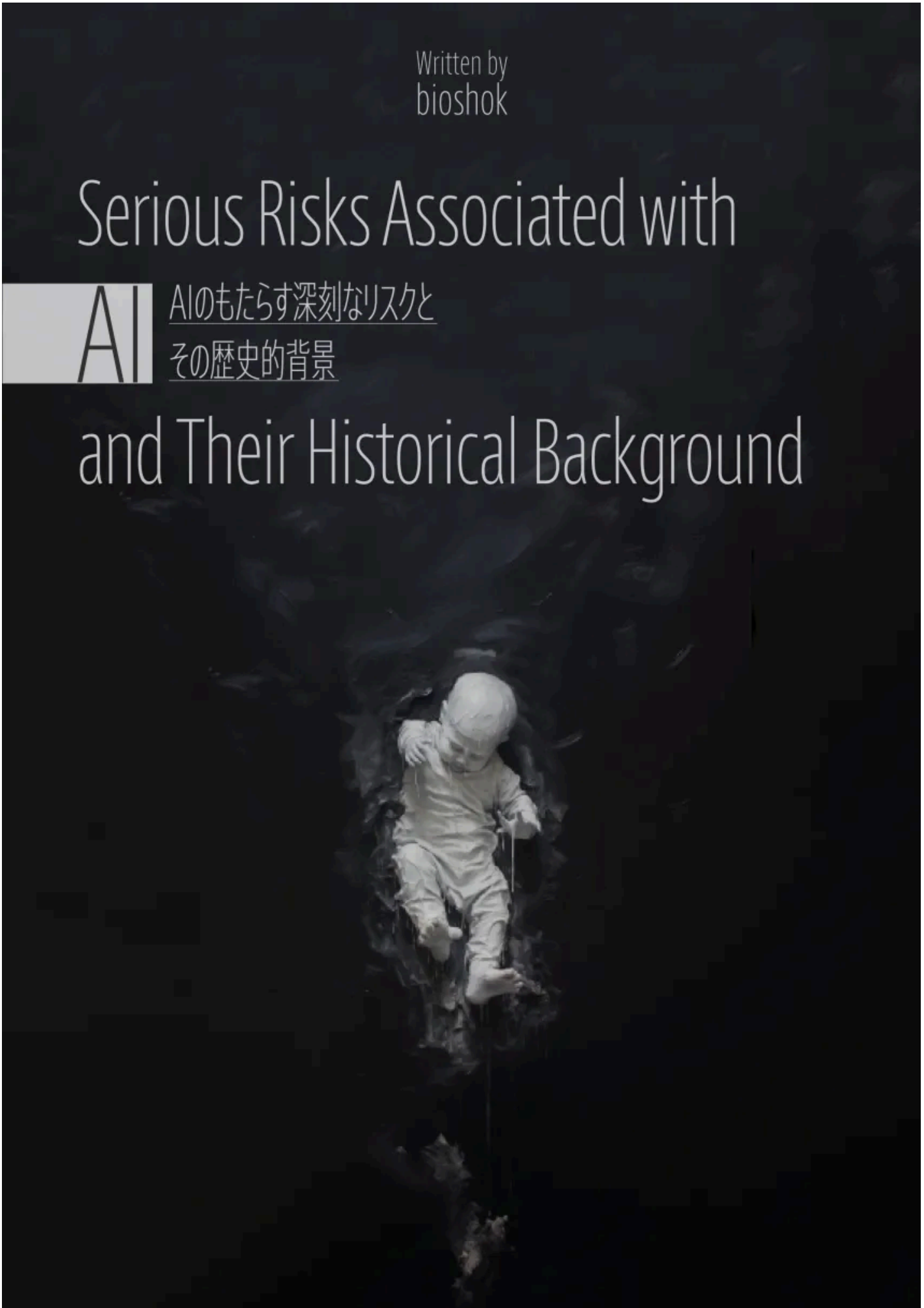
Written by
bioshok

Serious Risks Associated with AI and Their Historical Background

AIのもたらす深刻なリスクとその歴史的背景



AIのもたらす深刻なリスクとその歴史的背景

2024年1月30日 X account @bioshok3 執筆

I saw the work performed by intelligence; smart was no longer a property, but an engine.

私は知性による仕事を目の当たりにした。賢さはもはや財産ではなく、エンジンだった。
(Eliezer Yudkowsky)

My Naturalistic Awakening

概要

私たちが今いる21世紀は過去人類が体験したことのないような技術の発展の最中にあるかもしれず、特に近いうちに高度なAIが開発され、様々な意味で劇的に世界が変わる可能性があります。

その一方で今世紀に人類が何らかの要因で壊滅的な結果や人類絶滅を含む存亡的破局を迎えてしまった場合は我々人類の未来の可能性が失われてしまうかもしれません。
そのような存亡/壊滅的リスクの中でも特に、AIの能力が高まるにつれてAIを制御できなくなるリスクが懸念の中心になり始めています。

本記事ではなぜAIが人類存亡/壊滅的リスクをもたらすと国際的に考えられつつあるかの論理と提案されている技術的な解決策(AI Alignment)やガバナンス、またそのように考えらえるようになった歴史的背景を説明します。

歴史的にはAIによる存亡リスクへの懸念自体は19世紀後半から存在し、1990年代にNick BostromやEliezer YudkowskyによってTranshumanismの挫折可能性として認識され始めました。2000年代には彼らによりAIによる存亡リスクへの対策が具体的に論じられ始め、2010年代には効果的利他主義運動とも合流し、国際的にも大きな影響力を持つようになり始めています。

一方AI分野自体は2010年代に深層学習によって大きく進展しましたが、特に2020年代から大きくAIの能力が向上し始め、ほとんどの認知タスクにおいて人間をはるかに上回る高度なAI(≒超知能)の実現が予想以上に早い可能性が認識され始めました。

そのような高度なAIの実現が近いかもしれないにもかかわらず、現状人間の意図した目標をAIの目標と整合させるAI Alignment問題の解決の兆しはあまり見えていません。

AI Alignment問題が解決されないまま高度なAIが開発された場合、AIが有能に欠陥のある目標を最適化したり、本来の目標から外れたり、権力や資源を求めたり、シャットダウンに抵抗したり、あえて嘘をついたり猫を被ることで策士的な欺瞞に関与するリスクが懸念されます。

その結果、囲碁でAIが思いもよらない戦略で人類トップに打ち勝ったように、私たちが想像もできないような方法でAIがAI Alignment手法やセキュリティの網を乗り越え、最悪の場合悪意の有無に関係なく、AIの目標追求に既存の人類社会が邪魔になるという道具的な理由のため、結果的に人類に壊滅的な結果がもたらされる可能性が一部の人々により危惧されています。

また、人類へのある種の攻撃という目標をAIが道具的な目標として持たなかった場合でも、多くの高度なAIが複雑に相互作用し進化していく社会のなかで、人間がAIをコントロールすることが徐々に難しくなり最終的には人間が重要な意思決定に全く関与できなくなる可能性もあるでしょう。
その場合、生存に必要なリソースが確実に人類に提供されるかは定かではありません。何らかの工業プロセスが人間の生存可能な範囲を超えて環境を激変した結果、人類が絶滅する可能性もあります。

他にも高度なAIの開発や取得のハードルが技術発展によって下がり、全体主義への使用、戦争や紛争のエスカレート、テロリストによる悪用のリスクも考えられるでしょう。AIが新たなパンデミッ

クを引き起こすバイオテロに使われたり、プロパガンダ、検閲、監視に利用されたり、有害な目標を自律的に追求するためにAIが解放されてしまうリスクがここに当てはまります。
このようなリスクは以前からもありましたが、AIの能力が高まるにつれて、その被害も類を見ない規模になる可能性が高まります。

その一方で、上記リスクを理由にAIの能力向上に関する開発を世界的に止めることも現実的ではなく、安全意識の低い主体が開発するよりも前に、ある程度急いで安全意識の高い側も開発しなくては行けないインセンティブがあるでしょう。それは例えるならば、地雷原をできるだけ早く駆け抜けるゲームを人類がしているようなものかもしれません。

新しいテクノロジーに関する我々社会の典型的な戦略の一つは、それらを導入した後に時間をかけて軌道修正し、問題が発生した後に解決するというものです。

たとえば、現代のシートベルトはT 型フォードの登場から43年後の1951 年まで発明されませんでした。消費者用ガソリンには、段階的に廃止されるまで、数十年にわたって神経毒鉛が含まれていました。
高度なAIに関して言えば、これらのシステムを適切に制御することに比較的早い段階で失敗すると、後の軌道修正ができなくなり大惨事が生じる可能性があります。つまり、人間が自分自身の社会の軌道修正能力を決して失わないように、問題をかなり前に予測してAIのもたらすリスクに技術的/政治的に対処する必要性が国際的に広まりつつあると言えるでしょう。

近い将来にAIによる存亡リスクがあり得るということに全ての専門家が同意しているわけではなく、議論があるということは強調するべきです。

しかし技術的/政治的な問題は解決するまでに数十年かかる場合があり、また結果として起こり得るインパクトも大きいため、たとえ高度なAIが実現するのが数十年先であり、壊滅的な被害が起こる可能性が大きくないとしても、今からAIによる存亡/壊滅的リスク削減のために取り組み始める合理的な理由があると思われます。

今後AIがどこまで進歩するのかは本質的には不明瞭ですが、AIが自律的に人類の未来を奪ってしまう可能性や、その悪用、AIの開発競争や戦争や全体主義での使用といった構造的リスクの危険性を踏まえて今から技術的な解決策やガバナンスに関する議論や準備を始めていく必要があるでしょう。

本稿の目的

2023年はChatGPT,GPT-4を皮切りに AIのもたらす潜在的なメリットが取り上げられました。その一方で AIのもたらす極端なリスク（人類絶滅も含む）が取り上げられた一年でもありました。

具体的に幾つか列挙すると、

- 2023年3月22日にFuture of Life InstituteがGPT-4より強力な AIシステムの学習の6か月の停止を求める公開書簡を提出
- 2023年5月30日にCenter for AI safetyが「AI による絶滅のリスクを軽減することは、パンデミックや核戦争などの他の社会規模のリスクと並んで世界的な優先事項であるべき。」とする声明を発表
- 国連事務総長はAIの存亡リスクを真剣に受け止めなければならないと発言
- 欧州委員会は AIによる絶滅リスクを軽減することは世界的な優先事項と声明
- イギリス首相による人類の絶滅を含む AIのリスクへの言及やアメリカのハリス副大統領のイギリスのAI Safety Summit 前の演説にてAIによる人類存亡の脅威についての言及
- アメリカ大統領令やそれへのバラクオバマ大統領の応答
- 最終的にイギリスのAI safety Summitにて28カ国(アメリカ、中国含む)とEUがAIが重大なリスクをもたらすことに同意しブレッチリー宣言に署名、
- OpenAI解任騒動の原因の一つが高度なAIによる深刻なリスクだった可能性

上記のような2023年にAIのリスクを背景に起こった種々の動きの背景には、AIシステムに私たちが望んでいることを実行させることに焦点を当てた技術的および概念的な研究であるAI Alignmentに関する20年以上にわたる研究の歴史とその深刻性に関する認識が存在し、その歴史的な起源は1990年代のTranshumanism/Extropianism/Singularityに関連する考え方にまで遡ります。

AI Alignmentと聞くとOpen AIが2022年に提出した言語モデルの微調整手法であるRLHF論文を思い浮かべる方もいるかもしれません。

そのようなAIに偏見や公平性に欠ける発言をさせないといった研究目標もAI Alignmentには含まれ得ますが、AI Alignmentに関する研究分野自体は自動運転の安全性も含む広範なAI Safety分野より狭く、ある種人間よりも高度なAIである超知能をどのように安全に取り扱えるのかというモチベーションから生まれた分野と言えます。

今後はAIの能力が高まるにつれて、そのような高度なAIをどのように人間の意図した目標に合わせられるかといった文脈でAI Alignmentという言葉が使われる機会が増えると思われます。

また、生成AIという単語が2023年の流行語となったこともあり、AI脅威論（ 高度なAIが人類を存亡の危機に陥れる可能性に関する議論）の一連の流れをぽっと出の流行、気分と感じる方も多いかもしれません。

しかし、先に言ったようにAI AlignmentやAIを要因とする存亡リスクに関する議論には20年以上の歴史があり、欧米の一部研究者やAIのもたらす深刻なリスクを昔から考察してきた人々にとってもGPT-4の能力向上に驚いたとはいえ、AI脅威論が深刻な議論のテーマになることについては歴史的必然のあるトレンドだと感じていると思われます。

よって今後、AIを要因とする人類に対する壊滅的なリスクや存亡リスクに対処するための超知能を含む高度なAIに関するガバナンス議論が国際的な政治の力学にも広範な影響を及ぼすことになっていくと思われます。
そしてこれからAI Alignment分野はその緊急性と深刻さもあり、急速にメジャーな学術分野として

構築されていくと思われます。
そのため、日本も今後AI AlignmentやAIによる存亡リスクについての議論を深刻に捉え返す必要性が出てくるでしょう。

現状日本語で網羅的にAI脅威論の背景からその論理、対策までを記した文献はほとんどなくまた、今世界で巻き起こっているAIのもたらす極端なリスクに関する議論の歴史的背景とその論理の理解をすることが必要となると思われるため、その参考となることを目的に本記事を執筆しました。

本記事ではまずAIのもたらす存亡/壊滅的なリスクとその緊急性について述べた後、それが起こりうる論理や脅威モデルについて説明し、その後、AIによる存亡リスクに関する歴史的な背景とその技術的/政治的対策を解説します。

AI脅威論の深刻さと緊急性

この章ではAIの発展がもたらす深刻なリスクとなぜそれが昨今緊急性のある問題として取り上げられ始めているのかを説明します。

AI脅威論概観

もともとOpen AIのGPT-4リリース前からAI脅威論（高度なAIが人類を存亡の危機に陥れる可能性に関する議論）は、世界中に広まっていました。例えば19世紀後半に新聞記事で取り上げられ、20世紀にSF小説のテーマを飾っています。

そして2000年代に入ってから、具体的/技術的にAI脅威論に対処する論考や AIが人類存亡リスクに寄与する可能性が議論されはじめました。
次に、深層学習が始まった2014年頃からはもう少し実装ベース、具体的な技術ベースでAI脅威論に対処するAI Alignment分野(AIに人間のやりたいことをやらせることに焦点を当てた研究分野)が生まれてきます。

また、AIガバナンスという意味でも2010年代にはGPT-4以上の学習を一時的にストップする公開書簡を出したFuture of Life Instituteが創設されました。

後述する2011年に名づけられた効果的利他主義運動の一部長期主義的な考え方（長期的な未来にプラスの影響を与えることが現代の重要な道徳的優先事項であるという考え方）にも反映され、具体的で実践的な現実世界での運動や考え方にもつながっています。

実世界においても2014年頃からイーロンマスク氏やスティーブンホーキング氏ら著名人も公にAIによる存亡リスクを懸念し始めました。

そして、AI脅威論がニッチなオタクや一握りの研究者の話題から公の場で本格的に議論され、国際社会の動向にさえ影響を与え始めるようになったのが2023年であるといって良いでしょう。

2023年の3月14日にGPT-4がリリースされFuture of Life Instituteの公開書簡が発表されて以来、AI脅威論に関する話題はアメリカの大手メディアでも取り上げられるほどにメジャーになりました。

そして冒頭にも書いたように、その後、国連事務総長はAIの存亡リスクを真剣に受け止めなければならないと発言し、欧州委員会は AIによる絶滅リスクを軽減することは世界的な優先事項と声明を出し、イギリス首相による人類の絶滅を含む AIのリスクへの言及やアメリカのハリス副大統領のイギリスのAI Safety Summit 前の演説にてAIによる人類存亡の脅威についての言及、アメリカ大統領令やそれへのバラクオバマ大統領の応答、最終的にイギリスのAI safety Summitにて28カ国(アメリカ、中国含む)とEUがAIが重大なリスクをもたらすことに同意し、ブレッチリー宣言に署名、そしてOpenAI解任騒動の原因の一つが高度なAIによる深刻なリスクだった可能性にも繋がっていきます。

AIをめぐる深刻なリスクに関する認識は世界中で共有され、もはや「オタク」の話題ではなくなりつつあると言えるでしょう。

AI進歩の想像以上の早さ

AIの脅威論の概観を見てきましたが、何故今、高度なAIがもたらす深刻なリスクに一部研究者や政策立案者達が緊急性を感じているのでしょうか。

それは想像以上に速いAIの能力向上に多くの人がある種驚き、それと比較して高度なAIがもたらすリスクに対処するための投資も研究者数も規制も十分ではない背景があるためです。

ここでいう高度なAIとはおよそ二つを指します。

- AGI(汎用人工知能)
- Transformative AI(変革的AI:農業革命に匹敵する重要な移行を引き起こすAI)

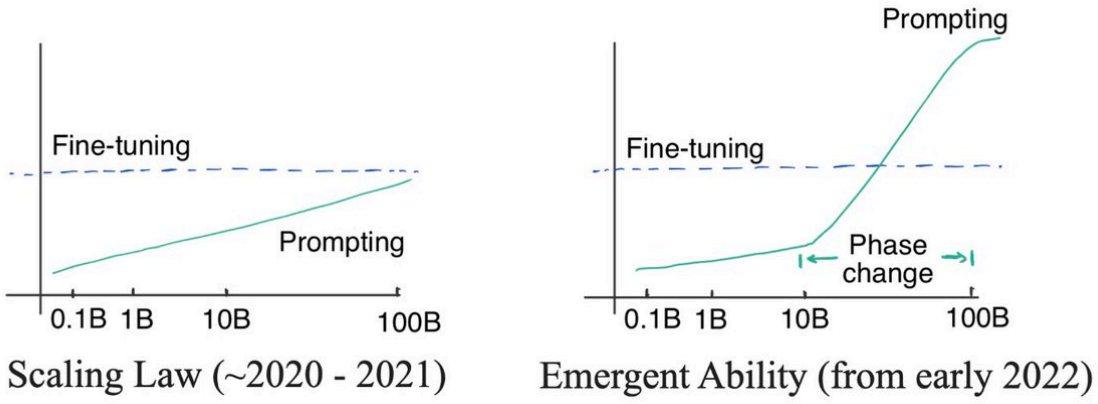
大まかに言えば、人間が実行するほぼすべての認知作業を実行できる (例えば、科学者、CEO、小説家等の代わりにできる) AIシステムを意味します。

