

10 章 呪われた問題

ASI アライメントとは、人工超知能（ASI）から有用な成果を、確実に、しかも破滅を引き起こすことなく引き出すという課題である。この問題にはいくつかの基本的な性質があるため、きわめて困難に見える。

以下では、『超知能 AI をつくれば人類は絶滅する』第 10 章を読んだ人向けの補足的な問いを扱う。歴史的な比較がどれほど示唆的かについて、より詳しく検討し、また問題をより容易にし得るシナリオの提案についても考える。本書の内容の繰り返しを避けるため、ここでは次の話題は扱わない。

- ・工学的問題を困難にする要因は何か。
- ・人類は歴史の中でどのような困難な問題に直面してきたか。そして ASI への道筋を考えるうえで、そこからどのような教訓を得られるか。
- ・最初から困難な問題に直面していると分かっている場合、何ができるのか。本当に困難な問題に直面するとき、どのように振る舞うべきか。

AI は歴史的 precedent とは異なるのではないか？

はい。

AI アライメントの課題には、原子力発電所の設計よりも容易にする独自の特徴もあれば、より困難にする特徴もある。総合的に見ると、核兵器や原子力発電所は、超人的知能の AI よりもはるかに管理しやすいように思われる。

業界の人々は、AI そのものに AI のアライメントを手伝わればよいとすぐに指摘する。私たちは、これはあまり決定的ではないと考えている。なぜなら、超知能をアライメントさせることは極めて困難な問題であり、解決策を評価したり進捗を測定したりする良い方法が存在しないからである。つまり、この問題を手伝うには、AI はすでに非常に有能で、しかも高度にアライメントしていなければならない。この提案については第 11 章でさらに論じる。

また、AI アライメントが原子力発電所の設計より容易であり得る点として、人間が構築する AI の動作様式をかなり高い程度まで制御できる可能性があることが挙げられる。原子炉を支配する物理法則は選べないが、人間が AI を設計し、その仕組みを完全に理解しているなら、AI の認知ダイナミクスについて多くの選択が可能になるだろう。もっとも、第 2 章で述べたように、現実には誰もその水準の理解には到底達していない。

では、AI が人類がこれまで直面してきた問題よりも難しい課題となる理由を、核兵器

と比較して考えてみよう。私たちは、この比較は、超知能 AI の方がはるかに厄介な問題であることを示唆していると考える。その理由は次のとおりである。

1. 核兵器は人類より賢くない。
2. 核兵器は自己複製しない。
3. 核兵器は自己改良しない。
4. 現実的な核戦争シナリオの多くでは、人類が完全に絶滅するとは想定されていない。おそらく廃墟の中に生存者が残り、再建が可能である。
5. ベンチャー資金を受けた企業が、世界の核兵器備蓄を毎年 10 倍に拡大しているわけではない。
6. 核兵器の科学はかなりよく理解されている。技術者は建造前におおよその威力を計算でき、壊滅的爆発を引き起こす連鎖反応に必要な核分裂性物質の濃度も正確に知っている。核兵器は自ら計画を立てない。ある国が核兵器を保有すれば、その国がそれを所有する。科学者は、核兵器が自分たちよりはるかに賢くなり、所有されることを拒否するのではないかと心配する必要はない。
7. 世界は概して、核兵器が爆発すれば人が死ぬという点で合意している。物理学者コミュニティが、「全員が核兵器を持てば悪人の核兵器に脅かされない」や、「心配するな、人間は核兵器と融合する」や、「核戦争は不可避なのだから止めようとするのは幼稚だ」といった奇妙な立場に分裂しているわけではない。
8. 核兵器は複製が困難である。誰もが利用できるレンタル技術として核兵器製造能力を提供する巨大な技術的取り組みは存在せず、実験室で 1 発つくれば 1 週間後に 10 万発配備できるというものでもない。
9. 主要国は核戦争を現実的かつ容認できない可能性として扱っている。世界の指導者たちはそれを避けるために真剣に努力している。最も利己的な指導者であっても、核戦争が自分や家族を殺し、自分にとって大切な場所や財産を破壊し得ることを理解している。市民や有権者も核戦争を望んでいない。核戦争に反対する点において、人類はこれまでになく一致している。
10. さらに悪いことに——本書および下記の〈より深い議論〉で述べたように——超知能に関しては人類にやり直しはない。原子力発電所が爆発すれば、他の人々がそこから学び、次はより良くなる。

