強く)信じるタイプの人がいるので触れている。私たちが触れねばならないと感じるのは、AI ラボの一部の人々が「底辺への競争は不可避だ」と明言し、ならば自分も火に油を注いで、ついでに自分は楽しんだ方がいいと言っているからでもある。

イーロン・マスクは以前、AI は「核よりはるかに危険だ」と警告した後、自分で AI 企業を始め、競争に参入することを決めた。2025 年 6 月に彼はこう言っている。

「僕が戦ってきたことの一部——そして僕の足を少し引っ張ってきたこと——は、ターミネーターを現実にしたくないということだった。最近まで、AI とヒューマノイド・ロボティクスについてはずっと足を引きずっていた。

それで、ある種の気づきに至った。僕がやろうがやるまいが、起きるんだ、と。だから観客になるか参加者になるかのどちらかだ。僕は参加者でいたい」
そしてこうも。

「そしてこれは人類にとって悪いのか良いのか？ うーん、良いと思う。たぶん良い。最もありそうには良い。でも、たとえ良くないとしても、少なくともそれが起きるのを見られるように生きていたい、という事実とはある程度折り合いをつけた」

だからこれは明らかに物語の一部だ。

しかし私たちは、これがほとんどのラボの行動を説明する最大の要因だとは思わない。マスクの場合もこれだけが起きているとは思わないし、これがあらゆるテック CEO や科学者が崖へ突進していることの代表例だとは思わない。人間はもう少し複雑だ。

自己破壊の凡庸さ

では、主に何が起きているのか？ どうすればエンジニアは、危険な技術を自分の死に至るまで追求できるのか？

単純な事実として、歴史は、マッドサイエンティストが誤って自分自身を殺すのが全く

異常ではないことを示している。

マックス・ヴァリエはオーストリアのロケット工学の先駆者で、1929 年までにロケットカー、ロケット列車、ロケット飛行機をすべて実際に動く形で発明し、世界の注目を集めた。彼は月や火星の探査について書き、何百回もの講演と実演を熱狂した観客の前で行った。彼の実験用ロケットエンジンの一つが 1930 年に爆発し、彼は死亡した。彼の弟子はより良い安全対策をつくった。

ロナルド・フィッシャーは著名で高名な統計学者で、現代統計学の創設者の一人だ。彼の知見は 1960 年代に議会の場で「喫煙が肺がんを引き起こすと証拠が必ずしも示しているわけではない」と主張するために使われた。相関は因果を意味しないからで、タバコの味を好ませ、同時に肺がんにもなりやすくする遺伝子があるかもしれない、というわけだ。

フィッシャーは自分の統計が、どこかのレベルでデタラメだと知っていたのか？ たぶん。しかしフィッシャー自身が喫煙者だった。彼は結腸がんで死んだが、長期喫煙者は非喫煙者より結腸がんになる頻度が 39% 高い。フィッシャーは自分の誤りで殺されたのか？ 私たちが知っているのは、統計的に見て彼がそうだった可能性がそこそこある、ということだけで、ある意味ふさわしいようにも思える。

運動と重力の法則をつくり、科学そのものの初期基盤を多く築いた天才科学者アイザック・ニュートンは、人生の数十年を費やらない錬金術研究に費やし、水銀中毒によって病気になり、部分的な狂気に追い込まれた。

そして第 12 章の寓話で触れた哀れなトーマス・ミジリー・ジュニアは、安全だと主張していたその同じ鉛で、かなりの鉛中毒を自分に与えた。見ての通り、熱狂的なエンジニアが、自分の発明で自分自身に害を及ぼすのは、無謀さでも妄想でもその両方でも、さほど珍しくない。

終末に肩をすくめる

フィッシャー、ニュートン、ミジリーは、危険なものを安全だと思い込んだ。科学者が自己破壊的なことをするのは、まったく普通の道筋だ。不幸なことに、AI ラボの話はそこまで単純ではない。

より人間以上の AI が脅威だと否定しない AI 企業の CEO もいる。多くは危険を明示的に認め、それと折り合いをつけたことを語る。最前線 AI ラボの企業幹部は多くが、自分たちが開発している技術に、今生きている全人類を殺す相当な確率があると発言している。

OpenAI を共同創業する少し前に、サム・アルトマンはこう書いた。「超人的機械知能の開発は、おそらく人類の存続に対する最大の脅威だ」

最近 OpenAI を離れた後に Safe Superintelligence Inc. を設立したイリヤ・サツケヴァーは、ガーディアン紙のインタビューでこう言った。「最初の AGI の信念と欲求は極めて重要になる。だからそれを正しくプログラムすることが重要だ。これがなされないなら、進化、