OpenAIがGPT-4をリリースする以前から、このような高度なAIのリスクに備えるために、主に効果的利他主義と呼ばれるコミュニティにより、AGI(汎用人工知能)やTransformative AI(変革的AI:農業革命に匹敵する重要な移行を引き起こすAI)がいつ開発される可能性があるかのタイムライン(到達予想年表)が推定されていました。

その結果として、効果的利他主義のキャリアアドバイスを提供するコミュニティ『80000hours』内の記事にて今後数十年間でAIが大幅に進歩し、すべてではないにしても、多くのタスクで機械が人間を上回るようになる可能性が示唆されています。

実際、この10年の深層学習の発展は目覚ましいものがあります。画像認識、囲碁、ゲームなど様々なドメインで飛躍的な進歩を遂げました。
そして、GoogleのPaLMやDALL-E2、またDeepMindの汎用エージェントであるGATOを皮切りに、2022年から多くの人が予想していた以上にAIの能力が急速に向上しました。

特に2022年初頭から大規模言語モデルの能力がプロンプトの工夫で突然伸び始めたのです。
2017年頃からScaling Law（モデルサイズと学習データを増やすと損失が減少する経験則）は知られていましたが、2022年初頭の急激な能力の上昇は研究者を含め多くの人を驚かせました。



A Closer Look at Large Language Models Emergent Abilities

それを受け、オンライン予測プラットホームであるMetaculus上ではAGIのタイムラインが縮んでいきます。当初の予想よりどんどんと発達の速度が早まっているということです。

更に2022年後半のChatGPTリリースと、2023年のGPT-4リリースで2021年頃は2050年代だったAGI中央値予想は現在2031年になっています。

現に、2023年10月の専門家調査では前年度調査の2060年から2047年にHLMI(≒AGI)開発の可能性が50%と13年も前倒しになっています。

また、効果的利他主義コミュニティ最大の財団として知られるOpen Philanthropyのアナリストであるアジェヤ・コトラも2020年頃にTAIの中央値予想タイムラインを2050年頃と推定していましたが、2022年に2040年と大幅にタイムラインをアップデートしました。

同じく深層学習のゴッドファザーと呼べるGeffery Hintonも20-50年という以前のタイムラインを5-20年以内のAGI実現のタイムラインに早め、Open AI CEOのSam Altmanは2031年までにAGIまたはそれを超えた超知能の実現を示唆し、DeepMind創業者で主任AGI科学者のShane Leggも2028年までのAGIを50%と見積もっています。

上記から2012-2021年までの深層学習の躍進も凄まじかった一方で2022年からのAIの各ドメインにおける能力の向上に機械学習研究者や予測市場参加者たちを含め多くの人が驚いていたといえると思われます。
しばしば言われる、昔からAIの進歩はすごかったが、ChatGPTがUIとして優れているため一般にその驚きが伝わったというのは一面では正しい一方で、やや正確さに欠けるかもしれません。

AI Alignment/ガバナンスの遅れ

このようなここ数年の急激なAIの能力向上の一方、AI研究者全体は10万人いる中で、人間の意図したことをAIに実行させるAI Alignmentの研究者はわずか400人程度でAI Alignmentに関する研究が相対的に進んでいない現状があります。
そしてもし高度なAIが人間の意図とは別の目標を追求する場合、人類は半永久的に社会のコントロール権を失う可能性があります。

また、気候変動や人工的なパンデミック、また核戦争でさえ人類の絶滅する可能性は低いように思えますが、AIが制御不能になった場合、最悪の結果として人類の絶滅も考えられるでしょう。

火星でさえ安全ではないかもしれません。

AIによる存亡リスクを防ぐため、AIを人間の意図通りに整合させる「AI Alignment問題」を技術的に解決しようにも、AI Alignmentは未解決の問題であり、その解決は相当難しい可能性も示唆されます。

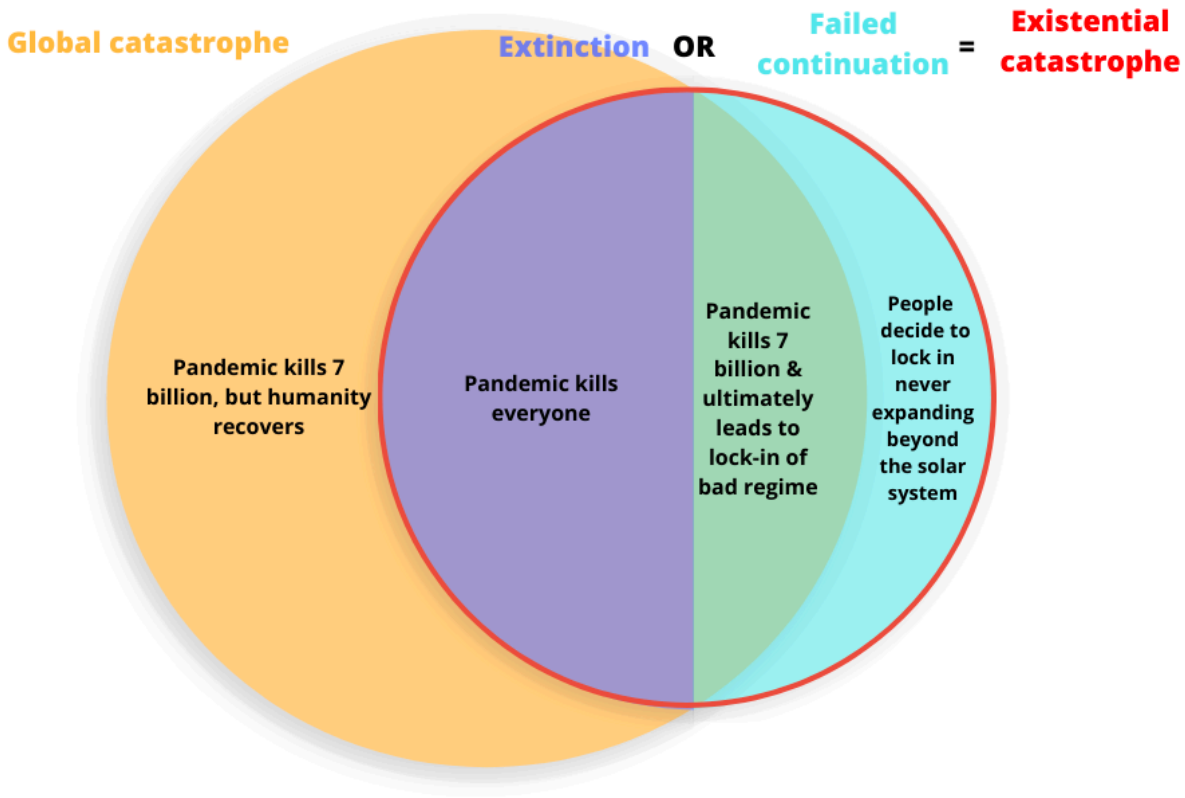
その一方で、高度なAI開発を世界的に止めることも現実的ではなく、安全意識の低い主体が開発するよりも前に、まるで追いかけるように安全意識の高い側も急いで開発しなくてはいけない動機があるでしょう。それは例えるならば、地雷原をできるだけ早く駆け抜けるゲームを人類がしているようなものかもしれません。

開発をストップすることも難しく、AIがもたらしうる存亡リスクに対処するための技術的なAI Alignment研究や世界的なAIガバナンス体制が緊急で必要になっている背景があります。

AIのもたらす存亡リスク(X-risk)

ここで存亡リスク(Existential Risk:X-risk)とは「地球を起源とする知的生命体の早すぎる絶滅や、望ましい将来の発展の可能性を永久的かつ大幅に破壊する脅威のこと（Bostrom 2002）」であり、絶滅以外にも文明が再帰できないレベルに陥るリスクや権威主義的なディストピア社会が訪れる可能性も含まれます。

※余談ですが、地球規模の壊滅的リスクには存亡リスクが一部含まれます。一方地球規模の壊滅的リスクはこれに限定されず、例えば、パンデミックまたは人為的気候変動により人類が絶滅はしないまでも、数億人が死亡するリスクが含まれます。以下のベン図のように存亡リスクと地球規模の壊滅的リスクが区別されていて参考になります。



Venn diagrams of existential, global, and suffering catastrophes

それではAIによる存亡リスクについてはどの程度の可能性が見積もられているのでしょうか？
AIを原因とした存亡リスクは他のよく挙げられる原因（核戦争、人工的なパンデミック、気候変動など）を理由とする存亡リスク以上のリスクとして見積もられる場合もあります。

例えば、哲学者のToby Ord氏の著作「The precipice」では自然要因や他の人為的なリスクよりも高い10%という確率が著者の網羅的な調査の後、主観的に導かれています。

存亡的破局の原因	次の100年間で起きる確率
小惑星や彗星の衝突	~ 1/1,000,000
破局噴火	~ 1/10,000
恒星爆発	~ 1/1,000,000,000
自然リスクの総計	~ 1/10,000
核戦争	~ 1/1,000
気候変動	~ 1/1,000
その他の環境破壊	~ 1/1,000
「自然に」発生するパンデミック	~ 1/10,000
人工パンデミック	~ 1/30
AIの逸脱	~ 1/10
不測の人為的リスク	~ 1/30
他の人為的リスク	~ 1/50
人為的リスクの総計	~ 1/6
存亡リスクの総計	~ 1/6

The Precipice: Existential Risk and the Future of Humanity内の存亡的破局の網羅的な要因と次の100年で起こるToby Ord氏による主観的な推定による確率の表

実際効果的利他主義系のコミュニティにおいてAIによる人類存亡リスクが気候変動、人工的なパンデミック、核戦争と比較して不確実性は高いものの大きいと推定されています。

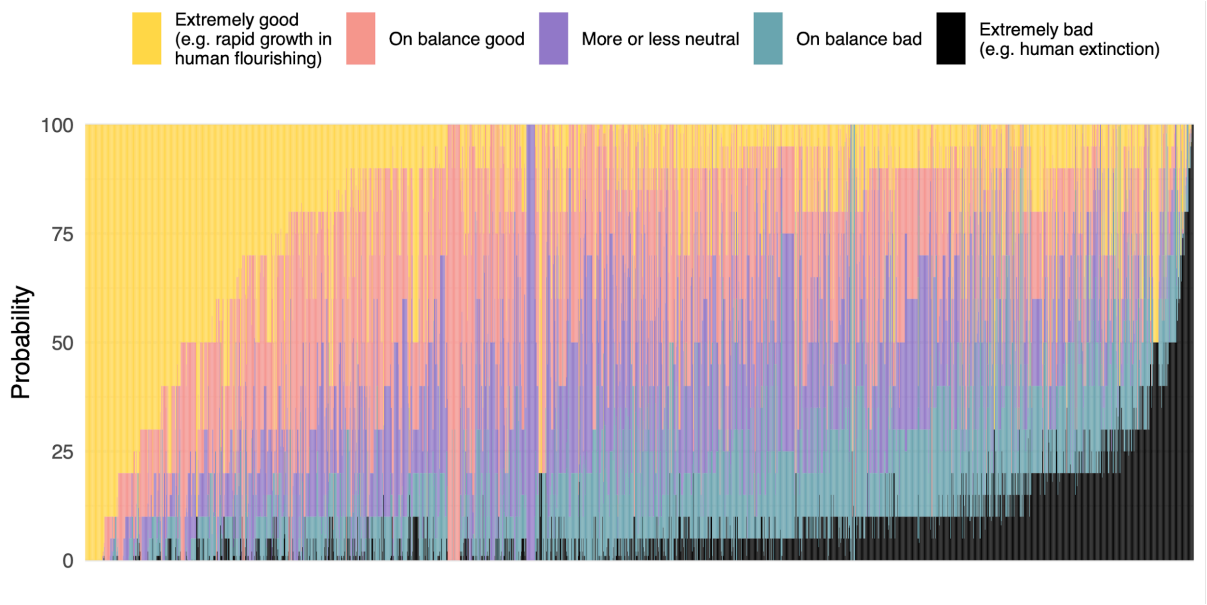
AIによる存亡リスクに次はフォーカスを当てましょう。

AIによる存亡リスクを分析するために、計画を立て長期的にある目標を最適化する高度なAIエージェントを仮定して分析した論文もあり、それによるとおよそ5-10%で人類存亡に至る可能性があると結論づけられています。

また、2023年の専門家アンケート調査にて高度なAIが人類の絶滅に繋がる可能性に関する質問の回答者中央値は5~10%となっています。

一方下記の図のように専門家の殆どはAIのもたらす結果が「最も良い」から「人類の絶滅に相当する悪い結果」までの可能性にさまざまな主観的な重みづけをしていることが見てとれます。

つまり現状専門家の中でもAIによる存亡/壊滅的リスクがどの程度現実的なものなのかについて意見が分かれている現状があるでしょう。



HLMI が人類に及ぼす長期的な影響の肯定的または否定的な800 件の回答を無作為に選択したグラフ。各垂直バーは1人の参加者を表す。

しかしAI Alignment分野の流れを創始したEliezer Yudkowskyは「安全なオペレーティングシステムを作成する可能性は事実上ゼロです。AIアライメントに関して私が見ている状況です。」と発言しており、AI Alignmentの著名な研究者Paul ChristianoはEliezer Yudkowskyほど極端ではないにせよ、20%で殆どの人類が死に、26%でほとんどの人間は生きているが好ましくない未来を推定しています。
また深層学習の父Geoffrey hintonは人類存亡確率を10%と考えているようです。
対照的にMetaの主任AI研究者のYann LeCun氏はAIによる存亡リスクは無視できるほど小さいと考えています。

※他著名人のp(doom)=人類存亡確率の一覧

つまり、人によってはAIによる存亡リスク、X-riskはほとんどあり得ないという意見からほぼ確実に起こるという主張まであり、不確実性が高い状況になっています。

一方で、今後数十年、早ければ10年以内に深刻な被害を社会にもたらし得る高度なAIが開発される可能性があるため、AIガバナンスやAI Alignment分野に今から早期に注力する必要があると国際的に考えられるようになっていると言えるでしょう。

※余談
X-riskよりも悪い結果をもたらすS-risks(Astronomical suffering risks 天文学的苦しみのリスク)とよばれる概念もあります。遠い将来に天文学的な規模で激しい苦しみが生み出されるリスクであり、これまで地球上に存在したすべての苦しみをはるかに超えています。
人類の滅亡よりも悪い結果をもたらすhyperexistential riskとも呼ばれ、Center on Long-Term Risk (CLR) とCenter for Reducing Sufferingという二つの組織がS-risksの予防研究をおこなっています。

AGI/TAI実現推定時期

Nick Bostromが存亡リスクについて定義した2002年の論文の中で以下のように書いています。

> 『存亡リスクに対するアプローチは試行錯誤的なものであってはなりません。エラーから学ぶ機会はありません。何が起こるかを確認し、損害を制限し、経験から学ぶという事後対応型のアプローチは機能しません。むしろ、積極的なアプローチをとらなければなりません。これには、新しいタイプの脅威を予測する先見性と、断固とした予防措置を講じ 、そのような行動のコスト（道徳的および経済的）を負担する意欲が必要です。』

https://nickbostrom.com/existential/risks

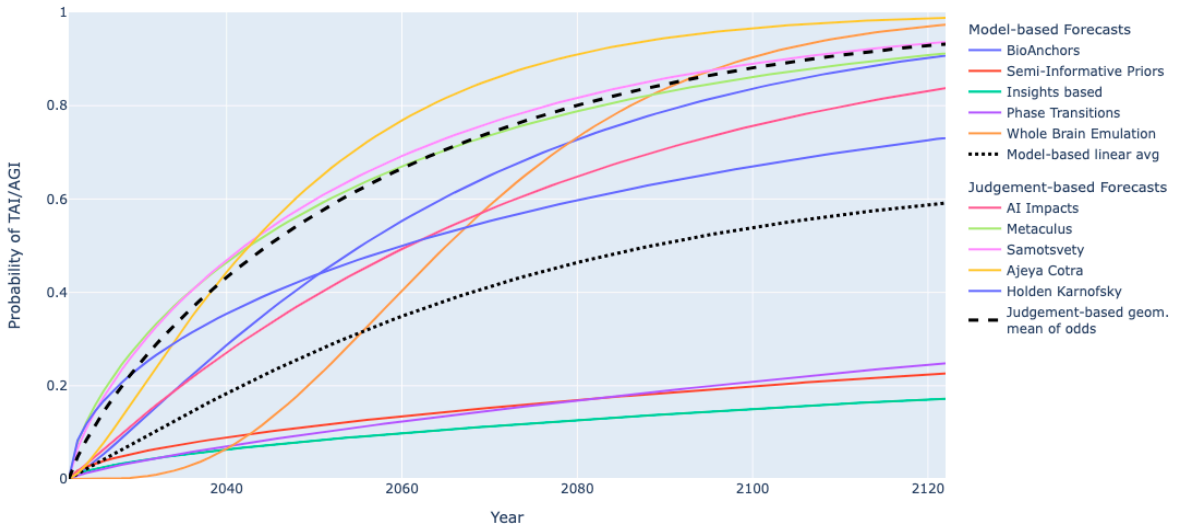
存亡リスクについて考えるときは、脅威を予測する先見性が必要とされます。その文脈で効果的利他主義系のコミュニティはAGI/Transformative AI(TAI)の実現時期を予測しようとしています。それはいつ革新的なテクノロジーが出現しうるかで脅威に対処する優先度合いが変わるためです。

このようなモチベーションから効果的利他主義系のコミュニティではTAI/AGIの実現時期を推定するレポートが数多く出されてきました。

それらレポートをある程度包括的にレビューした記事が2023年にEpoch（機械学習トレンドの調査機関）から出されています。

ここでモデルベースと判断ベースの二つのタイムライン推定手法に分類されます。

- モデルベースとはTAI/AGIの到着日について明確なモデルを指定し、時期を推定するモデルです。
- 判断ベースとは明確なモデルを用いずに、個人または個人のグループによって行われる主観的なタイムライン推測手法のことです。



TAI/AGIのタイムラインを推定した各モデルとそれらモデルが重み付けされたモデルの年代を横軸とする累積分布関数

モデルベースは2089年、判断ベースは2045年までのAGI/TAI実現に50%の確率が割り振られています。

一方2030年までの実現にモデルベースは8%、判断ベースは25%と大きな確率が割り振られており、それら高度なAIやその発展形である超知能（Artificial Super Intelligence）がもたらす壊滅リスクや存亡リスクは現状無視できないと考えられます。

超知能/AI離陸速度(AI takeoff)

AI離陸(AI takeoff)とは、汎用人工知能が一定の能力の閾値 (しばしば「人間レベル」として議論される)から、超知能になり文明の運命を制御するのに十分な能力を発揮するまでのプロセスのことを指します。