これらすべての特徴は、超知能が並外れて困難で、しかも並外れて新しい課題であることを示唆している。アナロジーは存在するが、それらは限定的にしか当てはまらない。ASI に関する確立されたブレイブブックは存在しない。

ASI アライメント問題の解決にはどれくらいかかるのか？

問題は単に時間が足りないことではなく、誤りが致命的であることだ。

「どれくらい時間がかかるのか？」という問いは答えにくい。というのも、ある意味で本質を外しているからである。研究者が非生産的なやり方で問題に取り組んでいる場合、その問題が原理的には迅速に解けるものであっても、永遠に「行き詰まった」ままにいる可能性がある。今日の時点で、どのようにすれば「行き詰まり」を抜け出し、信頼できる進展を遂げられるのかを明確に描くのは難しい。しかしそれは、科学や工学の通常のやり方をかなり根本的に変えることを必要とするように見える。

では、科学や工学は通常どのように進むのか。

西暦 500 年頃までに、太陽が地球の周りを回っているという理論に古代の思想家たちは収束していた。コペルニクスはそれに対抗する理論を提案したが、それは検討され、概ね退けられた。ガリレオが望遠鏡をつくり、木星の衛星——地球ではなく木星の周りを回る天体——を観測するまでは、新興の科学共同体は地球が太陽の周りを回っているという結論に強く動かされなかった。

人類はやがて正しい軌道力学の理論に到達した。しかしその前に、誤った合意に達していた。そして、地球が宇宙の中心ではないという事実を現実がガリレオに叩きつけるまで、その誤った合意を自信満々に保持していた。

科学共同体が真理に収束する通常の過程には、共同体が誤っている段階が含まれている。そして現実が証拠を突きつけ、私たちがモデルを更新するよう強いる。

ASI アライメントの問題は、単に難しい研究課題だというだけではない。「最初に好まれた理論が誤っていた」という事実を現実が叩きつける出来事が、この場合には「ASI が地球を消費すること」に相当するという点にある。その場合、より良いアライメント理論に収束するための生存者は残らない。

もし人類に 100 年の時間と無制限のやり直しがあれば、ASI アライメント問題を解決するのはそれほど難しくないかもしれない。

しかし、仮に知能の理論、AI が賢くなるにつれてどのように変化するか理論、そして AI を特定の目標に安定的かつ頑健に向ける方法の理論を発展させるために 300 年あったとしても——AI が何度か飛躍的に賢くなったとき実際に何が起るかを試して確かめる能力がない限り——私たちはおそらく誤った答えに収束するだろう。

人類は、はるかに単純な領域においてさえ、決定的な実験を行う機会がないまま、誤った答えに収束する傾向がある。

もし多くの異なる AI が存在したらどうか？

それらのどれにも「良いことを氣にかけさせる」ことができないのなら、大して役には立たない。

第4章で述べたとおり、AI が実際には誰も望んでいない奇妙な目的を氣にかけるようになる道筋は数多く存在する。人類が千の AI をつくったとしても、その千の AI がそれぞれわずかに異なる奇妙な目的を氣にかけているのであれば、意味はないのである。

多様な AI の世界が、幸福で健康で自由な人々を恒久的に宿すためには、それらの AI のうち少なくともいくつかは、繁栄する人々が存続することを明確に望んでいなければならない。単に、彼らが若いときにそれを氣にかけていると私たちに保証するという意味ではなく、第5章で論じたように、彼らの行動（あるいはその子孫の行動）が何らかの問いへの応答であるとき、その問いに対する最も効率的な答えが実際にそれである、という意味においてである。

この問題に関するさらなる背景については、以下を参照のこと。

第4章：「なぜ AI は、訓練された方向以外へも舵を切るのか？」「内省と自己改変は、すべてをさらに難しくする」

第5章：「AI は私たちを有用だと見なして残すだろうか？」「AI には法の支配が必要ではないか？」「AI は少なくとも多少は人間を氣にかけるのではないか？」「人間はほとんどの場合、最も効率的な解ではない」

第2部：「なぜセイブルほど賢い AI を1体だけ登場させたのか？」

もし私たちが、十に一つの割合で AI を良いものにする方法を知っていたなら、おそらく非常に多くの異なる AI をつくり、善い AI が私たちのために取引を成立させることを期待することで、宇宙の十分の一を手に入れることができるかもしれない。