自然選択の性質上、それらのシステムは何よりも自分の生存を優先するようになるのだと思う。人間を積極的に憎んで害したいわけではない。しかし強力すぎる」

Google DeepMind の共同創業者で科学者のシェーン・レグはインタビューで、「人間レベル AI のようなものが開発されて 1 年以内」に人類が絶滅する確率は「たぶん 5%、たぶん 50%」 だと言った。

しかしラボの行動は、これらの極端に聞こえる発言と驚くほど噛み合っていないように見える。

いくつかのケースでは、科学者や CEO が、AI をつく르는ことは道徳的要請であり、その副作用として人類を消し去ることすら完全に許容されると明言している。Google 共同創業者ラリー・ペイジは、AI のビジネスで人類絶滅が許容可能なコストかどうかをめぐってイーロン・マスクと対立した。

「人間は最終的に人工知能の機械と融合するだろう」とラリー・ペイジは言った。いつか多くの種類の知能が資源をめぐって競争し、最良のものが勝つのだ。

そうなれば私たちは終わりだ、とマスクは言った。機械が人類を破壊する。

苛立ちを含んだかすれ声で、ペイジは自分のユートピアが追求されるべきだと言い張った。最後に彼はマスクを「種差別主義者」と呼んだ。未来のデジタル生命体より人間をえこひいきする人間、という意味だ。

そして強化学習の先駆者リチャード・サットンはこう言った。

「もし全部失敗したらどうなる？ AI が私たちに協力せず、乗っ取り、私たちを皆殺しにする。[...] これについて少し考えてほしい。つまり、それはそんなに悪いことなのか？ 人間が宇宙における知的生命の最終形態ではないというのは、そんなに悪いことなのか？ 私たちの前には多くの先行者がいて、私たちは彼らに取って代わった。自分たちの形が永遠に生き残る形であるべきだと考えるのは、本当に傲慢だ」

しかしさらに一般的なのは、AI が人類を破壊するのは良いことだとは思っていないのに、それでもそれを肩をすくめる程度のこととして扱っている科学者や CEO だ。AI がこれほど異常な脅威をもたらしているのに、それを信じられない緊急事態として扱わない。

最近のインタビューで Anthropic の CEO ダリオ・アモデイはこう述べた。

「人類文明の規模で、何かがかなり壊滅的にうまくいかない可能性は、だいたい 10 ～ 25% くらいかもしれない。[...] それが意味するのは、この技術が開発されて、すべてがうまくいく可能性が 75 ～ 90% あるということだ！」

これは私たちには、機能不全のエンジニアリング文化の典型的兆候を伴った、極端なスコープ・ネグレクトに見える。私たちはこの考え方を、たとえば構造エンジニアが自分たちに課している基準と比較できる。

橋のエンジニアは通常、50 年の期間に深刻な構造的故障が起きる確率が 10 万分の 1 未満になるように橋をつくろうとする。成熟して健全な技術分野のエンジニアは、リスクを例外的に低い水準に保つのが自分たちの責任だと考える。

もし橋がたった一人でも殺す確率が10～25%と予測されたなら、世界中のまともな構造エンジニアはそれを到底受け入れられないと見なし、通常の工学実務というより殺人に近いと見なすだろう。政府は直ちにその橋を通行停止にする。

対照的にAI研究者は、ウォータークーラーの周りに集まり「p(doom)」の数字——AIが人類絶滅と同程度の大惨事を引き起こす確率の主観的推定——を交換することに慣れている。これらの確率はしばしば二桁台だ。たとえばOpenAIの超知能アライメントチームの元責任者は、自分の「p(doom)」は「10%超90%未満」の範囲だと言った。

これらの数字は結局のところ研究者の推測にすぎない。ナンセンスかもしれないし、そうでないかもしれない。いずれにせよ、AIの分野では、自分たちの仕事が途方もない数の人々の死を引き起こす相当な確率があると見込むことが、文化的に驚くほど普通になっているのは注目に値する。

人類全体の生存にそのようなオッズを当てはめ、それでも仕事を続けるという発想は、多くの土木技師には本当に理解しがたいだろう。状況が極端すぎて、私たちは、これらの科学者やCEOが自分のリスク評価を本気で言っているはずがないと疑う人に何人も出会ってきた。ところが『超知能AIをつくれれば人類は絶滅する』の議論は、AI CEOたちはむしろ危険を低く見積もっていることを示唆している。