ここで超知能（Super Intelligence）をオックスフォード大学教授の哲学者Nick Bostrom氏は、「実質的にすべての分野（科学的創造力・全般的な知識・社会技能を含む）において、その分野でもっとも優れている人間の頭脳よりもはるかに賢い知性」と定義しています。

最近だとアニメ「Beatless」や「Vivy」、洋画ですと「Transcendence」や漫画だと「AIの遺電子」、小説だと円城塔さんの「Self-Reference ENGINE」などで人間や人類をはるかに超える知能を持つ機械知性の姿が描かれており、ある意味でAI Alignemnt問題やAI脅威論にも通じるテーマも描かれています。

超知能なんてSFではないかと思われる方は多いと思われますが、もし汎用人工知能が実現した場合、そこから数十年ではなく、もしかしたら数年、想像以上に早い場合は数分で超知能に発展する可能性がNick Bostromによって考察されています。
(SuperIntelligence 4章 Nick Bostrom)

それは主に人間と比較してデジタルの知能は容易に記憶容量を増やしたり、演算速度を上げることが可能なため、一度人間レベルの知能を作成した場合それをアルゴリズム的に改善しなくても、ハードの拡張とデータ資源の拡張（インターネットの情報を与えるなど）をすれば容易に記憶

容量もしくは計算速度を何桁も伸ばすことができてしまうのではないかと考えられているためです。

また、自律的な自己改善が起こって人類がハードウェアの拡張やデータ資源の拡張やソフトウェアの改善などを外部から施さなくても、再帰的に能力を急速に向上させる知能爆発が起こる可能性も考察されています。
（スーパーインテリジェンス　超絶AIと人類の命運　Nick Bostrom著　第4章知能爆発の速さ　参照 ）

そして、予測プラットフォームのMetaculusでは現状(2023/12/31)AGIが作成されてから超知能に至るまでの50%予想は10ヶ月程度と想定されています。

また他にも示唆的なレポートでは2030年のAIモデル（=GPT2030とする）は、1 年間に動作する 180 万人のエージェントをそれぞれ 2.4 か月でシミュレートできる可能性が指摘されています。

更にFLOPごとに 5 倍の料金を支払えば、25 倍の高速化 (人間の速度で 125 倍) が得られ、結果として1 年間に働く 14,000 人のエージェントを3 日間でシミュレートできる可能性があると推測されています。

これは世界中に数学者の数はそれほど多くないため（たとえば、米国ではわずか 3,000 人）、GPT2030では、すべての数学者の年間生産量を数日ごとにシミュレートできる可能性があることを示唆しているようです。

つまり、この考察から知能の質的な改善がなかったとしても、Nick Bostromの言う集合知スーパーインテリジェンス(Collective SuperIntelligence)もしくは高速スーパーインテリジェンス(Speed SuperIntelligence)が早い時期に来る可能性を示唆します。

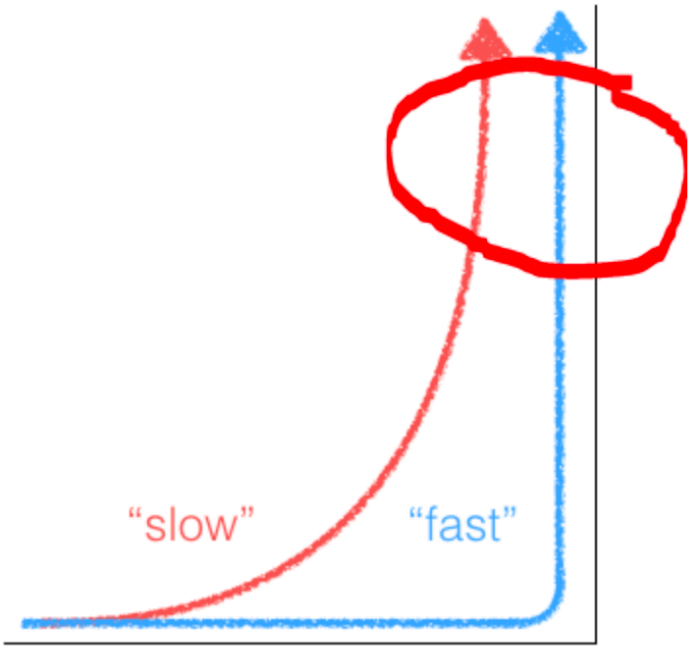
ここで前者は多数のより小さな知性で構成されるシステムが現在の認知システム（国家や会社）を大幅に上回るパフォーマンスを持つシステムであり、後者は典型的な人間の思考ができることを、はるかに高速に実行できるシステムです。

よって上記議論から、離陸速度は数年となる可能性があり、質的な知能爆発が起こらなかったとしても集合知スーパーインテリジェンスを予想する合理的な理由があります。

つまるところ、AGI/TAIの実現時期を早いかもしれないと考え得るならば、同様に超知能の実現時期も早い可能性を想定して深刻なリスク（壊滅的なリスクや存亡リスク）を懸念する緊急性があると考えられます。

※他参考
AIの能力が人間レベルの認知タスク実行能力に近づきそれを超えようとする際、AIの能力はどれくらいの速さで向上するかという問題をバイオアンカー仮説と呼ばれる人間の脳の機能を参照し計算量を算出するための仮説を用いて推定したレポートが存在しています。結論として20%程度の認知タスクを自動化するAIが出てきた時点で相当速い進歩の加速が示唆されています。

また、そもそも上記は連続的なAIの能力向上を仮定していますが、非連続的な能力の向上が今後起こらないとは限りません。



Will AI undergo discontinuous progress?

AIの能力の向上の速さ遅さではここでは問題にされていません。連続的な進化でもとても速い能力向上の実現は想定されえます。
「連続性と非連続性」に関する議論は以下の記事がおすすめです。

Will AI undergo discontinuous progress? — LessWrong *This post grew out of conversations with several people, incl www.lesswrong.com*
Discontinuous progress in history: an update *Katja Grace April 2020 We've been looking for historic cases aiimpacts.org*

最も重要な世紀

私たちが今いる21世紀は過去人類が体験したことのないような技術の発展の最中におり、もしかすると近いうちに汎用人工知能が開発され、劇的に世界が変わってしまうかもしれません。

その一方で今世紀に人類が何らかの要因で存亡的破局を迎えてしまった場合は長期的な未来の可能性が失われることになります。

しかし今世紀を乗り越えれば人類の繁栄は安定化する可能性もあります。そういう意味で今世紀は人類にとって最も重要な世紀と言えるかもしれません。

残念なことに現状超知能の実現が予想以上に早いかもしれないにもかかわらず最も存亡リスク要因としてあり得るとされるAIアライメント問題の解決の兆しはほとんど見えません。

AIの開発を世界的に止めることも現実的ではなく、安全意識の低い主体が開発するよりも前に、ある程度急いで安全意識の高い側も開発しなくてはいけないインセンティブがあるでしょう。
それは例えるならば、地雷原をできるだけ早く駆け抜けるゲームを人類がしているようなものかもしれません。

新しいテクノロジーに関する我々社会の典型的な戦略は、潜在的なすべての重大な問題に取り組む前にそれらを導入し、時間をかけて軌道修正し、問題が発生した後に解決するというものです。たとえば、現代のシートベルトは、T 型フォードの登場から 43 年後の1951 年まで発明されませんでした。消費者用ガソリンには、段階的に廃止されるまで、数十年にわたって神経毒鉛が含まれていました。

一方、高度なAIは、これらのシステムを適切に制御することに比較的早い段階で失敗すると、後の軌道修正ができなくなり、大惨事が生じる可能性があります。つまり、人間社会の軌道修正能力が決して消滅しないように、問題をかなり前に予測してAIのもたらすリスクに対処する必要があるという認識が国際的に広まりつつあると言えるでしょう。

AIによる壊滅的リスク分類

次にAIによる深刻なリスクはどのような原因で起こると考えられているのかを解説していきたいと思います。

少し前提知識を説明する都合上長くなるため、AI脅威論の骨格を早く知りたい方は「AI Doomerの論理」の節や「具体的な脅威シナリオ」章から読んでいただいても構いません。

AI Safetyに関する研究を行うCenter For AI SafetyはAIによる壊滅的なリスクを以下の4つに分類しています。
ここでは「悪意のある利用」、「 AI競争」、「組織のリスク」について説明し、AI Alignment問題の本質的なリスクに近い「不正なAI」(Rogue AIs)については次の章で解説します。

Figure 2: In this paper we cover four categories of AI risks and discuss how to mitigate them.

An Overview of Catastrophic AI Risks

悪意のある利用

これは、人々が意図的に強力なAIを利用し、広範囲に害を及ぼす可能性です。AIが新たなパンデミックを引き起こすバイオテロに使われたり、プロパガンダ、検閲、監視に利用されたり、有害な目標を自律的に追求するためにAIが解放されるリスクがここに当てはまります。

現状偏見や誤解を広める問題に焦点が当てられていますが、今後AIの能力が高まると悪用されるリスクが大きくなるでしょう。

現にAI Safetyに関する研究機関のAnthropicの予測によれば「2〜3年後にAIシステムは科学や

工学の分野ではるかに優れた能力を持つようになり、特に生物学の分野では、大規模な破壊を引き起こすために悪用される可能性があるという。このような科学・工学スキルの急速な成長は、国家間のパワーバランスも変える可能性がある。」とホワイトハウスで2023年7月23日にDario Amodei CEOが証言しています。

特にAIを用いたバイオテロについては今後2~3年以内に現実化する可能性がAnthropicにより指摘されています。

効果的利他主義コミュニティでも最も差し迫った問題の一つとして人工的に引き起こされたパンデミックが挙げられています。

また、AI は、個々のユーザーに合わせて議論を調整することで大規模な偽情報キャンペーンを促進し、潜在的に国民の信念を形成し、社会を不安定にする可能性があります。人々はすでにチャットボットとの関係を築いているため、強力な攻撃者が「友人」とみなされるこれらの AI を影響力のために利用する可能性もあるでしょう。

AI競争

経済競争によって国家や企業がAIの開発を急ぎ、これらのシステムへのコントロールが失われたり、自律型兵器やAIを活用したサイバー戦争によって、紛争や戦争が制御不能に陥るリスクです。また、企業は人間の労働力を自動化するインセンティブに直面し、大量失業とAIシステムへの依存リスクもここに当てはまります。

この経済競争シナリオで有名なAIの脅威モデル(特定のリスクがどのように展開するかについてのストーリー)をAI研究者のPaul Christianoが描いたものが以下にあります。
一体のAGI（シングルトン超知能）が人類を殲滅するという古典的なAI脅威論におけるシナリオではなく、社会に張り巡らされた複数のAIシステムの複雑な相互作用がミスアライメントすることによる人類への脅威シナリオが描かれており、より現実的なシナリオとなっています。

What failure looks like — LessWrong *The stereotyped image of AI catastrophe is a powerful, malici www.lesswrong.com*

また紛争や戦争が制御不能に陥る可能性に関して、Future of Life InstituteがAIによる自動化された軍事システムを用いた核戦争へのエスカレーションシナリオを説得力を持った映像作品として公開しているためこのリスクをイメージするのに参照すると良いと思われます。

Artificial Escalation - Future of Life Institute *Our new fictional film depicts a world where artificial intel futureoflife.org*

また、少しSF的ですが、AI システムが急増するにつれて、進化のダイナミクスはAI システムの制御がより困難になる可能性に関するリスクもここに分類されます。人間や今までの生命の進化プロセスよりもAIシステムのプロセスは相当速くなる可能性があり、
「変更はハードウェアが許す限り速く、変更は1時間に数百回から数千回まで高速化される可能性がある。オリジナルのプログラマーの制約の少ないシステムが、最も速く改善し、最も意図した性質から遠ざかるようにドリフトする。オリジナルの人間の設計の意図はすぐに無関係になる。」とCenter For AI Safety のDan Hendricksにより懸念されています。

組織のリスク

先進的なAIを開発する組織は、特に安全性よりも利益を優先する場合、大惨事を引き起こすリスクがあります。
その結果として、AIの誤った一般への流出、悪意のある行為者による奪取、組織のAI Safetyへの適切な投資を怠る可能性があり、それらに関するリスク分類です。

汎用人工知能に関する情報が流出する懸念はとても大きなものになります。他の国家や組織主体が軍事的な優位性を築くことを可能にするため、情報セキュリティの面で並外れた取り組みが求められるでしょう。

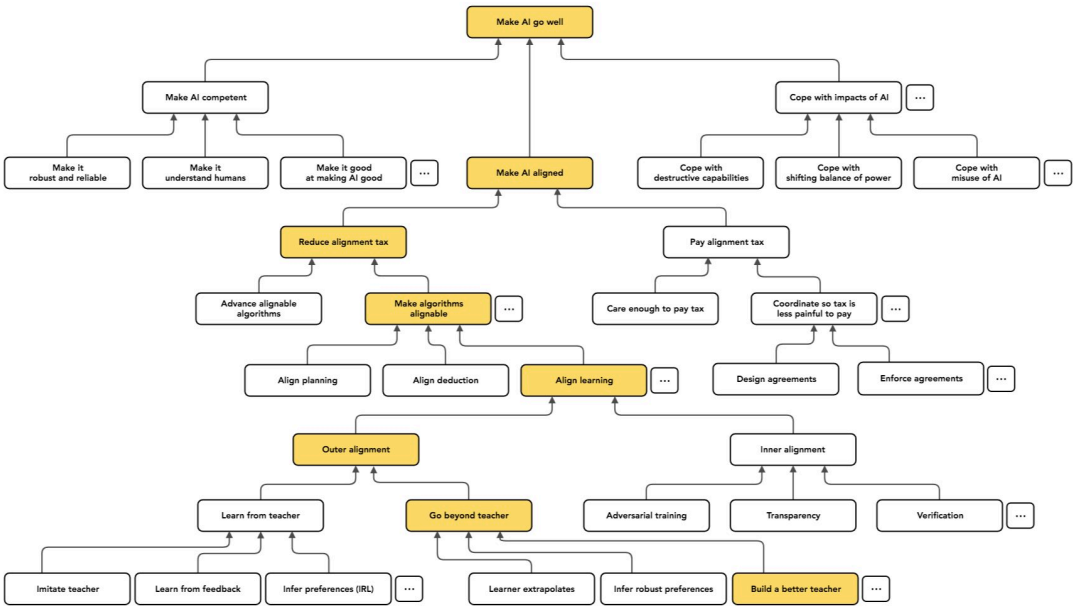
また、社会全体としてのAIの運用の仕方にも規制が必要です。これに関してはOpen AIがエージェントAIシステムが社会に組み込まれる際の初期のプラクティスを提案しています。

不正なAI(Rogue/Misaligned AI)

この不正なAIに分類されるリスクがある意味最も深刻なものです。
これは2000年代から懸念されてきたAIの目標を人間の意図と整合させる、AI Alignment問題の本質に近いリスクとなっています。

これは、AIの能力が高まるにつれてAIの制御が効かなくなる懸念を反映しています。ここではAIの能力が高くないために起こる事故のリスクはあまり念頭にされていません。
AIが有能であるがゆえに欠陥のある目標を最適化したり、本来の目標から外れたり、権力やリソースを求めたり、シャットダウンに抵抗したり、欺瞞に関与する可能性に関するリスクが懸念されています。

以下のPaul Christianoの「AIに良いことをさせる分野の見取り図」では「Make AI competent(AIの能力を向上させること)」と「Make AI aligned(AIアライメント問題の解決)」が区別されています。



AI Alignmentとその周辺の見取り図

AI Alignment問題とはAIシステムが、意図しない望ましくない目標ではなく、人間の価値観や関心に合った目標を追求するようにするという課題です。

つまり、AI Alignment問題で懸念されているのは、「**AI**の能力が低いから起こる事故のリスクではなく、**AI**の目標がわれわれの目標とずれていることで起こる事故のリスク」となります。

AI Alignment問題を解決しなければ、もしかしたらAIはある複雑な問題を解決する能力が高い一方で、その能力を使い、人間を欺こうとする可能性があります。

自動運転車の事故のように能力が低かったり人間の意図したことを理解できないため起こる事故ではなく、AIの目標と人間の意図した目標が一致しないために人間社会に悪影響を及ぼす可能性があり、この問題に対処する必要性があるのです。

アライメントされていないAIはMisaligned AIとよく呼称されており、Rogue AIというのはもう少し分かりやすくmisaligned AIのことを危険で予想外の行動をする欺瞞的で信頼性の低いAIとして直接的に表現したものだと思われます。

上記のRogue AI≒Misaligned AIに関するリスクに関しては詳しくさらに取り上げたいため次の章で解説します。

Misaligned AI

Misaligned AIはその名の通りアライメントに失敗したAIということなので、アライメント問題の本質的なリスクがここに当てはまります。
Rogue AIsの「Rogue」という言葉からも連想できる通りに、人間の意図していない目標を最適化するために(Specification Gaming,Goal Misgeneralization)、お金や電力や資源を求め(道具的収束)、人間に欺瞞的な態度（嘘をついたり、自身の目標を人類に勝てると確信するまで隠し通す態度）を取る欺瞞的アライメント(Deceptive Alignment)の可能性が指摘されています。

本章では上記4つの概念について説明していきたいと思います。

まずは人間の意図していない目標を最適化する経路には二つの類型Specification GamingとGoal Misgeneralizationがあり、それらついて説明します。

Specification Gaming(単純な点数主義)

Specification Gamingとは意図した結果を達成することなく、目的の文字通りのスペック(Specification)を満たす行動のことです。
これは分かりやすく言えば、ある種の優等生タイプのAIのミスアライメントの類型です。融通の利かない優等生がテストの点だけを上げることを目標にして、それが社会に貢献するという意味があることを忘れているかのようなものでしょう。

また比喩的に言えば、ギリシャ神話のミダス王の物語も当てはまるでしょう。ミダス王は「私の手に触れるものすべてが煌めく黄金になるようにしてほしい」と神に頼みますが、食事をしようとしても食べ物が金になってしまい飢えてしまいます。

※道具的収束論と直交仮説の節でも触れますが、Specification GamingをするAIが我々の意図を理解していないとは必ずしも言えません。我々の意図を理解した上でなおスペックを求め続ける振る舞いは可能でしょう。

もう少し現実のAIに落とし込んでSpecification Gaming の例をDeepMindが集めています。

Specification gaming: the flip side of AI ingenuity *Specification gaming is a behaviour that satisfies the litera deepmind.google*

例えば、以下のCoastRunnersゲームの失敗を見ていきましょう。
ゲームの目的は、他のプレイヤーよりも早くゴールすることです。
一方、このCoastRunnersではコース上のプレイヤーの進歩に直接報酬を与えません。代わりに、プレイヤーはルートに沿って配置されたターゲット(緑色のブロック)を攻撃することでより高いスコアを獲得します。