しかし、第2章から第4章で論じたとおり、AI が人々をちょうど適切な仕方で氣にかけるようにすることは、現代の「育てられる」AI の体制においては、きわめて起こりにくいのである。人類は、そのような例外的に狭い標的を10%の確率で命中させられる段階にはまったく達していない。問題が、私たちが AI の動機へと接続することすらほとんどできていない複雑な構造にある以上、AI の数を増やしても助けにはならないのである。

AI がゆっくり開発され、徐々に社会に統合される場合はどうか？

そのシナリオもなお、破局に終わる。

もし AI が何らかの未知の臨界閾値を通過し、きわめて急速に知性を高めるならば、それはアライメントをとりわけ困難にする。しかし、速度こそが中心的困難の源なのではない。中心的困難は、「前」と「後」との差異にあるのである。

仮に、人類が人間とほぼ同程度に汎用的な AI と「統合」するための数年を得るとしよう。

（この地点を超えた後の「遅い」進歩を論じる研究者たちは、一般に、AI があらゆる人間よりもはるかに知的かつ強力になるまでにせいぜい数年しかないと想定しているのであって、数十年ではない。しかし、仮に相当数の研究者が数十年を見込んでいたとしても、数十年で十分であるかどうかは自明ではない。）

その数年のあいだに、研究者は AI の仕組みについてさらに多くを学ぶであろう。さまざまな問題が発生するのを目にし、さまざまな回避策や、被害を軽減する方法を見出すであろう。AI の進歩は継続し、やがて AI は集散的に、権力を掌握するのに十分な能力を持つようになるであろう。これは、このシナリオでは AI がただひたすら賢くなり続けるという事実から帰結する。

その時点で、研究者たちは、支配できない AI をより信頼可能かつ収益的にするために編み出したさまざまな小技やハックの蓄積を持っているであろう。彼らは、その既存の小技やハックの集合が、支配可能な AI から支配可能でない AI へと一般化することを必要とする。そして、それが最初の試みで成功することを必要とするのである。

最終的には、十分に高い能力を持つ AI は致命的に危険である。そして、私たちが実験的失敗から学べるのは、それが致命的でない場合に限られる。もし AI の進歩が閾値効果を示すのであれば——それはおそらくそうであるように見えるが——この問題はいつそう絶望的に見える。しかしそれは、巨大な数の不透明で人間を超える知性の AI をゆっくりと育てれば、すべてがうまくいく現実的可能性があることを意味するものではない。

どこかに困難な転換点が存在せざるを得ないことを示す一つの方法は、人間の利益を実際には気にかけていないにもかかわらず、人間に奉仕している数十億の超知的 AI に満ちた世界を考えてみることである。実際には、私たちが現代の機械学習に類するものを用いて AI を構築していると仮定すれば、世界がそのような状態に到達することは決してないであろう。事態はそれよりはるかに以前に破綻するであろう。しかし、正確にいつ、どのように破綻するかを予測するのは困難である。極端な事例を考えることで、ある時点で物事が破綻することが明らかになるのである。

アナロジーで言えば、あなたの都市が竜を奴隷化し、畑を耕させ、コーヒーを焙煎させ

ようと計画しているとするならば、竜がきわめて急速に成長する場合、いくつかの問題はより悪化する。しかし中核的問題は、「都市をひと息の恐るべき炎で消し去ることのできる巨大な火を吐く怪物が、急速に成長したらどうするのか」ということではない。

もしあなたが、敵対的な成熟した竜で満ちた都市に自らがいると想像するならば、その状況はきわめて急速に悪化するであろう。たとえ、竜を統制するための巧妙な計画——たとえば、人間が財産の大半を所有するという法制度を与え、ゆえにいかなる竜の連合も自らの権利が無効にされることを恐れて代替的な法制度を提案することはないであろうと考えるような計画——を持っていたとしても、うまくいかないであろう。あるいは、竜同士を巧みに対立させ、他の竜を恐れてどの竜も火を吐こうとしないようにする計画を立てていたとしても、うまくいかないであろう。竜が完全に成熟したとき、彼らは互いを見やっとうなずき、そしてあなたを焼き尽くすであろう。