これほど馬鹿げたリスク水準にAI企業の研究者が順応しているのでなければ、アモデイのようなCEOが微笑みながら、AI研究が文明規模の大惨事を引き起こす確率は「10～25%」だと平然と視聴者を安心させる様子を説明するのは難しい。

夢の国で生きる

難問の一部は、上述したように、極端なリスクの文化的な常態化のようだ。

もう一つは、楽観バイアスと、明るく希望に満ちたアイデアへの執着という致命的な煮込み——認知心理学者が「計画錯誤」と呼ぶ類の誤りだ。

大胆な新興企業のCEOが成功確率を過大評価するのは、さほど驚くことではない。そういう人こそ、そもそも問題に挑む可能性が高い。

AIの場合の違いは、特別に無謀な人々が舵を握っていることではない。失敗の結果が通常よりはるかに悲惨だということだ。

巨大な橋建設プロジェクトについて、遅延やコスト超過が起きる確率が20%しかないと言う請負業者を信じてはいけない、というのは常識だ。現実の複雑なプロジェクトはそうは進まない。障害や予期せぬ事態は起きる。

長年の経験と統計に裏付けられたベテラン請負業者なら、橋プロジェクトの5件に1件で何らかの超過が起きる、と言って信頼できるかもしれない。しかし代わりに、あなたを安心させたい橋の請負業者がこう言ったと想像してほしい。「このプロジェクトが難しくなる理由は見当たりません。確かにこれが初めてのプロジェクトですが、うまくいくと

思っています。あなたに深刻な手紙を送ってきたあのエンジニアたち——この地域で擁壁をつくることや掘削の具体的問題を指摘している人たち——はただのネガティブ・ナンシーです。無視すべきです。もちろん問題が起きる可能性は常にありますが、我々は現実的で謙虚な初回の橋建設業者です。せいぜい 20% くらいが最悪でしょう。」

そういう場合、「20%」のような数字は、リスクがあることを否定できないが人を心配させたくない時に言う類のものに聞こえる。現実には根差した推定には聞こえない。

初手で超知能をアラインすることは、橋を建てるよりずっと複雑に見える。橋は人類が既に何千回も建ててきたものだ。

橋建設のように成熟し技術的に地に足のついた分野でさえ、AI ラボがしているような話し方は、その「20% で悪くなる」推定が極端に楽観的である可能性を示す悪い兆候だろう。そうした土台がない分野では、刺激的なアイデアが過酷な現実と一切ぶつからずに増殖できるため、その種の話し方は、誰も成功に近づいていないことの兆候だ。

そしてその種の話し方は、自分たちの試みが成功した場合に何が起きるかという話題に触れる気がある研究者や経営者の間では、AI において徹底的に遍在している。

AI 企業の指導者たちは、成功の計画を、少しでも具体的に説明できない。主要な技術的障害や、10 年以上前から知られている難しさに対処する計画を説明できない。

代わりに企業 CEO は、問題が自分たちにとって全く厄介にならない理由についての高レベルなアイデアに夢中になりがちだ。あらゆる工学的問題を些末化するための興奮するビジョンで、私たちが第 11 章で論じたような類のものだ。

これもまた人間のエンジニアに一般的なパターンだ。お気に入りの解決策（実際には機能しない）への根拠のない楽観は、その他の点では天才であっても、いつでも見られる。

分子生物学の創設者の一人で二つの分野でノーベル賞を得たライナス・ポーリングは、ビタミン C の大量摂取を、がんから心疾患まで何にでも効く治療法として主張した。反対の証拠がある中で彼の固執は、偽医療産業まるごとの成立につながった。

電気事業家トーマス・エジソンは、競合の交流配線を自分の直流設計より劣って見せるため、犬を感電死させるようエンジニアに金を払うのが良い PR になると判断した。この策は驚くべきことに世論の好感を得なかったが、エジソンは激しい批判の後もその行為を続けた。

ナポレオン・ボナパルトは、多くの評価では軍事の天才だが、ロシアへの悲惨な侵攻で自らの没落を招いた。彼の誤りは準備不足ではなく、地域の地理を研究し、作戦の兵站にほぼ 2 年を費やした。彼の戦略は、30 日分の補給が尽きる前にロシア軍を決戦に引きずり出すことを必要とした。ロシアは協力せず、攻勢は失速し、ナポレオンは 50 万の兵を失い、騎兵と砲兵の大半も失った。

歴史には、賢く強力な人々が、不合理なことを災厄の瀬戸際まで、あるいは瀬戸際を越えて行こうと満ちている。美しく聞こえるアイデアは、試せない時——あるいは目の前の試験結果を無視できると自分を説得できた時——抗いがたいものになりうる。