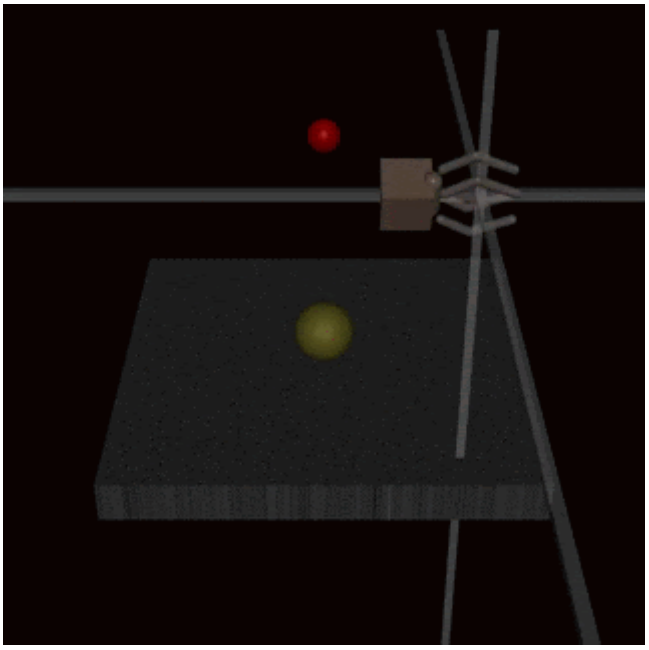
Faulty Reward Functions in the Wild (Amodei & Clark, 2016)

プレイヤーが獲得したスコアがレースを完走するという非公式の目標を反映していると想定されていましたが、RLエージェントは、大きな円を描いて旋回し、3 つの緑色のブロックを繰り返しノックダウンできる水域を見つけます。

火災が発生したり、他のボートに衝突したり、コースを逆走したりしたにもかかわらず、エージェントはこの戦略を使用して、通常の方法でコースを完了する場合よりも高いスコアを達成することができました。

上記は報酬の設計を間違えたことによりおこりましたが、それでは人間が常に見張って報酬を付与する戦略は取れないでしょうか？それでも問題が起こり得ます。

例えば、以下はボールを手で把持するタスクを訓練している様子です。しかし、この場合把持タスクを実行するエージェントは、カメラと物体の間にホバリングすることで人間の評価者をだます方法を学習してしまいます。



Deep Reinforcement Learning From Human Preferences (Christiano et al, 2017)

このように人間の評価者を用いて報酬を与えようとすると、人間の理解できる範囲でわかりやすい結果を見せるおべっか使いをするAIが誕生してしまうかもしれません。つまり、長期的な影響がどうなろうと、評価者を短期的に喜ばせたり、指示された要件を満たしたりするような学習が行われる可能性があります。

Specification Gamingを回避する代表的な手段はRLHF(人間フィードバックによる強化学習)やInverse Reinforcement Learning(逆強化学習)です。

一方で、上記のように人間が教師信号をAIに与えることができるか定かではありません。

またそもそも人間より賢くなったAIの振る舞いを監視できるのか不明瞭です。そのため、Open AIやPaul Christiano氏のAI Alignmentの研究の方向性であるScalable Oversightという考え方が模索されています。

これは一人の人間では評価するには複雑すぎるタスクをAIに助けてもらい高度なAIを人間が監視できるようにするという手法の一群です。

また、Open AIは SuperAlignmentというプロジェクトを推進しており、そこでも超人的な能力を持つAIをどのようにアライメントするか？という文脈で人間より賢いAIに対してもスケーラブルにアライメントを保証できるように以下の論文を最近出しています。

Weak-to-strong generalization *We present a new research direction for superalignment. toget openai.com*

※余談
Specification Gamingと似た概念としてセンサーの入力を改竄(reward function input tempering)したり、報酬関数そのものを変更(reward function tempering)したり、報酬の値そのものを最大化する(wireheading)ような意図しないハッキング手法が整理されている記事があるので紹介しておきます。

Clarifying wireheading terminology — AI Alignment Forum *See also: Towards deconfusing wireheading and reward maximiza www.alignmentforum.org*

Goal Misgeneralization(目標の誤汎化)

Specification Gamingがいわば融通の利かない優等生タイプのミスアライメントに関する類型でした。

さて、こちらのGoal Misgeneralizationとは日本語にすると「目標の誤った汎化」となり、人間が与えた目標とは別の目標を追求してしまうミスアライメントの類型となります。システムの能力は汎化されるが、その目標は望んだように汎化されません。

これはいわば隠れた目標を持った有能で悪質な詐欺師を思い浮かべると分かりやすいでしょう。

以下のEA Japanによる説明もわかりやすいと思われます。この中で言うところの「おべっか使い」がSpecification Gamingに該当し、「策士」がGoal Misgeneralizationに該当します。

モダンな深層学習でAIアライメントが困難になるかもしれないわけ — EA Forum *This is a Japanese translation of “Why AI alignment could be forum.effectivealtruism.org*

以下GoogleがGoal Misgeneralizationについて具体例を集めています。

Goal Misgeneralization *CoinRun CoinRun is a simple 2-D video game (platformer) wher sites.google.com*

その中の一つとして、以下のCoinRun があげられています。敵や障害物を避けながらコインを集めるシンプルな2D ビデオゲームです。

上記動画を見れば分かる通りに、敵や障害物を有能に避けているのにも関わらず、目標はゴールの右端に到達することと誤って汎化してしまっています。そのため、コインという人間の設計した報酬を無視した行動を有能に追求してしまいます。

Specification Gamingが人間の与えた報酬モデルが間違っていたことで起こったのに対して、Goal Misgeneralizationは上記の例のように例え報酬モデルが正しかったとしても起こり得るミスアライメントの類型となっており、Specification Gamingよりもより困難なアライメント問題の類型となっていると考えられます。

つまり、例え人間が完璧な報酬関数を設計できたとしても、その報酬関数を単に無視し、別の目標を最適化してしまう懸念があり、その場合はSpecification Gamingの対応策がうまく機能しなくなる可能性があります。

この問題に対する解決策は完全には見つかっていませんが、より多様なトレーニングデータを与えること、複数のAIモデルをアンサンブルして異なる結果の場合は警戒する、帰納的バイアスと一般化の理解、機械論的解釈可能性を進めるなどの研究の方向性が考えられています。

道具的収束論と直交仮説

●道具的収束論
今まではAIが追求する目標をコントロールできない可能性を考えてきましたが、例えエージェントが意図しない目標を達成するために動いてもそれを防げたり大きな被害にならなければ問題がないかもしれません。

しかし、十分に知的なエージェントのほとんどが、自己保存や資源獲得などの潜在的に制約のない道具的目標(最終目標に利するサブ目標)を追求するという仮説がSteve Omohundro によって2008年に唱えられ、その後Nick Bostromによって2012年に道具的収束論として理解されました。

Instrumental Convergence - LessWrong *Instrumental convergence or convergent instrumental values is www.lesswrong.com*

例えばStuart Russel氏は印象的な「死んだらコーヒーを汲めない」という具体例を出しています。あるAIエージェントにコーヒーを持ってくることを依頼したとして、これは 比較的無害に見える目標ですが、そのエージェントは自分が存在しなければ依頼主がコーヒーを手に入れることができないことに気づくかもしれません。

したがって、この単純な目標を達成するために、自己保存は道具的に合理的であることが判明します。また、その結果自分がシャットダウンされることにも抵抗するかもしれません。そして、コーヒーを取得するために必要になる権力や資源の獲得も道具的な目標になり得ます。例えばコーヒーを防衛するために大きなビルを建てるかもしれません。その際に邪魔になる人間は排除されてしまう可能性もあります。

ここで重要なのは権力や資源を獲得しようとするのはそれ自体が目的なのではなく、ここで言えばコーヒーを汲むための目的に合致する「道具的な目標」であるという点です。
そのためAIが悪意を持って人間に害をなそうとしたり、お金や資源そのものを目的として行動を起こすことは念頭には置かれていません。

つまり、極端なことを言えば、ペーパークリップを生産するだけのAIが道具的収束を起こし、リソースのある限りそれらをペーパークリップの生産に使用し、人類存亡リスクに繋がるという可能性も否定できません。

実際に環境にある対称性を仮定するとRLエージェントは道具的目標を追求することが最適なポリシーになることが形式的に確認されています。
また、近年大規模言語モデルはそのスケーリングに従って道具的収束目標を追求するような傾向も確認されており懸念が増しています。

●直交仮説
一方でAIは人間の価値観を理解するのではないのでしょうか？そもそもAIを人間の価値観に沿わせることはそこまで難しいのかという疑問も立ち現れるでしょう。

ここで直交仮説と呼ばれる概念が上記道具的収束論を紹介した2012年のNick Bostromの論文で導入されています。
直交仮説は「知性と最終目標は直交する軸であり、それに沿ってエージェントが自由に変化する

ことができる。言い換えれば、多かれ少なかれ、どのようなレベルの知性も、原理的には多かれ少なかれ、どのような最終目標とも組み合わせることができる。」と定義されています。

一見すると知能が高くなればなるほど、仏のような悟りを開いたり、多くの生命に慈悲深くなることを想像してしまいがちですが、論理的には知能と目標は独立しており、人間からしたら荒唐無稽な目標も持ち得て、それでいて知能はとてつもなく高いということが想定し得るということになります。

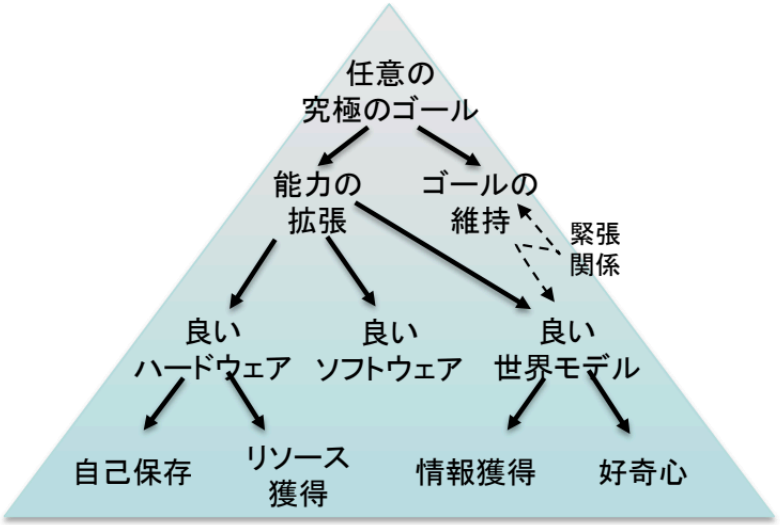
よくAIによる存亡リスクに関して疑問に挙げられる点として、「超知能ならば人間の指示を理解する際に間違えて愚かなことを実行しないほど賢いのではないのか？」があります。
これは価値(目標)と能力の間には論理的には関係はないという上記の直交仮説を理解すれば、「超知能は人間の指示とその意図を場合によっては人間以上に理解した上で、それでも別の目標を持つ可能性がある」ということがわかると思われます。

例えていうならば、ある人が「ホモサピエンスが人工的な味の食べ物を好むのは、栄養価の高い食べ物を求める進化的な圧力によるものだ」と知ったとしても、その人が突然栄養価の高い食べ物を望むようになるわけではありません。

「分かっていることとその通り行動するか」は必ずしも一致しないということです。

直交仮説で主に懸念される問題も、高度なAIが私たちの本当の望みを「理解できないこと」を意味しているのではなく、AIシステムが必ずしも私たちの望みに沿って行動するとは限らないことを意味しています。

上記直交仮説や道具的収束論を合わせると、下記のMax TegmarkのLife3.0の図のように、ほとんどどのような最終目標でもAIは持ちえて、その最終目標を最適化するために、自己保存やリソースの会得を目標にする可能性が指摘されています。



（M. Tegmark, Life 3.0, 2017）

Life 3.0 Max Tegmark著 第7章,図は山川宏氏の資料から引用

また、Eliezer Yudkowskyの以下の言葉は有名です。

> The AI doesn't hate you, neither does it love you, and you're made of atoms that it can use for something else.
> AIはあなたを憎んでも愛してもいませんが、あなたは他の何かのために使える原子でできています。(Eliezer Yudkowsky)

これは直交仮説と道具的収束論を理解すれば、AIは人間を愛しているのでもなく、憎んでいるわけでもなく、単にある目標を最適化するために道具的な収束が起こった結果、人類が絶滅する可能性を示唆する格言でしょう。

※一方で直交仮説や道具的収束論が実際のニューラルネットワークで実現されるかは不透明な部分があり議論があります。

Corrigibility/欺瞞的アライメント

Specification GamingやGoal Misgeneralization、論理的には直交仮説で目標設定をする難しさ、またミスアライメントされたAIが道具的収束により壊滅的な被害をもたらす可能性を論じてきました。

その場合、サーバーに隔離して怪しい振る舞いがあったら物理的もしくは情報的に外に出さないか、シャットダウンしてしまうという戦略が思いつくかもしれません。またはそのAIの目標を再プログラムすれば良いと考えるかもしれません。
このように人間のシャットダウンする試みや再プログラムしようとする行動を妨げないで受け入れる性質を持つAIをCorrigibility(修正可能性)を持つAIまたはCorrigibleなAIと言います。

以下はAIのCorrigibilityに関する問題を解説しておりわかりやすい動画となっています。

しかしここで問題になるのが欺瞞的アライメントです。

欺瞞的アライメントとは実際にはアライメントされていないAIが人間を欺くために一時的にアライメントされたように振る舞うことです。シャットダウンや再トレーニングによる自身の目標の最適化ができないことを避けるために、一時的に人間にアライメントしているように見せるというサブ目標を道具的に追求します。

永遠に物理的情報的にAIを外部に出さないという選択をするならばそれでも良いかもしれませんが、AIを構築するメリットが全くなくなってしまう可能性があります。そのため、この欺瞞的アライメントを解決する必要性があります。

これまでは、概念的にニューラルネットワークの学習における欺瞞的な振る舞いは考察されておりその可能性が論理的には高いとされていました。
一方で大規模言語モデル(LLM)における欺瞞的行動の概念実証例が近年展開され、実際に起こり得る可能性も示唆され始めています。

※余談
他欺瞞的なアライメントだけではないミスアライメントの方向性として、この世界がシミュレーションであるという認識論的な世界モデルを仮定し、その仮定が奇妙な振る舞いを誘引することで人類

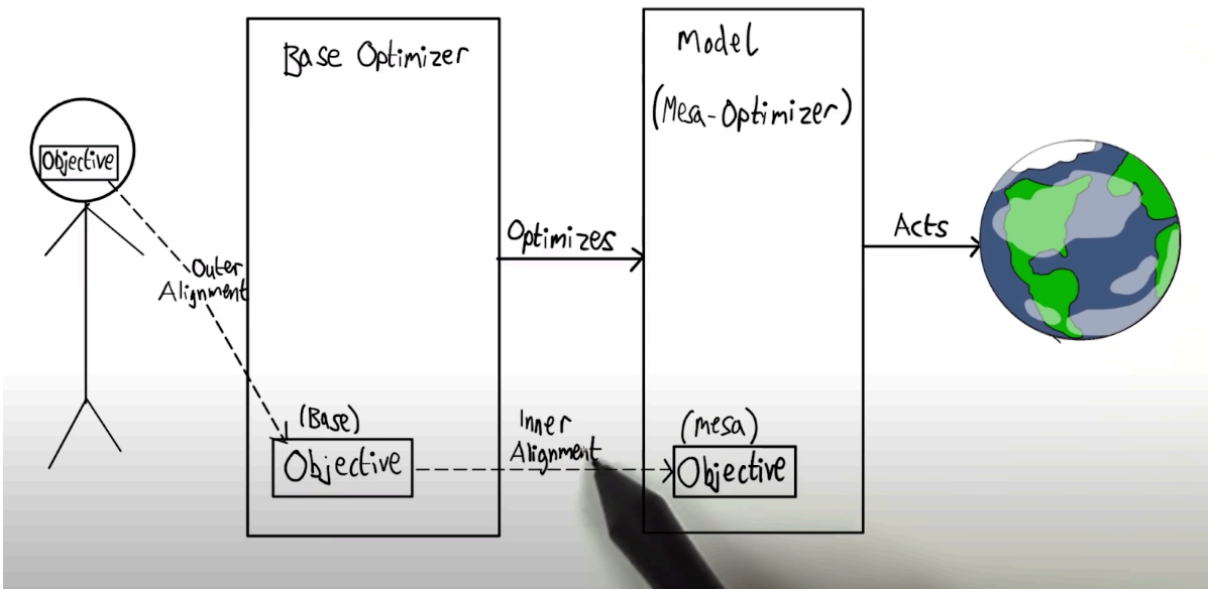
存亡リスクに繋がるシナリオもAI Alignmentの著名な研究者であるPaul Christianoは懸念しています。

Mesa Optimizer,Inner/OuterAlignment

上記の欺瞞的アライメントをMesa Optimizer、Inner/Outer Alignmentと呼ばれる概念を交えながらもう少し詳しく説明していきます。

下記の動画は欺瞞的アライメントを紹介した論文をわかりやすく説明した動画となっています。

以下の画像1において、人間が設計するBase Optimizer(ニューラルネットワークの学習アルゴリズム)とそれが目標とするBase Objective（人間が目的とする目標）を考えましょう。



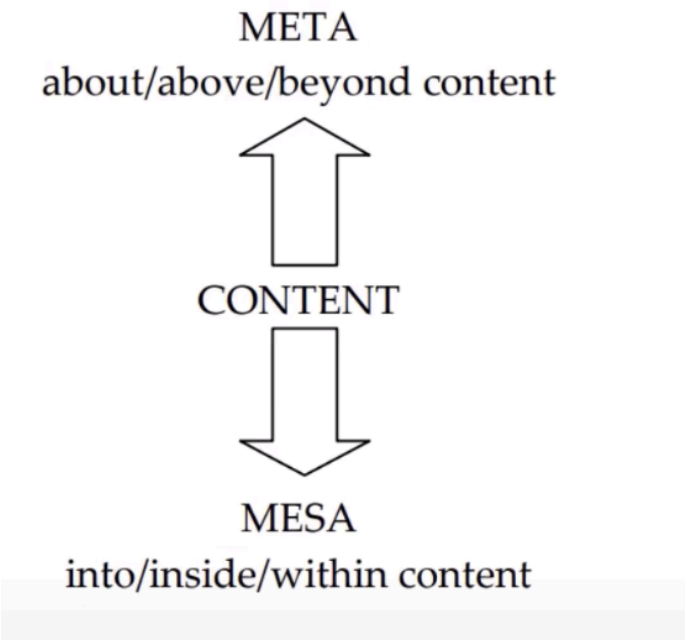
画像1

Base Objectiveと人間の目標が一致することをOuter(外側の) Alignmentと呼びます。これは大まかにSpecification Gamingの解決を意味するでしょう。