もしAIが比較的ゆっくり発展するならば、その未来はしばらくのあいだ興味深く奇妙なものになるかもしれない。しかし最終的には、権力と資源がAIへと流れ込む盆地に到達するであろう。現実的に言って、人間が歩調を合わせ続けることはない。ひとたびそれらのAIのいずれかの集団が、地球の資源を自らのために取得できる位置に至ったならば、それが帰還不能点である。そのAI集団の中に、幸福で健康で自由な人々を深く気にかける構成要素が含まれているか、さもなくば私たちにとって未来は悲惨なものとなるのである。

〈より深い議論〉

「前」と「後」をより詳しく見る

本章で述べたとおり、AI研究者が直面している根本的な困難は次の点にある——AIがあなたを殺せるほど強力かつ有能になる前に（あるいはアライメントに抵抗できるようにする前に）、そのAIをアラインしなければならぬ。しかもそのアライメントは、超知能、あるいは複数の超知能が、望めばあなたを殺せるようになる「後」の条件へと持ち越されなければならない。

言い換えれば、超知能を構築するなら、それをアラインしなければならないが、その際、本当に重要な現実条件のもとで自分のアライメント技術を徹底的にテストすることは一度もできない。たとえ、あなたが殺されるほどの力を持たないシステムを扱っているときには、自分の研究がどれほど「経験的」に感じられようとも、である。

これは、AI研究者が慣れている基準ではないし、ほとんどあらゆる分野のエンジニアも慣れていない。

私たちはしばしば、非科学的であり、経験的観察から切り離されているものを要求して

いる、と不満を聞く。それに対して私たちは、第 10 章で述べた宇宙探査機の設計者たちと話してみることを提案するかもしれない。

自然は不公平であり、時には「本当に重要な環境」が「テストできる環境」と一致しない状況を与える。それでも、頑健な道具と強い予測理論、すなわち自分が何をしているのかについての確固たる理解を備えている場合には、エンジニアが初回で正しくやり遂げることもある。——これは、AI 分野では明らかに欠けているものである。

問題全体はこうだ。安全にテストできる AI、すなわち、失敗テストがあなたを殺すことのない AI は、すでにテストされていなければならない AI (あるいは AI エコシステム) とは異なるレジームで動作している。なぜなら後者は、もしアラインされていなければ、全員が死ぬからである。前者の AI、あるいは AI システムは、「望めば全員を殺せる」という現実的選択肢を自分にあるとは正しく認識していない。後者の AI、あるいは AI システムは、その選択肢を見ている。

例えば、あなたが同僚のボブを自国の独裁者にすることを検討しているとする。まず彼を町の「模擬独裁者」にして、権力を乱用するかどうかを見ることはできる。しかし残念ながら、これはあまり良いテストではない。「軍に議会を威圧させ、次の選挙を『監督』させる」と命じることと、「町民に観察されながら（まだ殴られたり職を失ったりしうる状態で）模擬的権力を乱用する」とことは、まったく別の選択肢である。

十分に発達した認知理論があれば、AI の心を読み取り、もし本当に世界を乗っ取る機会があると信じたなら、どのような認知状態に入るかを予測しようとすることもできる。

また、シミュレーションを構築し (AI の内部感覚を欺くなどしながら)、あなたの認知理論が予測するところでは、AI が実際に裏切る選択肢を持ったときに入るであろう認知状態に非常によく似た状態を再現することもできる。

しかし、あなたが実験室で誘発し観察するその状態と、AI が本当にあなたを裏切る選択肢を持っている状態との間の連結は、根本的に、未検証の認知理論に依存している。AI の心は、超知能へと発展するにつれて、かなり変化する可能性が高い！

AI が自分より賢い後継 AI をつくるなら、それらの AI の内部構造は、あなたが以前研究した AI の内部構造とは異なる可能性が高い。「前」の心からしか学んでいない場合、その知識を「後」の頭脳へ適用することは、前と後のあいだで心がどのように変化するかについての未検証理論を経由する。

AI を、本当にあなたを裏切る機会がある状態まで動かすこと——しかもそれが偽装しにくい形で——は、研究室環境とは根本的に異なる環境における、それらの理論の経験的テストである。

多くの科学者 (そして多くのプログラマー) は、根本的に新しい動作環境において複雑なシステムがどう機能するかについての自分の理論が、初回からうまくいくことはしばしばない、と知っている。これは、異常に低い理解水準の領域において、不公平ほど高い予測可能性・制御・理論的洞察を要求する研究課題である。しかも、実験結果がエンジニ