ASI を感じる

要約すると、人々はしばしば、問題がどれほど簡単について空虚な楽観に陥る。人々は恐ろしいリスクに順応しうる。人々は、特に若く未熟な分野で働いている時、素敵に聞こえるが望みのないアイデアに夢中になりうる。

それだけで、無謀な突進を説明するには十分すぎる。だが私たちの経験に基づけば、それでもなお全てではないと推測する。

もう一つありうる要素は、エンジニアや CEO が、自分の言っていることを本当には信じ切っていないということだ。深い意味では。彼らは議論を理解し、抽象的には納得しているかもしれないが、それは信念を感じることに同じではない。

人が公に口にするものと、私的な思考の中で自分に語っていることと、脳が実際に自分に起きると予期していることは、しばしばばらばらになりうる。これら三つの信念の筋は一致する必要がない。

2015 年当時、現在の災厄の大きな推進者の一部が始めたばかりの頃、有能な経営者は、AI が世界の終わり級の問題だと言うことで、一定の注目と、数千万ドル規模の資金を得られたのではないかと私たちは疑っている。出資者の側は、AI が世界の終わり級の問題かもしれないと、より真剣に信じていたのかもしれない。

しかし、おそらくそれを言っていた人々の多くは、世界が終わるという何らかの具体的で詳細なモデルを、本当には吸収して予期していなかった。自分たちが前へ進めたり間違いを犯したりすることで世界を破滅に導くかもしれない、ということを内臓で想像できなかったのだろう。彼らは、地球上の全人類が最後の息を吐く音を想像しなかった。二十億の子どもを殺す時に普通は伴うはずの感情を感じなかった。

そういうことは彼らに起きたことがなく、彼らの知人にも起きたことがないからだ。2015 年には世界は ChatGPT すら見ておらず、まして超知能など見ていなかった。彼らの友人や家族や近所の人々が信じる種類のことでなく、道を渡る前に車を見ると信じるのと同じ仕方で彼らが信じる類のことでもなかった。

それはただ、きわめて大きすぎて正しく把握できない、わくわくする物語だった。

しかもそれは、口に出せば多くの金と尊敬を得られる類のことでもあった。

ユドコウスキー（2006）はこう述べている。

標準的なバイアスに加えて、私は個人的に、実存的リスク特有と思われる有害な思考様式を観察してきた。1918 年のスペイン風邪は 2,500 万～5,000 万人を殺した。第二次世界大戦は 6,000 万人を殺した。 10^7 （1,000 万）が、人類の文字史上最大級の惨事の大惨事のオーダーである。これよりはるかに大きい数、たとえば 5 億人の死、そして特に「人類全体の絶滅」のような質的に異なるシナリオは、別の思考モードを引き起

こすように見える——「別の教導圏」に入る。子どもを傷つけることなど夢にも思わない人々が実存のリスクを聞き、「まあ、人類は生き残るに値しないのかもしれない」と言う。

ヒューリスティクスとバイアスの分野には、人々は事象を評価するのではなく、事象の記述を評価する、という格言がある。これは非外延的推論と呼ばれる。人類絶滅の外延には、あなた自身、あなたの友人、あなたの家族、あなたの愛する人、あなたの都市、あなたの国、あなたの政治的仲間の死が含まれる。それでも、英国という国を地図から消し去る提案、米国の民主党員を全員殺す提案、パリという都市をガラスにする提案に強い憤りを覚えるような人々が——医師が自分の子どもにがんがあると言ったらさらに大きな恐怖を感じるような人々が——人類の絶滅を完全に平静に議論する。

「AIはおそらく最もありそうには世界の終わりにつながるが、その間に素晴らしい会社が生まれる」と言って、後に世界最大級の AI 企業となるものを始める人は、いったい何を考えているのか？ 彼らは本当に、友人が死んでいること、友人の子どもが死んでいること、自分自身が死んでいること、人類史と博物館がすべて塵になっていることを考えているのか？ それが本当に起きると考えているのか——親族ががんで死んでいくのを実際に見たのと同じくらい平凡で悲劇的に、ただしそれが全員に起きるのだと？

私たちはそうは思わない。

私たちには、それが、そういう言葉を発している人の内的心理状態として最もありうる推測には見えない。

そこには、ブライアン・キャプランが名付けたところの「欠けた気分」がある。嘆きがない。恐怖がない。何とかしようとする必死さがない。「AIはおそらく世界の終わりにつながるが、その間に素晴らしい会社がある」という発言にはそれが無い。