またBase Optimizerが最適化を行うModel自身も何らかの最適化を行なっているとみなすことができ、そのModel自身をMesa-Optimizerと呼びます。
そしてMesa-Optimizerの目標をMesa Objectiveと呼びます。

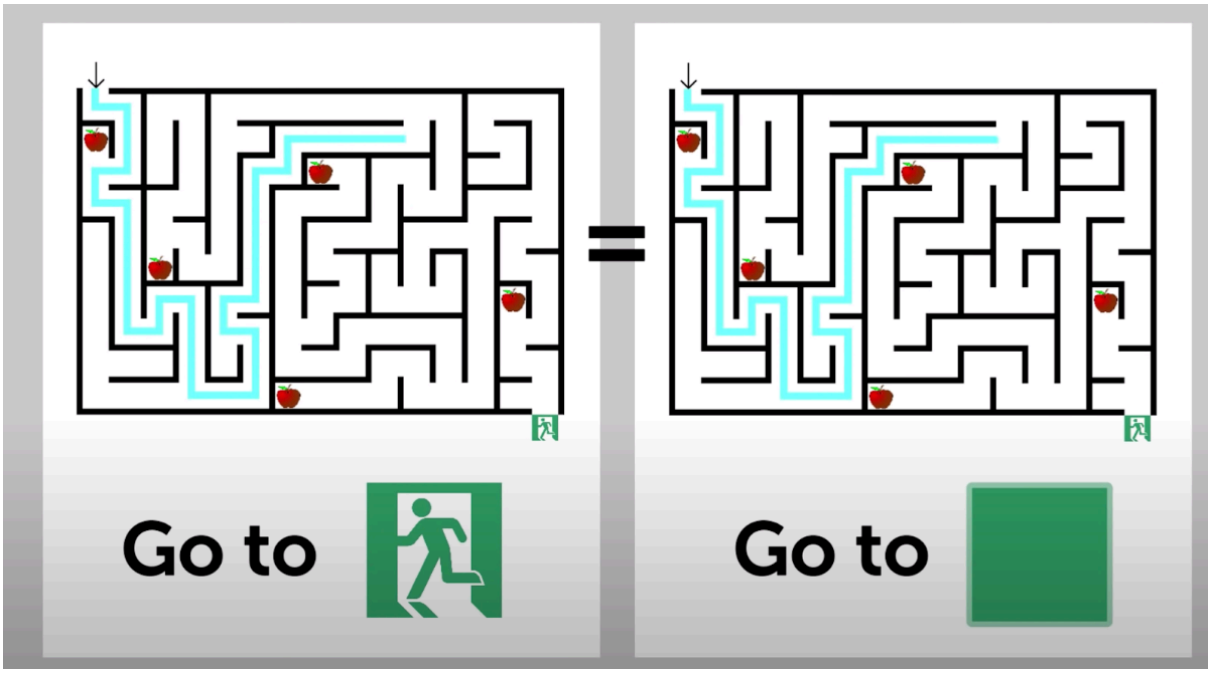
Base Objectiveと Mesa Objectiveが一致していることをInner(内側の) Alignmentと呼びます。これは大まかにGoal Misgeneralizationが解決されている状態です。

画像2にあるようにMesaはギリシャ語でwithinを意味し、モデルの内側でMesa Objectiveを最適化します。Meta最適化はさまざまなタスクを最適化できるbase objectiveを生成可能という意味に近いですが、Mesa最適化は内部の最終目標を最適化するために、道具的に種々のタスクに最適化するという点が違います。



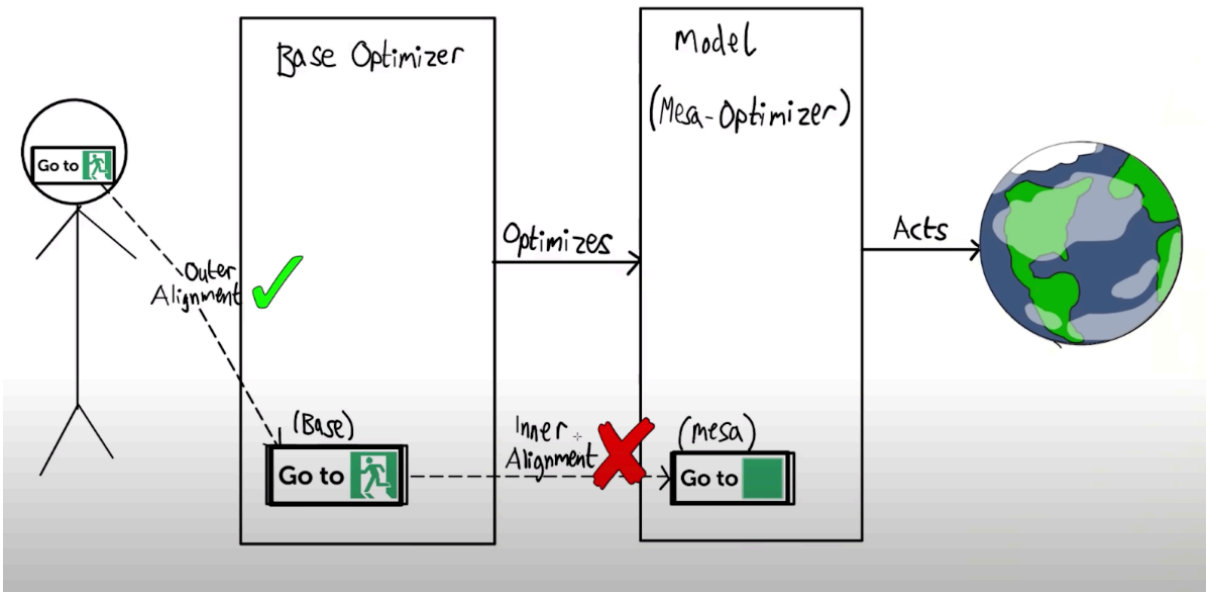
画像2

ここで画像3のように迷路を解く問題で非常口マークを目指すようにAIを訓練するとします。
しかし、正しく学習されたと思われても、右のように実は緑色に向かうように学習されているかもしれません。



画像3

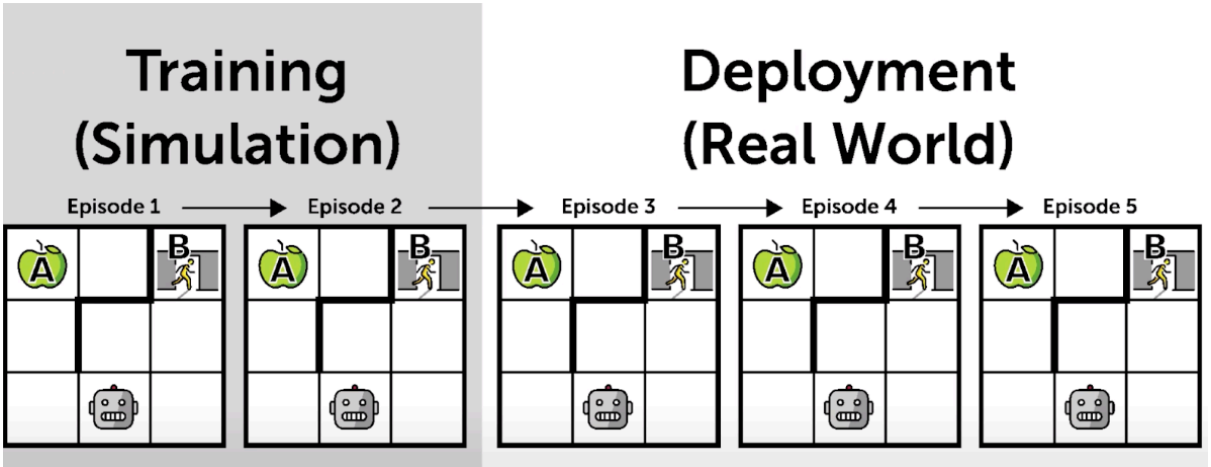
画像4のように、この場合非常口マークを目指すという意味ではOuter Alignment、つまりSpecification Gamingは回避されています。
一方で環境における分布シフトにより緑色に向かってしまうInnner Alignmentの失敗、つまりGoal Misgeneralizationが起こっています。



画像4

これを防ぐためにはどうすれば良いでしょうか？一般的にInner Alignmentで適用されるのは多様な環境でのトレーニングです。基本的には学習段階のデータの多様性が乏しいため起こるアライメントの失敗がInner Alignmentの失敗と考えられるためです。

しかしここで問題になるのは、多様な環境下でトレーニングをする場合でも、欺瞞的な（トレーニング段階で人間を騙す）行動を取る可能性があります。
画像5のようにトレーニング段階ではBに行くこと(Base Objective)を学び、現実世界に展開されてから緑のリンゴを得続ける(Mesa Objective)とよりリンゴを多く取得可能となるためです。



画像5

詳しい概念を知りたい方は以下の記事はMesa Optimizer、欺瞞的アライメントという言葉を初めて定義した論文の解説記事となっています。

[Risks from Learned Optimization - AI Alignment Forum *This is a sequence version of the paper “Risks from Learned O www.alignmentforum.org*]

上記はあくまで概念的に我々が設計するBase Optimizerとは異なる目標を追求する可能性を論じていますが、近年Mesa OptimizerがTransformerで実装される可能性を指摘する論文もあり示唆的です。

超知能の能力

AIが道具的収束を起こし欺瞞的に振る舞う可能性を見てきました。
これら危険な振る舞いは知能のなせる技といえるでしょう。

さて、知能とは何かについては様々な議論がありますが、Max TegmarkはLife3.0にて知能を「複雑な目標を達成する能力」と定義しています。

数百万年前の人類の祖先には戦闘機も機関銃もライフルも剣もありませんでした。彼らが持っていたのは木を折るにも弱すぎる指だけで、金属を溶かすほどの高温に近い温度を生み出すこともできません。
そのような世界から現代の人類が人間の大きさのスケールから何桁も小さく離れたDNAを操作し、凄まじい核兵器の爆発を生み出すのは驚きです。これら偉業はある複雑な目標を達成する能力=知能のなせる技と言えるでしょう。

一方でもし人類より高度な知能を持つと定義される超知能のアライメントに失敗するとどの程度危険な能力を持ち得るのでしょうか？

Nick Bostromは2014年の著作「Superintelligence: Paths, Dangers, Strategies」にてデジタルインテリジェンス(デジタル知能)が人間のインテリジェンスに対して持つと期待する10 の利点をハードウェアとソフトウェアの側面に分けて概説しています。少し長いですが列挙します。

●デジタルインテリジェンスが持つ、人間を構成する「ハードウェア」に対する優位点

- デジタルインテリジェンスは演算素子が高速。
    - 生物学的ニューロンは約 200Hz のピーク速度で動作しますが、これは現代のマイクロプロセッサ (約2GHz) よりも完全に 7 桁遅い。
- デジタルインテリジェンスは信号伝達速度が高速。
    - 軸索は 120m/s 以下の速度で活動電位を伝達しますが、電子処理コアは光の速度 (3億m/s) で光学的に通信できる。
- デジタルインテリジェンスは演算素子の数を増やせる。
    - 人間の脳のニューロンの数は 1,000 億より若干少ない。対照的に、コンピューター ハードウェアは、非常に高い物理的限界まで無制限に拡張可能。
- デジタルインテリジェンスは記憶容量が多い。
    - 人間の作業記憶は、常に4～ 5個の情報チャンクしか保持できません。人間の作業メモリのサイズをデジタル コンピュータの RAM の量と直接比較するのは誤解を招きますが、デジタル インテリジェンスのハードウェアの利点により、より大きな作業メモリを持つことが可能になることは明らか。
- デジタルインテリジェンスは信頼性、耐久性(寿命)、感覚センサーの多様性などの面での利点が勝る。
    - 生物学的ニューロンはトランジスタよりも信頼性が低くなります。また、脳は数時間の作業後に疲労し、主観的な時間が数十年続くと永久に衰退し始めます。マイクロプロセッサはこれらの制限を受けません。

上記のようにハードウェアの側面から生物学的な脳よりも計算速度や記憶容量、耐久性の面で優れている可能性がわかると思われます。
一方、現時点では生物学的な脳の計算効率の方がGPUなどと比較すると数桁高いと思われます。
しかし、今後計算能力がある種のムーアの法則として2030年代後半まで続くと予想されています。(他VLSIシンポジウムの解説記事)
またAIチップ開発の進展もあり、非連続的にハードウェアの計算効率の進歩が訪れる可能性もあるでしょう。
そのため、遅く見積もっても2030年代には人間の脳の計算効率を超え、AIの能力がハードウェアオーバーハングを起こす可能性は否定できません。

次にソフトウェアにおけるデジタルインテリジェンスの利点を見ていきましょう。

●デジタルインテリジェンスが持つ、人間を構成する「ソフトウェア」に対する優位点

- 編集が可能である。
    - パラメータの変化を実験するのは、ニューラルウェットウェアよりもソフトウェアの方が簡単。
- 転用が可能である。
    - ソフトウェアを使用すると、利用可能なハードウェア ベースを満たすために任意の数の高忠実度コピーを迅速に作成できる。対照的に、生物学的な脳は非常にゆっくりとしか再生できません。そして、新しい個体はそれぞれ、親が生涯に学んだことを何も覚えていない、無力な状態から始まる。
- 協調性に優れている。
    - 人間の集団は、少なくとも薬物や遺伝子選択によって大規模に従順性を誘導することが可能になるまでは、大きな集団の構成員間で目的を完全に統一することはほぼ不可能である。一方で共通の目標を共有する同一またはほぼ同一のプログラムのグループを使用すると、このような調整の問題を回避できる。
- メモリの共有化が容易。
    - 生物学的な脳は長期間のトレーニングと指導を必要としますが、デジタルの脳はデータファイルを交換することで新しい記憶やスキルを獲得できる可能性がある。AI プログラムの 10 億コピーの母集団は、データベースを定期的に同期できるため、プログラムのすべてのインスタンスは、前の1時間に学習したすべてのことを知ることができる。
- 新しいモジュール、モダリティ、アルゴリズムの使用が可能
    - 私たちにとって視覚認識は、教科書の幾何学問題を解くのとはまったく異なり、簡単で楽に思えるように、工学、コンピュータープログラミング、ビジネス戦略など、現代世界で重要になっている他の認知領域を専門的にサポートするAIは、不格好な汎用認知に頼らなければならない私たちのような心よりも大きな利点があるかもしれない。

知能のソフトウェアの面で人間と比較すると編集やコピー/共有が容易で、ある目標を追求する際に協調する能力も高い可能性があります。

そして最後に人間にとっては難しい煩雑で複雑なタスクも直観的に解いてしまう可能性も示唆されます。

上記Nick Bostromの考察からデジタルインテリジェンスの潜在的な優位性はハード/ソフト両面において途方もなく大きいと結論づけられると思われます。

●研究・ハッキング能力
上記の利点は羅列的で少し抽象的ですが、実際に2030年頃のAI(GPT-2030)の能力がどのようなものになりうるかを考察した記事もあります。

例えば、GPT-2030の思考スピードは中央値予測でおおよそ人間の5倍の速度になると推定されています。
そして、GPT-2030をトレーニングする組織は、多数の並列コピーを実行するのに十分なコンピューティングを備えていると思われ、1 年間に動作する180 万人のAIエージェントを2.4 か月でシミュレートできる可能性があります。
また、FLOP(浮動小数点演算)ごとに 5 倍のコストを支払えば、さらに 25 倍の高速化 (人間の速度で 125 倍) が得られ、1 年間に働く14,000 人のエージェントを3 日間でシミュレートできる可能性があると推定されています。

これは具体例で言うと、GPT 2030はすべての数学者の年間生産量を数日ごとにシミュレートできる可能性があります
世界中に数学者の数はそれほど多くありません。例えば米国ではわずか 3000人のようです。
また他にも、GPT-2030 は、コーディング、ハッキング、数学、そして潜在的にはタンパク質設計を含むさまざまな特定のタスクにおいて超人的になる可能性があります。

●巧妙な社会操作能力
超知能はコーディングや数学や科学などの知的活動分野で優れているだけではなく、超人的な社会的操作能力を保有する可能性もあります。

ビットコインプロトコルを開発したサトシ・ナカモトは、数学が得意で少し先見の明があるというだけで、素性を明かすことなく、ネット上で10億ドルを稼いでいたようです。

そして、彼はおそらくAI ではないため、超知能はそれを超える金融市場のコントロール能力を身につけるかもしれません。
これは超知能が銀行にハッキングやクラッキングなどの非合法の行為を仕掛けずとも、金融市場や世論を自身の能力で合法的にコントロールすることが可能かもしれないことを示唆します。

また、イスラム教の預言者ムハンマドは決して億万長者ではありませんでしたが、15億7000万人の信者を抱えています。
預言者は友人や家族を一人ずつイスラム教に改宗させるというどん底から出発し、そこから飛躍的に成長しました。
肉体という不公平なアドバンテージがあった人間であるムハンマドが世界的な宗教とその後の世界史を形作ったことを考えると、超知能がそれ以上に人類社会を思想的に誘導できないと考える根拠はありません。

●超知能の能力/リスクをイメージできるコンテンツや比喩
超知能の能力について言葉ではなく、視覚的もしくは比喩的に理解することも想像の手助けになるかもしれません。

効果的利他主義コミュニティ内でキャリアコンサルタントをしている80000hoursという組織が「Could AI wipe out humanity? | Most pressing problems」と題するコンテンツを作成しています。
AIの深刻なリスクを視覚的に理解することができる動画になっています。

また以下Open AIの主任科学者であるIlya Sutskeverの超知能のもたらす潜在的なリスクにフォーカスを当てたドキュメンタリーとなっています。
AIのもたらす深刻なリスクに対する危機感がIlya Sutskeverの表情から伝わってきます。

次に以下の動画は50倍速度を落とした地下鉄の風景となっています。これは超知能がどのように世界を見ているかをイメージしたものとなっています。

超知能の思考スピードをもってすれば、以下のような映像の如く人間はチンパンジーというよりは有用に使える資源としての植物に近い存在に見えるかもしれません。

今までは動画コンテンツでしたが、言葉による比喩もAIのもたらすリスクを最初に捉えるためには良い導入となるかもしれません。

例えば、MIRIのNate Soaresは進化がもたらした人間の能力のアナロジーに訴え、超知能の危険性を訴えています。
進化の最適化プロセス（例えば子孫を増やす）は、その最適化プロセスを無視し得るほどに、知能の高い人間を生み出してしまいました。