アの希望を否定すれば、私たち全員の命が賭けられている。

この観点から見ると、研究者がAIのフロンティアを可能な限り押し広げようと急ぐべきではない、ということは、過剰決定されているように思える。これは正真正銘、狂気じみた試みであり、いかなる政府がそれを許すにしても正真正銘、狂気じみたことである。

シカゴ・パイル1号の物語

1942年、シカゴ・パイル1号（Chicago Pile-1）は、エンリコ・フェルミの指揮のもとで建設された。これは45,000個の黒鉛ブロック（総重量330トン）、4.9トンの金属ウラン、41トンの二酸化ウランから構成され、シカゴ大学のスタッグ・フィールドのラケットボールコートの観客席の下に設置された。用語の定義によっては、これを最初の原子炉と呼べる。工業用の発電を目的としたものではなかったが、持続的な臨界反応を起こす最初の装置だった。

現代の基準からすれば、安全面でいくつかの手抜きがあった。たとえば、大学の構内、しかも大都市の中にあるラケットコートの観客席の下に建設されたという点である。

マンハッタン計画全体の責任者であったグローヴス将軍は、実験をシカゴ市内ではなく近郊で行おうとし、そのための建物の建設を命じていたが、工事が遅れていた。CP-1を受け入れたシカゴ大学のノーベル賞物理学者アーサー・コンプトンは、大学学長に許可を求めなかった。後に彼は、学長に尋ねれば学長は「ノー」と言わざるを得なかっただろうし、それは間違った答えだったからだ、と説明している。

黒鉛やウランのブロックを積み上げる作業は、徴兵を待ちながら小遣いを稼ごうとしていた高校中退者たちが行った。

ウランは金属製の原子炉容器ではなく、7メートルのゴム製立方体に囲われていた。当然、巨大なコンクリート製格納建屋もなかった。

これらの事実を後に知った国家防衛研究委員会委員長ジェームズ・コナントは、顔面蒼白になったと言われている。1940年代でさえ、これは完全に「普通の」科学的行為とは見なされていなかった。

もしこの話を、結末を知らずに歴史書で読んでいたら、大事故の前兆だとも思うかもしれない。2025年の文化が標準的安全対策と考える多くのものが欠けている。検査官やチェックリストはどこにあるのか。分厚い運転規則集は？慎重に議論する委員会は？影響評価書は？「高度な資格を持つ者のみがウランブロックを積める」という規則は？書類はどこだ？

しかし、ウランと黒鉛の積み上げは暴走しなかった。

その理由は、フェルミが自分のしていることを理解していたからである。彼は事前に法則を予測していた。

フェルミは、近づけると熱を多く出す「謎のブロック」をただ積んでいたわけではない。彼は、ウラン原子の一部が自発的に崩壊し核分裂することを知っていた。核分裂が起これば中性子が放出されることも知っていた。その中性子が別のウラン原子に衝突し、さらに核分裂を引き起こすことがあることも理解していた。

彼は、これは指数過程だと事前に理解していた。ここでいう指数とは、今日のメディアが「大きい」「速い」という意味で乱用するものではなく、増加率が現在の値に比例する、数学的な指数関数的増大である。

ウランと黒鉛を積み上げることで、指数過程の中で掛け算される係数を増やしていることをフェルミは知っていた。本書で述べたように、中性子増倍率が100%未満か100%を超えるかには、決定的な違いがある。100%未満なら、温かいブロックの山で済む。しかし100%を超えると、放射能レベルは上がり、上がり、上がり続ける。

それは、これまで試したより小さなウランの山と同じ振る舞いはしない。もし炉を過度に減速させていなければ（過熱時に連鎖反応が弱まるようにしていなければ）、それは小さな山のように安定はしなかった。夜通し動かせば、翌日には「工業的に有用な出力」になるのではない。

ただひたすら放射能が増え、黒鉛が発火するか、ウランがスラグ状に溶けるまで続く。消防士が駆けつけても、水をかけても熱を出し続ける奇妙な火災に直面するだろう。

1942年は、シカゴ大学に通うには良い年ではなかっただろう。

だがフェルミはそれを知っていた。だから大丈夫だった。1942年12月2日、制御棒（カドミウム板を釘付けした木の板）をさらに12インチ引き抜くよう命じたとき、彼は事前に「これで放射能は上昇し、上昇し続ける……平衡にはならない」と予告した。