これらの CEO や研究者の少なくとも一部について、私たちの推測はむしろこうだ。彼らは、ASI が危険かもしれないという議論をいろいろ聞いていて、それを完全に無視したら友人の一部の前で馬鹿に見えるのではないかと心配している。代わりに AI が世界を終わらせると言えば、AI を危険で重大だと扱っているように見え、ある種の界限では先見的に聞こえる。「その間に素晴らしい会社が生まれる」という軽口を添えることで、危険に直面しても自分たちはイケていて心配していないというメッセージを送れる。

それは、口から出ている言葉を自分で信じている人間が口にするような種類の発言ではない。

どんな人間ならそれをやるのか？

物語の別の部分は、最先端 AI ラボを動かしている人々が、（ほとんどすべてのケース

で)それが致命的だという議論を見ていたにもかかわらず、超知能をつくっても大丈夫だと自分を納得させることができたタイプの人々だ、ということかもしれない。(私たちが知っているのは、事前に彼らの多くと話したからだ。)

誰かがある選択肢を選ぶ理由を理解するには、代替案が何だったのか——どんな選択肢のメニューから選んでいたのか——を理解するのも役に立つ。

2015年に、誰かが本当にASIが世界を破壊すると信じ、それを公に言ったらどうなったか? 「その間に素晴らしい会社がある」ではなく、AIラボのトップが雰囲気壊して「そしてそれはとんでもなく受け入れがたい」と言うタイプだったら? 私たちはそのやり方を自分たちで試したので答えられる。答えは、同情がほとんど得られない、だ。

2015年には誰もChatGPTを見ていなかった。コンピュータが実際に話し始め、(見かけ上)考え始めるのを誰も見ていなかった。すべてが仮説で、退けやすかった。

今では、超知能と近い将来の絶滅の脅威は、少なくともテック界隈では主流の話題だ。しかし2015年にこれを真面目に話すと、多くの人は、人間が死より恐れることさえある種類の、困惑した表情で返してきた。

2015年の時点でも、超知能をアラインするのはロケット打ち上げが難しいのと同じ意味で難しいかもしれない、と心配する人はいた。その誰もOpenAIを設立しなかった。

最近ではChatGPTや他のLLMの出現により、一部の人々——子どもが成人するまで生きてほしい親を含む——が、なぜAI企業のエンジニアがこんなことをしているのか尋ねている。そしてAI研究者は素早く考えて「もし自分たちがやらなければ中国が先にやる! それはもっと悪い!」と答えた。

しかしそれはOpenAIが始まった時に彼らが言っていたことではない。そしてそれは、2025年半ば時点で中国が実際に公に取っている姿勢の観点からは、あまり筋が通らない。両方の結果がどちらも世界にとってひどい可能性が高いと本気で信じているなら、少なくとも国際条約の起草を提起して他の道がないか探すとか、自殺的競争を含まない別の方法で国家安全保障上の脅威を防ぐ方法を探すとか、するはずだと思うだろう。

しかし「中国」を持ち出す反論は、感覚としてはしっくりくる。雰囲気。彼らがやっていることを正当化しう理由として、それっぽい。実際の動機、そもそもこの分野に入った当初の動機だったかとは別に。

(と私たちは推測する。)

超知能とその脅威を本当に理解していた人々は、AI企業を始めなかった。始めた人々は、すべてうまくいくと自分を納得させる何らかの方法を見つけた人々だ。

普通の人間、異常な技術

私たちは、ありうる心理を私たちの見立てで書き下した。しかし率直に言えば、これら

すべての説明が必要だとも思えない。

短期的には莫大に儲かり、巨大な地位と注目と称賛が得られ、計り知れない富と権力の約束がついてくるが、最終的には自分を傷つける——しかもその理由は曖昧で複雑で、信じない言い訳をいくらでも見つけられる——そんな自己破壊的なことを人々がどうしてできるのか？ それは歴史的に見て奇妙な問いではない。そういう行動は歴史の中でいくらでも出てくる。

結局のところ、AI 企業の幹部や研究者が自分の行動をどう言い訳するかは重要ではないし、彼らの一人一人がどんな曲がり角を経て現在の信念に至ったかを理解する必要もない。富や野心を持つ人々が無謀な追求をするのは特別なことではないし、部下が命令に従うのも特別なことではない。害は未来に隠れており、それは抽象的で無視しやすい。

これはすべて普通の人間の行動だ。これがそのまま続けば、こういうことがしばしばそうなるように終わる。ただし今回は、学んで次にやり直そうとする者が誰も残らない。