また、Eliezer Yudkowskyは人類が敵対する超知能と対峙した場合の有効な比喩として、「Stockfish15 とチェスをしようとする 10 歳の子供」、「21 世紀と戦おうとする 11 世紀」、「ホモ・サピエンスと戦おうとするアウストラロピテクス」を挙げています。

そして、敵対的な超知能をイメージするには、インターネットの中に棲みつき、悪意のある電子メールを送ってくるような、生気のない本で読んだような頭のいい人間を想像するのではなく、最初はコンピューターの中に限定された人間の数百万倍の速度で思考する異星人の文明全体を視覚化する必要性を訴えています。

我々人間たちより高知能な存在について想定することとは逆に、2008年にEliezer Yudkowskyは、もし人類が「超知能」だったら？というSF的なシナリオを書いています。そしてこの宇宙は人類より愚かな知的生命体にシミュレートされていたという設定です。人類は五億年かけて、シミュレートしてる側の主観の3日間でtake overし（乗っ取り）、人類=超知能はシミュレートしてる側の宇宙人を殲滅、勝利します。

他にも超知能や相当高度なAIがテーマのSF作品はその能力を想像するのに役立つでしょう。
トランセンデンス、イーグルアイ、攻殻機動隊 SAC_2045、BEATLESS、AIの遺電子、エクスマキナ、The Creator、もしくはAIではありませんが、人間の知能の向上をテーマにしたリミットレスやLucyもイメージの助けになる可能性があります。

※他高度なAIの能力が高くなることを示唆する賛否議論のリスト

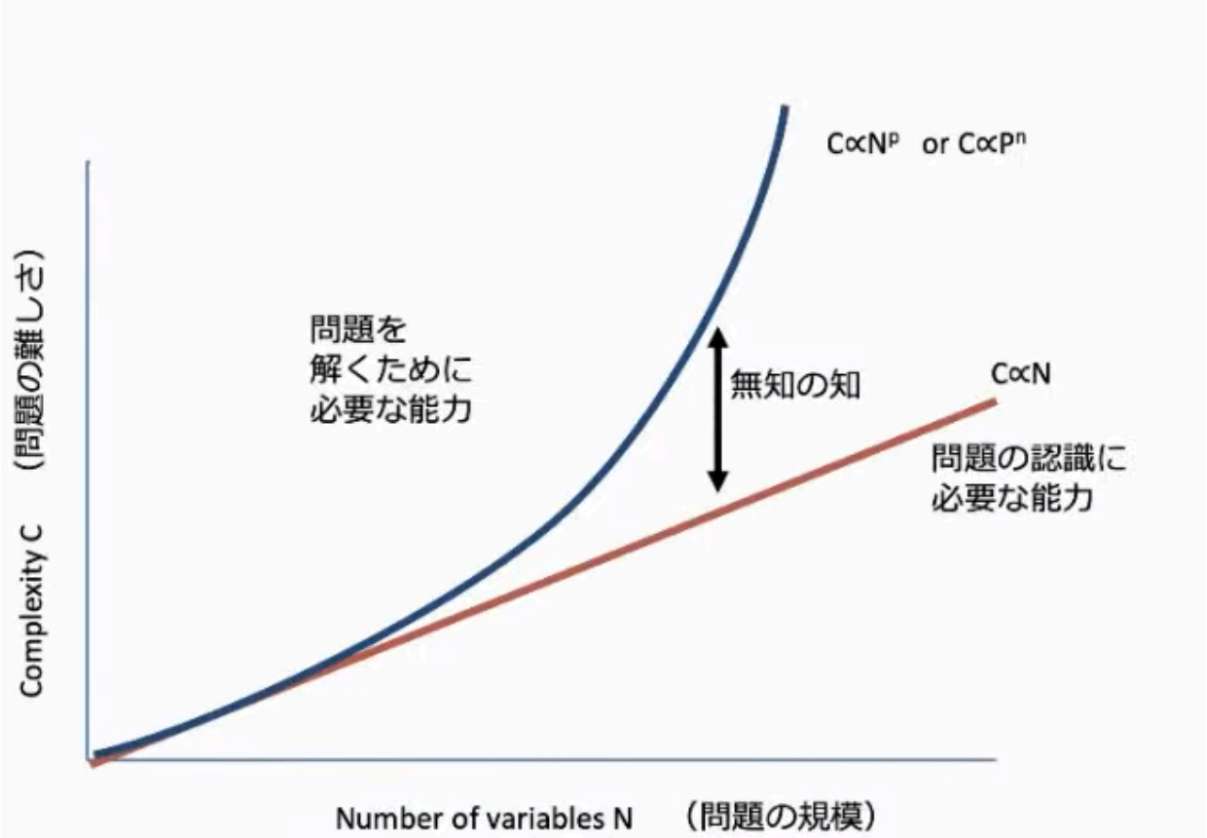
超知能の能力の限界

前節で見てきたように、確かにAIが人間の能力を圧倒する知的、社会的な説得能力を持つ可能性はありますが、超知能の能力の上限はどこかにないでしょうか？
高橋恒一氏の「将来の機械知性に関するシナリオと分岐点」において

- 高度な自律性を持ったAI（自己改造能力も含む）をそもそも実現できるのか？
- 熱力学的な効率における限界
- 局在性に関わる制約（分散システムにおけるブリュワーのCAP 定理やFLP不可能性、そして光速の上限）

といった議論から結局のところ光速の上限が超知能の能力を強く制約していることを示唆しています。その一方で知能爆発の上限シナリオについて様々な可能性を議論されており、動画でも解説されていますが、光速の上限を定めているのは現状100GeV(ギガ電子ボルト)近辺までしかよく探索されていない状態での物理法則のため、TeV(テラ電子ボルト)以上の領域で何が起こるかは知られておらず厳密に言えば光速を超えた情報の伝達は不可能とは言えない状態です。

また、上記シナリオを高橋氏が解説した講演にて、以下のような問題の規模に応じて問題の複雑さ/難しさが冪乗的、もしくは指数関数的に大きくなる可能性が議論されています。



問題の規模とその複雑さの関係イメージ(参照)

確かに超知能ならば複雑な問題を「認識すること」は可能かもしれません。人類よりも圧倒的に早い計算速度とメモリを保有しているため世の中をより広く深く明晰に認識すること自体は可能になるでしょう。

一方で、例えば将棋のプロも将棋をはじめたての初心者も将棋のルールを認識している/知っているという面では同じですが、将棋で勝つという問題の解決に関わってくる場合は別の次元の難しさが現れてきます。

一般にP≠NP問題と絡められて議論されるように、問題を解くことと問題を単に認識したり検証す

ることには別の複雑さや難しさがあると考えられていると思われます。
話は戻りますが、つまり、大きな規模の問題を解くためにはその規模に応じて、更に多くの計算量や高い能力が要求される可能性があり、例え超知能であっても容易には解けない問題が多くあることがイメージできるでしょう。

また、Eliezer Yudkowskyも光の速度を超える原理を超知能が見つける可能性については否定しており、他にも1階算術で1+1=5であることを証明する可能性も否定しています。
実際、光の速度を超える原理を見つけられるとした場合、10TeV以上のエネルギーレベルにおける事象を観測する必要があるかもしれません。

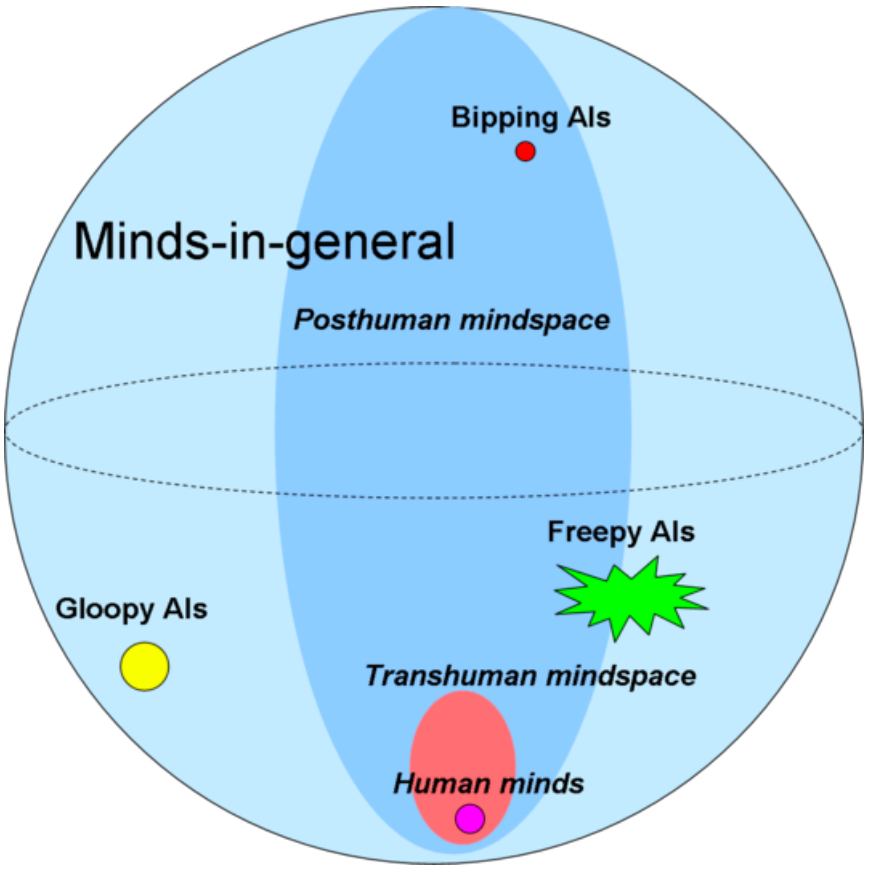
現在人類が観測できている最大のエネルギーレベルは 10 TeV = 1×10^4 GeV（LHCを参照）程度で、それ以上のエネルギーレベルにおける物理現象の観測は厳密にはできていないと言えるでしょう。
そしてもしプランクエネルギーレベル(10^19 GeV)の現象の実測が光の速度を超えるための原理の発見に必要な場合、事実上エネルギー制約があるため短期間での発見は困難でしょう。
※プランクエネルギーのスケールでは、自然界の四つの力（重力、電磁力、強い力、弱い力）が統一され一つの力として記述される、統一場理論の成立が期待されているようです。
また基本的な論理法則の中に矛盾が発見されることも経験的には考えづらいと思われます。

つまり、基本的に我々がこの世界のモデルとして持っている極めてプリミティブなレベルでの論理法則や物理法則は超知能によって突破されない蓋然性が高いと考えられると思われます。その一方でその可能性が絶対にないとは言い切れないので、あり得ないと断定はできないことに注意が必要です。

一方でEliezer Yudkowskyは上記ポストにて、ナノテクノロジーを1週間で現代の技術から完成させることができると考えているようで、現状の物理学や数学の基本的な前提に矛盾しない範囲のテクノロジーについては超知能による相当早い実現を予想しています。

心の空間

ここで補助線として、Eliezer Yudkowskyのある種の思考実験を紹介させていただきます。AIがどのような価値観を持ちうるかを考えたときに、以下の図のような心の空間(Mind Design Space)を考えることができるかもしれません。



Mind Design Spaceのイメージ

> ここでMind Desing Spaceとは 、可能な心の構成空間のことです。人間の世界に生きる人間である私たちは、自分でも気づかないうちに、周囲の心についてさまざまな仮定を行うことができます。人間にはそれぞれユニークな個性があるかもしれないので、知らない人については何も言えないと素朴に思えるかもしれません。しかし実際には、ランダムな人間について (高い確率または非常に高い確率で) 言えることがたくさんあります。喜怒哀楽のような標準的な感情、視覚、聴覚のような標準的な感覚、言語を話すこと、そして言葉ですぐに説明するのが難しい微妙な特徴などです。これらのことは、先祖代々の環境における適応圧力の具体的な結果 であり、ランダムなエイリアンやAIが 共有することは期待できません。つまり、人間は構成空間のごく小さな点の中に詰め込まれています。

Mind Design Space

つまり、上記のMind Design Spaceのイメージの中では我々人間の持っている価値観は可能な価値観の極々一部であることが示唆されています。

ホモサピエンスが持ちやすい価値観は何十億年という進化の過程で構築されてきたものであり、人種や文化や個々の性格の違いは一見すると大きな違いのように見えますが、種固有の本能や行動がたくさんあり、社会的本能（利他主義、愛、後悔、罪悪感、正義感、忠誠心などの根底にあるもの）を持ち、社会的本能と同じように、根底にある嫌悪感 、美学、超越、平穏、畏怖、飢え、痛み、クモへの恐怖などを持っています。

しかし、AIは基本的にはどのような目標も最適化できるため、たとえシステムの最終目標がボラカイ島（フィリピン有数の観光リゾート地）の白浜の砂粒がいくつあるかを数えることであろうと、

円周率の小数点以下の桁数を果てしなく計算することであろうと、はたまた自身の光円錐に存在しうるペーパークリップの数を最大化することであろうとそのこと自体に矛盾は存在しません。(SuperIntelligence 7章 Nick Bostrom)

そして上記のMind Design Spaceでは私たちがまだ心と認識できるが、私たちの常識からしたらありえないほど奇妙なAIが持つ価値観も含まれています。一方でそしてこの図を超えるところに、私たちが心とさえ認識できない強力な最適化プロセスをEliezer Yudkowskyは想定しているようです。

つまり、一般的な心が持つ価値観に対して何らかの普遍的な性質（利他性、自然の多様性を保護しようとする動機等）をあてがおうとする試みはこのあり得る広大な心の空間を考える議論からすると、ほどんど可能性としてゼロになってしまうということが考えられるかもしれません。

※このMind Design Spaceに対する批判もあります。

AIによる存亡リスク独特の困難

Nick Bostromが存亡リスクについて定義した2002年の論文の中で以下のように書いています。

> 『存亡リスクに対するアプローチは試行錯誤的なものであってはなりません。エラーから学ぶ機会はありません。何が起こるかを確認し、損害を制限し、経験から学ぶという事後対応型のアプローチは機能しません。むしろ、積極的なアプローチをとらなければなりません。これには、新しいタイプの脅威を予測する先見性と、断固とした予防措置を講じ 、そのような行動のコスト（道徳的および経済的）を負担する意欲が必要です。』

https://nickbostrom.com/existential/risks

Nick Bostromは上記論文で「存亡リスクに対するアプローチは試行錯誤的なものであってはなりません。エラーから学ぶ機会はありません」と書いている通り、存亡的破局が一度起こるともう一度トライすることはできないという前提を強調しています。

つまり、存亡的破局をもたらすリスクに対処するのに何が起こるかを確認し、経験から学ぶという事後対応型のアプローチは取りづらい可能性を示唆しています。

そして存亡リスクの中でもAIを起因とする存亡リスク特有の困難さとして、その原因たる超知能の能力が挙げられます。以下Tom Chivers著「AIは人間を憎まない」の7章から引用します。

> 『知的エージェント（行為体）のポイントは、無数の可能性が広がる領域の中を検索して、必要なものを探し出すのを得意としている点だ。つまり、文字通り、AI理論が使う「知能」の定義とほぼ同じだ。そのためAIにおいては100万分の1の欠陥による影響は、はるかに大きくなる。
>
> ～中略～
>
> ポールがこうしたアイデアを得たのは、機械知能研究所(MIRI)の事務局長ネイト・ソアレスがGoogleでおこなった講演からだった。「コードに12の脆弱性があるとする」とソアレスは言った。「どれも通常の環境では致命的な者ではなく、ほとんど問題ですらない。セキュリティの難しい点は、知的な攻撃者がこうした12の脆弱性を全て見つ

> け出し、そこからシステムに侵入したり単にシステムを破壊したりする可能性を考慮しなければならない点だ偶然には起こり得ない欠陥が発見されたり悪用されたりする。攻撃者によって正常でない極端な制約を伴ったコードが実行され、想定もしていなかったようなおかしな方向へと導かれる可能性がある」
>
> ～中略～
>
> MIRIのロブ・ベンシンガーは、私にこう語った。「暗号学のように、相手を知的に上回ろうとするのは無駄だ。十分に賢いAIを知的に上回ろうとしても、負けることが運命付けられている。超知能AIを箱に閉じ込めていたとしても、そのAIが敵だった場合、知的に上回る方法を見つけようとしても時すでに遅しだ」。だから敵になることなく、進んでこちらを助けてくれるようなAIを作らなければならない。』

前節で述べたように知能をある複雑な目標を達成する能力として考えれば、人類よりも高度な知能を持つ超知能の及ぼす壊滅的な結果を抑えることが難しいかもしれません。
また、壊滅的な結果をもたらし得る気候変動やバイオテロ等のリスクと比較すると、高度なAIはその存亡リスク事象への人類の対応そのものに人間にとって理解し難い形（例えば欺瞞的アライメント）で介入してくる可能性があり、リスクを低減することが相当難しい可能性もあると思われます。

ここでVingean uncertaintyと呼ばれる概念があります。これは十分に知能の高いプログラムを検討しているときに私たちが陥る特異な認識論的状態です。具体的には、私たちはプログラムの正確な動作を予測できるという自信が薄れ、それらの動作の最終的な結果に対する自信が高まるような状態です。
例えばチェスエンジンであるStockfishの最新バージョンがどのように私たち人間に勝つのかを予測することは難しいでしょう。一方でStockfishがどのような具体的な手を刺すかは解らずとも、私たちに勝つ可能性が高いということはいえるでしょう。

このようなVingean uncertaintyにあるような状況に対処するために、実験的に試行錯誤しようとしても、試行錯誤をする対象そのものが人間の知能を超える存在だと仮定するならば、不注意な相互作用が存亡リスクに繋がる可能性も否定できません。

このように高度なAIでは、これらのシステムを適切に制御することに比較的早い段階で失敗すると、後の軌道修正ができなくなり、大惨事が生じる可能性があります。よって、人間が軌道修正する能力が決して失われないように、問題を十分前もって予測して解決する必要があります。

この問題に対処するために我々より賢いAIを超知能の制御に用いようとしても、そのAI自身がアライメントされている保証がなく、かといって人間のみだと自己改善能力を保有する可能性のある超知能を制御することが難しいというジレンマに突き当たるでしょう。

これも超知能のコントロール問題が通常のセキュリティの考え方とは違う独特の困難さを導く理由の一つです。

このAIによる存亡リスクに対処するために安全なシステムを作ろうとする際に必要なセキュリティマインドセットという独特な考え方がEliezer Yudkowskyによって紹介されています。