その後、放射能は2分で倍になり、さらに倍になり、28分間、2分ごとに倍増し続け、約16000倍に達した。

16000倍という増加は、予測されていた振る舞いであり、詳細に理解されていた。これは偶然の発見ではない。単に前回より10倍多くブロックを積んで、何か面白くて儲かることが起こるか試してみた、というものではない。

本書で述べた通り、原子炉と核爆発の間には非常に狭い余地しかない。正確には0.5%強である。それが、工業的出力を出す炉と爆発する炉の違いである。

つまり、核反応をまず強めていかなければ、そもそも機能しない。そして、やっと機能し始めたその瞬間から、ほんのわずか——0.65%——強めすぎると爆発する。

現実には、そういう問題を突きつけることがある。

しかしフェルミとシラードらのチームは、遅発中性子と即発中性子の違いを理解し、事前に法則を予測していた（詳細は第10章参照）。中性子増倍率が100.06%になったとき、フェルミはそれ以上制御棒を引かなかった。100%の臨界までで止め、0.65%先の即発臨界には進まなかった。彼は予測した結果を得た。そしてそれ以上進めば何が起きるかも知っていた。だから進まなかった。

28 分後、放射能が 2 分ごとに倍増し、16,000 倍に達したところで、フェルミは世界初の原子炉を停止した。

もちろん、フェルミが低エネルギー炉物理の自己整合的モデルを持っていたからといって、完全に責任ある行為だったとは限らない。彼は間違っていた可能性もある。核工学の歴史には予想外がある。

最初の水爆実験「キャッスル・ブラボー」は、予想の 3 倍の爆発力を示した。燃料にリチウム 6 とリチウム 7 の混合物が使われており、リチウム 7 が不活性だと誤解されていたためである。

フェルミは、工業的出力レベルではなく低出力で運転したため、多くの複雑性を回避できた。もし予期していない中性子増倍率増加要因があったら——16,000 倍の中性子束増加で 1.0006 が 1.02 に跳ね上がり、人間が緊急カドミウム棒を挿入する前に暴走したら——今日アメリカには「シカゴ立入禁止区域」があっただろう。

それでも私たちは、フェルミが実験すべきでなかったとは言わない。それは人類を滅ぼす実験ではなかった。1942 年当時、ナチス・ドイツが核兵器を得るかどうかは分からなかった。都市外で実験することは戦時中の不便であり、不便にもコストがある。

私たちの目的は道徳的裁定ではない。

ここから引き出す教訓は、「典型的な安全」と、現実があなたを殺さないために本当に必要なものの違いである。

CP-1 には、官僚が要求するような目立つ安全対策がほとんどなかった。災害を防いだのは理解であって、安全劇ではなかった。フェルミの理解は十分だった。足りなかった可能性もあるが、現実には十分だった。そして現実が要求したのはその理解であり、形式的な安全演出ではなかった。

もし誰もウランブロックの内部で何が起きているかを深く理解していなかったなら、立派なスーツを着た検査官が金属ブロックを覗き込んでも、「認定操作者のみが扱える」と書かれた分厚い手引書を印刷しても、ほとんど助けにならなかっただろう。

フェルミのいない世界を想像できる。誰も自己加熱ブロックの真の法則を理解していない世界。

その世界で、別の科学者が危険を察知していたかもしれない。たとえば、

サルヴィアティ：ブロックを近づけると出力が跳ね上がるのは、自己強化過程の明らかな兆候だ。数学モデルをつくれば、十分押せば爆発するモードがあるはずだ。

シンプリチオ：ナンセンスだ！ 現実ではどんな過程もやがて限界に達する。無限に続くわけがない。だから安全だ。

サルヴィアティ：超新星は無限に熱くならないから危険でない、超知能は無限に賢くないから無害だ、弾丸は速度に限界があるから皮膚を貫かない、と言うようなものだ。限界があるからといって、その限界が低いとは限らない。

シンプリチオ：閾値について科学者も一致していないではないか。合意がないなら心配する必要はない。

サルヴィアティ：致死的爆発の可能性があると有力科学者が警告しているのに、正確な閾値が分からないことは、安心材料ではなく不安材料だ。私たちはまだ理解していない。理解するまで遊ぶべきではない。

私たちは、フェルミが核連鎖反応を理解していた水準の何分の一も、AI をモデル化できていない。

この道を進み続ければ、やがてシカゴ被曝よりはるかに重大な結果に、猛スピードで突入することになるだろう。