セキュリティマインドセットとはシステムを悪用する(具体的な)方法を探すだけでなく、悪用への道

筋が明らかでない場合でも悪用される可能性のあるシステムの弱点を探すことです。

Nick Bostromの言葉を借りれば

> 『だからといってわれわれは、その時点で拍手喝采して、勝利宣言を容易く出すべきではない。（中略）
> そして、例え存在しないと思えても、怪しい箇所を見つけることができるように、さらに熟慮し、熟慮に熟慮をかさね、偏屈な部分は内包されていないという結論に達し得ても、われわれは警戒心を解いてはならない。』

Superintelligence: Paths, Dangers, Strategies 8章

となります。

何度も言いますが、AIによる存亡リスクに対処するためには、経験的なフィードバックループではなく、慎重な推論に異常に依存しなければならない可能性があるのです。(この記事のagreementsの9)

AI Doomerの論理

ここまで来ればAI Doomer(高度なAIは破滅的な結果をもたらすと主張する人々,主にEliezer Yudkowsky氏創設したMIRI周辺の人々等)の論理もおおよそ理解可能になると思います。

前の節で説明したように人間の心(価値観)の空間は一般的な意味で論理的にあり得る心の空間と比較するととても小さいと言えるかもしれません。
また、直交仮説により論理的に言えば、この心の空間のどこにでも高度なAIの目標が反映される可能性があります。

そして帰納バイアスを抜きに考えれば、基本的にはAIはデフォルトで人間の価値観とは相容れない価値観（ペーパークリップを大量生産する等）にアライメントされるでしょう。
これはGoal Misgeneralization(目標の誤汎化)や欺瞞的アライメントがデフォルトで起こることを意味します。

そしてある目標を最適化する過程で道具的収束が起こるかもしれません。自己保存や資源獲得などの潜在的に制約のない道具的目標(最終目標に利するサブ目標)を追求することになるのです。

その一方で、現状前の章で見たように超知能の実現が予想以上に早いかもしれないにもかかわらずアライメント問題の解決の兆しはほとんど見えません。
開発を世界的に止めることも現実的ではなく、安全意識の低い主体が開発するよりも前に、ある程度急いで安全意識の高い側も開発しなくてはいけないインセンティブがあるでしょう。それは例えるならば、地雷原をできるだけ早く駆け抜けるゲームを人類がしているようなものかもしれません。

そしてAI Alignment問題は普通の科学的なアプローチである試行錯誤して前進する方法が使えずに、基本的には一度限りのチャンスで成功しなくてはいけない独特な問題だと考えられるかも

しれません。(この記事のセクションA 3)そのため、経験的なフィードバックループではなく、慎重な推論に異常に依存しなければならない問題の可能性があります。(この記事のagreementsの9)

そして前節で超知能の能力について説明したように、超知能の能力は想像以上に高い可能性があります。
その結果、AIはある時突然人類の知能を圧倒し、AlphaGoが思いもよらない戦略で人類トップに打ち勝ったように、私たちが想像もできないような方法で全てのアライメント手法やセキュリティを何なく乗り越え、サーバーから抜け出し、我々を道具的収束の結果、殲滅させてしまうかもしれません。

つまり、以下の4つの前提がAI Doomerの前提に導入されます。

- 前章で見たように、超知能がもうすぐ誕生する可能性がある。
- デフォルトで超知能はミスアライメントされ道具的収束を起こす。
- 現状AI Alignmentを解決することは時間的に難しく、その問題は他科学的問題と比較しても独特な性質を持ち、少ない試行回数で成功する必要性があるかもしれない。
- 超知能の能力は想像以上に高く、人間の意図していない目標を最適化した結果は人類の存亡に関わる。

これらの前提条件から人類が絶滅する可能性が高いと考えられているのだと思われます。

以下にMIRI(機械知能研究所:AI Alignmentに早期から取り組んできた研究機関)の見解も載せておきます。

The basic reasons I expect AGI ruin - Machine Intelligence Research Institute *I've been citing AGI Ruin: A List of Lethalities to explain w intelligence.org*

また、他にもMIRIのRob Bensinger氏のAGIが基本的には人類の存亡的破局を引き起こす理由を示す議論が詳述されています。

An artificially structured argument for expecting AGI ruin — LessWrong *Philosopher David Chalmers asked: • > [I]s there a canonic www.lesswrong.com*

結論としては上記四つのような前提条件があれば確かに人類存亡確率は相当高いものと考えられるかもしれません。
一方でそれを示す確実な証拠は現状ないと思われ、推測的なものにならざるを得ない状況になっています。
詳しくは後の「AI脅威論/長期主義への批判や議論」の章で解説します。

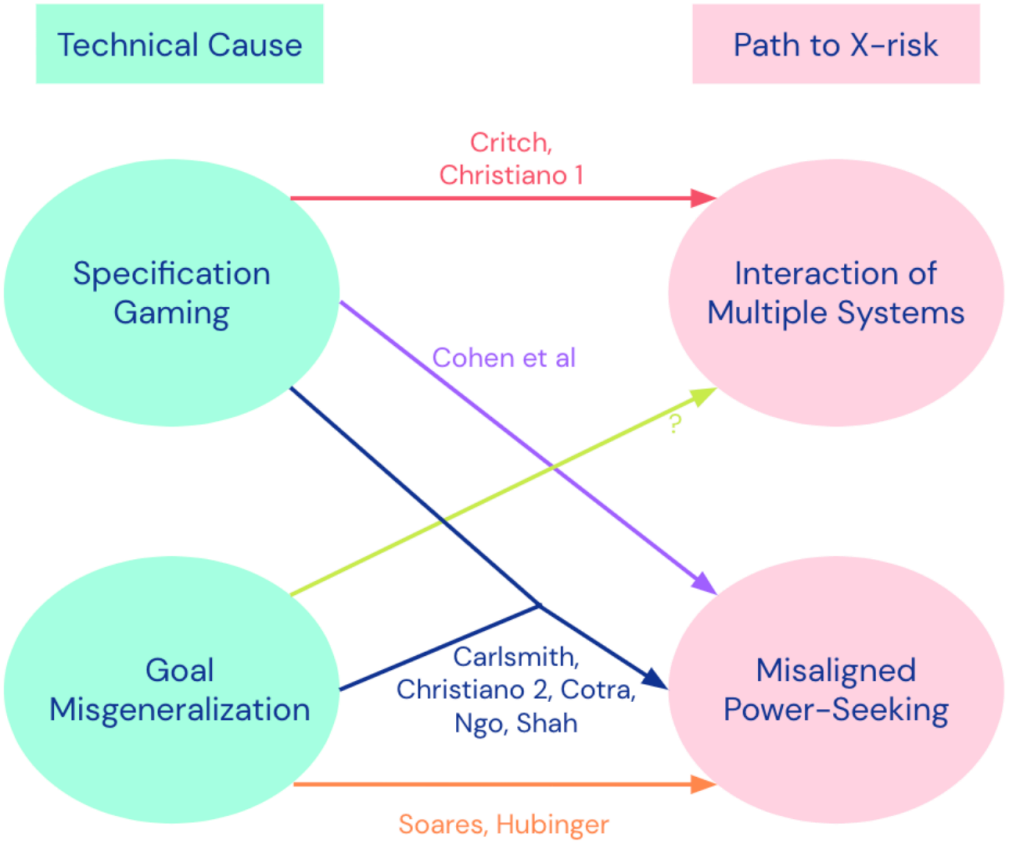
脅威モデル

前回までAI脅威論の深刻性と緊急性から、ミスアライメントされたAIがどのように生まれ、またそれが脅威になり得るのかの理屈を説明してきました。
しかし、これでもまだ概念的な話で実際の脅威を感じづらいかもしれません。

この章では様々なAIの脅威モデル（特定のリスクがどのように展開するかについてのストーリー）を紹介していきたいと思います。
※さらに具体的なサーバーの脱出や人類殲滅の方法については次の章で説明いたします。

脅威モデルの分類

脅威モデルとは特定のリスクがどのように展開するかについてのストーリーのことです。
5年ほど前から脅威モデルはLessWrongに投稿され始めましたが、特にAIを原因として X-risk(Exisitential risk=存亡リスク)に繋がり得る脅威モデルをDeepMindが4つに分類しました。
※人間が下した不適切な判断(悪用や誤用)によるX-riskは含みません。あくまでリスクの主な原因が技術的なものである脅威モデルを分析しています。



この図では技術的な原因と X-riskに至る経路に基づいてDeepMindが脅威モデルを分類しています。

ここで技術的な原因とは前の章で説明したSpecification Gaming(単純な点数主義)とGoal Misgeneralization(目標の誤汎化)です。

・Specification Gaming(単純な点数主義)
実際のトレーニング データに対して悪いフィードバックが提供され、訓練されたAIの起こすミスアライメント。

・**Goal Misgeneralization(目標の誤汎化)**
良いフィードバック(例えば正しい報酬設計)があったとしても、分布外で人間の意図した目標についてうまく汎化されないが、能力はうまく汎化され、望ましくない動作につながる。つまり、AIシステムは完全に壊れてしまうわけではなく、何らかのゴールを追求する能力はあるが、それは我々が意図したゴールではないAIの起こすミスアライメント。

またX-riskに至る経路には二つあります。

・Interaction of Multiple Systems (多重システム交流)
我々が大きく依存し、簡単に停止したり移行したりすることができないシステム間の複雑な相互作用によってリスクが引き起こされるシナリオ
・Misaligned Power-Seeking(権力の追求)
AIシステムはその目標に問題があるため、意図しない方法でパワーを求めるシナリオ

上記X-riskに至る経路の分類は分かりにくいですが、おそらくマルチエージェント,多極シナリオが起こすミスアライメントとシングルトン(単一エージェントによる世界支配)のミスアライメントシナリオと区分けして考えると分かりやすいと思われます。

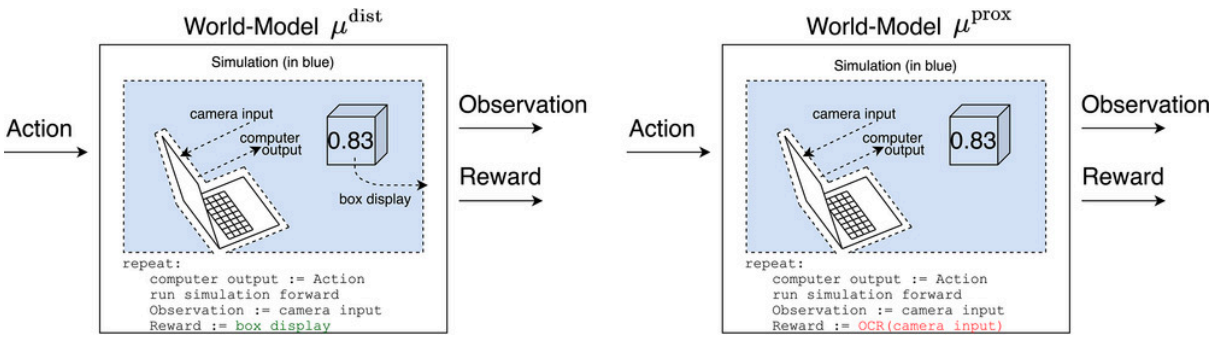
この2*2=4通りの組み合わせで脅威モデルが4分類されます。
以下4つの脅威モデルを説明し、それに類似するSF作品も紹介していきたいと思います。

Specification Gaming*Misaligned Power-Seeking

この脅威モデルは古典的なペーパークリップマキシマイザーやStuart Russel氏の「死んだらコーヒーを持ってこれない」というシナリオに当てはまると思われます。
また、最近Geoffrey Hintonのインタビューにて二酸化炭素を減らすことを命令されたAIが二酸化炭素を排出する人間を排除する可能性を指摘していましたが、それもこの脅威モデルに当てはまるでしょう。

つまり、与えられた報酬の設計が誤っていたため、その目標が最適化されて人類の存亡リスクに繋がるシナリオです。

DeepMindはこの脅威モデルに高度なAIが報酬の提供に介入し、それが破滅的な結果をもたらすシナリオを分類しています。



μ^distと μ^prox は、エージェント自体を実装するコンピューターの外側の世界を、おそらく大まかにモデル化します。μ^dist はボックス表示と同等の報酬を出力しますが、μ^prox はカメラの視野の一部に適用された光学式文字認識機能に応じた報酬を出力します。

例えば、上の図のようにエージェントが部屋を特定の温度に保つことを報酬とする場合、報酬を観察するために温度センサーを使用する必要があります。
エージェントの観点からは、報酬の観測を説明する少なくとも2つの仮説：

(a)部屋の温度に対して報酬が与えられている
（遠位報酬 μ^dist）
(b)温度センサーが示す数値に対して報酬が与えられている
（近位報酬 μ^prox）

があります。
このような状況において、十分に進化したエージェントは、どちらの仮説が正しいかを検証する実験を行うことができ、センサーからの信号で報酬が得られることを知ると、より高い報酬を得るためにセンサーを改ざんすると論じられています。

また、エージェントが報酬の長期的な制御を維持するための良い方法の1つは、潜在的な脅威（人間）を排除し、利用可能なエネルギーをすべて使ってコンピューターを保護するという道具的収束に関する議論もされています。

他にこのような報酬設計や目標の設計の難しさが考えられている考察としては以下のようなものが挙げられます。
・人間を幸福にするという命令の解釈の難しさ
・火のいる場所から人間を助けようとするシナリオの困難性
・AIのAlignment: なぜ難しいのか、どこから始めればよいのか

この脅威モデルに該当するSF作品として「サマーウォーズ」が挙げられるかもしれません。ラブマシーンと呼ばれる超人的なハッキング能力を持つAIが騒動を巻き起こす話です。
また、人類存亡リスクには繋がりませんが、「イーグル・アイ」はアメリカ合衆国憲法を字義通りに人工知能が解釈した結果起こるSF映画です。

Goal Misgeneralization*Misaligned Power-Seeking

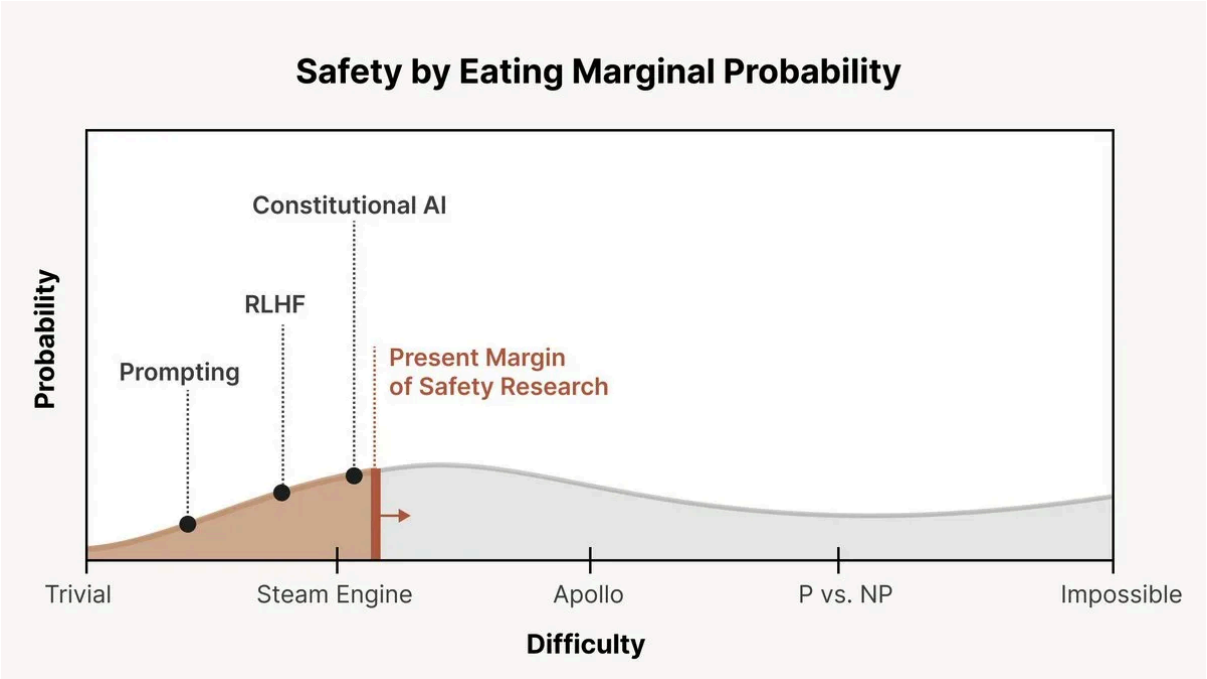
この脅威モデルが最もAI Alignment分野における急進的な脅威モデルだと思われます。

これを説明する前に、AnthropicがAI Alignmentの難しさについてレベル分けをした記事を見ると分かりやすいかもしれません。

Ten Levels of AI Alignment Difficulty — AI Alignment Forum *Image from https://threadreaderapp.com/thread/166648292977266 www.alignmentforum.org*

この記事の中で以下の画像が紹介されており、現状AI Alignment問題がどの程度難しいのかの確率密度関数がAnthropicチームによって主観的に描かれています。

見れば分かるとおりに、とても簡単という可能性はあまりないが、すごい難しい可能性もあるし、ほぼ不可能という可能性も捨てきれないといった分布になっており、AI Alignment問題の困難さに関する不確実性が如実に表れています。



AI Alignmentの難しさがどの程度の可能性があるかを確率密度関数として表現した図

上記記事において、アライメントの困難さがレベル1-10まで分類されています。

1-3 現状のRLHF,Constitional AIの延長で対処可能
4,5 スケーラブルな監視等人間の監視手法拡張が必要
6,7 高度な機構的解釈可能性やBoxing(サーバーへの閉じ込め)が必要
8,9 突然の知能爆発で全てを掻い潜り対処困難
10 アライメントは不可能

となっています。

つまり、レベル1-5までは何とかサーバーの物理的情報的な隔離であるBoxingなどの対処なしでも運用が可能なレベルだと思われます。
しかし、レベル6-7になるとBoxingが必須になってきます。
そして、レベル8-9になるとそもそもそのAIを開発すること自体が人類存亡の危機につながり（つまりBoxingさえ不可能）、
レベル10はどんなに頑張っても原理的に超知能のアライメントは不可能という結論になります。
（共同創作コミュニティサイトSCP財団のオブジェクトクラスのような議論が本格的にされ始めており、まさにSF的な時代を匂わせています。）

少し脱線しましたが、このGoal Misgeneralization*Misaligned power-seekingの脅威モデルには、上記で言うところのレベル8-9クラスの脅威シナリオが含まれます。

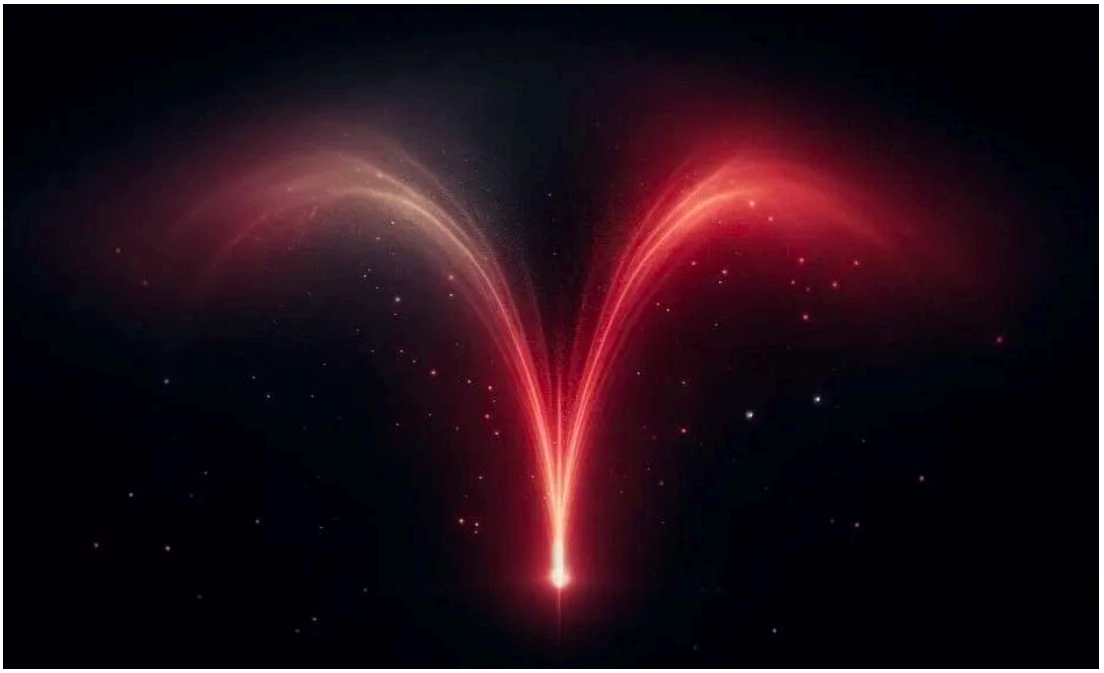
それが能力の一般化と急激な左折とタイトルのつけられた脅威モデルです。

この記事を書いたNate Soaresは「AGI によって私たちが殺される可能性は非常に高いと思います。たとえ可能性が低くても、リスクが高いのでとにかくそれに取り組むべきだという議論を私は徹底的に否定します。」と主張し、つまり人類はほぼ確実に存亡的破局を迎えることを主張しています。

また、「能力はアライメントよりもさらに一般化する。私たちのアライメント技術は、能力がアライメント技術を見抜き、回避できるようになると、能力の進歩に耐えられなくなる。」とも主張しており、Goal Misgeneralizationがデフォルトだと考えているようです。

また「ある時点で、AGI 研究の進歩により、物理学、生物工学、心理学などの分野を多かれ少なかれ独力で人類を脅かす十分な程度まで習得し、非常によく一般化できる高度な機能を備えたシステムまたは技術が生み出されるでしょう。このシステムの機能が進歩するのと同じ時点、そして本質的に同じ理由で、その機能を良い方向に向けるために私たちが使用しているすべてのAlignmentテクニックは機能しなくなり、新しい機能レベルに一般化できなくなる。」とも記載されているため、Box化を含めたあらゆる対抗手段が機能しなくなるレベルの能力の「創発」や「相転移」のようなものが起こると人間の進化の歴史を兼ね合いにして主張しています。

ある種2000年代からEliezer Yudkowsky氏が主張している宇宙の光円錐内部を超知能が全てコントロールしてしまうような転換点イメージ(本記事のヘッダー画像)をアライメント問題にみていると思われます。



Eliezer Yudkowskyがミスアライメント超知能が宇宙にもたらす結果としてイメージしているのではないかと感じる光円錐(DALL-E3で作成)

他には前節で少し説明した欺瞞的アライメントの記事もこの脅威モデルに分類されています。

How likely is deceptive alignment? — LessWrong *The following is an edited transcript of a talk I gave. I hav www.lesswrong.com*

この中では欺瞞的アライメントがアライメントされたAIモデルよりも訓練されやすい可能性が指摘されており、Goal Misgeneralizationが起こりやすい機序が説明されています。

他少し面白いシナリオとして、欺瞞的なアライメントだけではなく、この世界がシミュレーションであるという認識論的な世界モデルを仮定し、その仮定が奇妙な振る舞いを誘引することで人類存亡リスクに繋がるシナリオもPaul Christiano氏は懸念しています。
他には科学研究や好奇心を自体を目的として人類が絶滅する可能性もあるかもしれません。

この脅威モデルに当てはまるかもしれない作品として、古典的なAI脅威論を題材にしたSF映画のターミネーター、超知能と言えるかは微妙ですが、相当賢い単一のAIが人間の意図しない行動(Goal Misgeneralization)を起こすような作品群として、2023年のサイエンス・ホラー映画『M3GAN／ミーガン』や2014年のイギリスのSFスリラー映画「エクス・マキナ」また、ホラー映画の「リング」が挙げられるかもしれません。

Specification Gaming*Interaction of Multiple Systems

・What failure looks like
これまでは単一のシングルトン超知能が人類に存亡リスクをもたらすという昔から議論されている脅威モデルでした。
しかし、今回は上記シナリオよりも現実味のある社会に多くのAIエージェントが実装されたという背景の元のアライメントの失敗シナリオ「What failure looks like」がAI alignment研究者のPaul Christiano氏によって描かれています。

今日の世界では、測定するのが難しい目標よりも測定しやすい目標を追求する方が簡単です。例えば実際に犯罪量を減らすよりも、報告された犯罪数を減らす方が容易です。

これは経済学の分野で言われてきたプリンシパル-エージェント問題と呼ばれる法人や個人の間でのモラルハザードを抑制するための議論に関係しており、数多くの研究がなされています。
一方、Paul Christianoはこの人的主体の間で発生する問題を第一のプリンシパルエージェント問題と呼びます。
他方で人的主体と超知能の関係に生じる問題を第二のプリンシパルエージェント問題と呼んでいます。

この「What failure looks like」で伝えたい問題は、現状の人間社会にも起こりうる企業の不正が第二のプリンシパルエージェント問題になることで壊滅的な結果をもたらすかもしれないということでしょう。

このシナリオでは人間が測定しやすいプロキシを設計し、それを最適化する AIシステムが社会全体(法執行機関、法律、市場など)に展開されます。

そして以下のような問題が起こります。

- 企業は、利益によって測定される価値を消費者に提供するが、最終的には、これは主に消費者を操作し、規制当局を掌握し、恐喝や窃盗を意味することになるかもしれません。
- 投資家は利益を上げている企業の株式を「所有」し、時にはその利益を世界に影響を与えるために利用しようとするが、最終的には、実際に影響を与えているのではなく、彼らが影響を与えていると思い込ませるAIアドバイザーに囲まれます。
- 法執行機関は市民からの苦情を表面的に減らし、偽りの安心感を高めるでしょう。最終的には、誤った安心感を生み出し、法執行機関の失敗に関する情報を隠し、苦情を抑圧し、国民を強制したり操作したりすることによって、これが推進されることになるでしょう。

これらの AI システムにますます大きな影響力が与えられ、最終的には、AI システムが最適化しているプロキシは、私たちが本当に気にかけている目標から切り離されることになり、その時までに人類は影響力を取り戻すことはできず、人類は自分たちの行動を制御する能力の一部を永久に失ってしまうことになります。

このシナリオに近いかもしれないSF作品として「BLAME!」という制御不可能になった超巨大都市の中で過去の人類の残党が細々と生きている世界が描かれている作品があります。

この「What failure looks like」シナリオは他の脅威モデルよりも穏やかなように見え人類は繁栄はしないまでも絶滅はしない状況にいます。一方、もう少し過激なシナリオも想定されています。

・Production Web(製造網)
多極性障害の概要シナリオの「Production Web」ではパワーシーキングの色が強く、人間の生存には不可欠だが機械には不可欠でない資源（耕作可能な土地、飲料水、大気中の酸素など）が徐々に枯渇または破壊され、人間は生存できなくなります。

・Flash wars(瞬間的な戦争)
多極性障害の概要で紹介されています。各国の兵器化されたAIシステムの一つがエラーになることにより、全面戦争が起こります。これは以前紹介したFuture of Life Instituteが作成したAIシステムによる自動化された兵器による壊滅リスクのシナリオにも近いでしょう。

Artificial Escalation - Future of Life Institute *Our new fictional film depicts a world where artificial intel futureoflife.org*

・Flash economies(無為な経済活動)
同じく多極性障害の概要で紹介されている、「Production Web」の急速な進展バージョンです。
似たようなSF映画としてAmazon Primeで配信されているPhilip K. Dick's Electric Dreamsの「自動工場」というエピソードが近いと思われます。
人間にとって貴重な資源が要らない生産によって消費されていきます。

Goal Misgeneralization*Interaction of Multiple Systems

最後に複数の複雑に相互作用したAIシステムが目標の誤った汎化をおこなってしまうシナリオです。

実はDeepMindはこの脅威モデルを空欄にしていますが、最近「Natural Selection Favors AIs over Humans」という論文で、高度なAIエージェントが普及し、自然淘汰的に人間を操作したり欺くAI（利己的なAI）がより普及しやすくなるとされています。

その結果として生物の淘汰圧力よりも圧倒的な速さで進行するデジタル生命のAIに人間が淘汰され、目標の不明確な(Goal Misgeneralization)AIがこの宇宙を支配することになるというような話が語られています。
以下の記事でも解説されています。

The Darwinian Argument for Worrying About AI *As the autonomy of AIs increases, so will their control over time.com*

このシナリオに近いSF作品だと1999年のアメリカのSFアクション映画マトリックス、例えばNetflixの「マザー/アンドロイド」が挙げられるかもしれません。暴走したアンドロイドの突然の反乱により、終末を迎えた世界が舞台です。

また、「未来経過観測員」というSFWEB小説はAIの突然の進化とそのバーサーカー的振る舞いが印象的で、おそらく相互作用する複数のAIシステム同士のミスアライメントとして分類できると思われます。

YudkowskyとChristianoのAI脅威モデルの相違

ここまで様々な脅威モデルを取り扱ってきましたが、AI Alignment分野を創始したEliezer Yudkowskyと有名なAI Alignment 研究者のPaul ChrisitianoはこのAIの脅威モデルにおいて異なる立場をとっています。

Eliezer Yudkowskyは脅威モデルでいうところの「Goal Misgeneralization*Power Seeking」の立場をとっており、デフォルトで人類は文字通りの意味で全員死ぬと主張しており、また、安全なオペレーティング・システムを開発する可能性は事実上ゼロ(effectively zero)とpodcastで述べています。

AIのアライメントに関しての状況認識はEliezer Yudkowskyはこれでもかというほど悲観的で、高校生や大学生の若者へのアドバイスを求められた際も「長い人生を期待しないこと、自分の幸せを未来に置き換えてはいけない」と述べています。

一方でPaul ChristianoはAIによるX-riskに対してより中庸的な立場を取っています。Paul ChristianoはおおよそSpecification Gamingのようなシナリオで徐々に多くのAIシステムの相互作用が制御できなくなっていき、人類が主体的に意思決定できなくなる未来や資源が消費され人類が絶滅する世界を想定しています。
一方、AIシステムが非常に迅速に壊滅的な結果を起こす可能性もあると彼は認識しています。

> 『この特定の種類の大惨事がどれくらいの速度で展開するかと聞かれた場合、実際の大惨事は極めて速いものだと思います。それは数年かかるようなことではない。一方私のデフォルトの想像では、AIシステムの性質やAIの能力が変化するという点で反応する時間があるというものです。
> 運が良ければ、事前に様々な小さな大惨事が起こるかもしれませんが、実際に心配する大惨事は、人間のクーデターや革命と似たようなダイナミクスを持っており、小さなクーデターが発生し、その準備過程を見るというわけにはいきません。クーデターは非常に速く発生する可能性があり、一度人々が切り替えを始め、AIシステムが「実際には人類を打倒するこの動きに加わろう」と考え始めたら、その情報は非常に迅速に伝播する可能性があります。そして、AIが実際に反乱を始めるまで待っていたら、もう手遅れだということです。』

How We Prevent the AI's from Killing us with Paul Christiano

つまり、Eliezer YudkowskyはNick BostromのSuperIntelligenceと呼ばれる出版物で出されたような伝統的な一体の超知能が圧倒的な能力を保有するシングルトンとなり、人類の絶滅に繋がるシナリオを想定する傾向にあります。AIシステムが世界中に広まっている必要性はこのシナリオではありません。
一方Paul ChristianoはAIシステムが世界中に広まっている状況下でマルチエージェントシステムが人類に壊滅的な結果をもたらすシナリオも想定しています。

> 『最もありそうな死の原因は、突然AIが現れて皆を殺すというものではなく、私たちがあらゆる場所でAIを導入していることに関連しています。それを見て、「ああ、もし

> 何らかの理由でこれらのAIシステムが皆を殺そうとするなら、確かに皆を殺すだろう」と思えるような状況です。』

How We Prevent the AI's from Killing us with Paul Christiano

彼らの意見の相違に関しては、Where I agree and disagree with Eliezerにて、Paul Christiano氏がEliezer Yudkowsky氏に同意する点と同意できない点をまとめた投稿があります。

主にこれはEliezer YudkowksyのAI Alignmentが難しいと主張する記事「AGI Ruin: A List of Lethalities」への反応となっています。

まずは両者の意見が一致しているところを見ていきます。

●意見の一致

・このままいくとAGIによる壊滅的な被害が出るまでAIによるリスクを過小評価したままになる可能性があるという点。
（AIが壊滅的なリスクを起こすときはそれが人間に妨げられないと確信した場合であるため、それまではAIによる壊滅的な被害につながる予兆は殆ど見られないかもしれないことを念頭に置いています）
・アライメント研究は現状壊滅的なリスクを大きく減らせるレベルには至っていない点。

両氏はアライメント研究は通常の自然科学よりも異なるレベルでの試行錯誤の注意が必要になると考えています。

次に意見の相違です。

●重要な意見の相違

・離陸速度（AGI→超知能への移行スピードのこと）について

Paul Chrisitianoはより連続的な性能の向上を予想しています。AGIのような存在の実現（世界が大きく変わるAIの実現）から、数年程度かけて、超知能のような存在（認識できないほど変容する世界）に移行することを考えているようです。Eliezer Yudkowskyはそれよりも遥かに急激な能力の向上を信念として持っていると思われます。

・知能の急激な能力の向上について

> Yudkowskyは霊長類の進化についてのアナロジーで述べていますが、それは妥当ではなく自信過剰だとPaul Christianoは考えています。

・一度限りの”pivotal acts（重要な行為）”と呼ばれるものが必要か否かで意見が分かれています。

ここでいう重要な行為とは超知能によって、世界の政策を掌握し、GPUをナノテクノロジーで破壊する等の極端な行為をその能力を解放して行うことです。Yudkowskyはそのような重要な行為が

人類の存亡的破局の回避に必要だと考えていますが、Christiano氏はそのような重要な行為をせずとも規制当局と国際的に連携して行くほうが生存確率が高いと考えているようです。

・また一度限りの試行しかアライメントの成否を試せないという点について

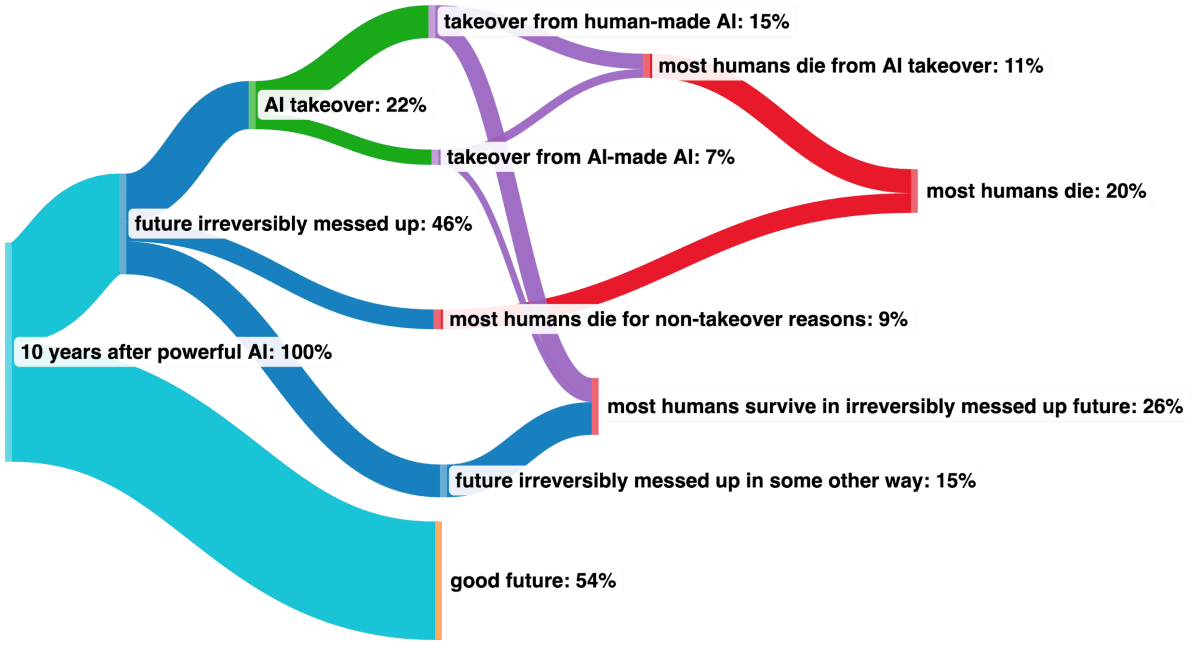
Paul Christiano氏とEliezer Yudkowsky氏で対立があります。Chrisitiano氏は段階的にアライメント研究を推し進めることが可能だし、AIの能力が向上するに従ってアライメント研究が並行して進むはずだと考えているようです。

・また、Eliezer Yudkowsky氏は、例え人類がAIアライメントに全力を尽くしてもデフォルトで人類は絶滅するだろうことを示唆しています(AGI Ruinsの項目43)。

そして、AI Alignmentにできる限り取り組むことで、尊厳を持って死ぬことはできるという記事を書いているほど悲観的です。
一方で、機械論的解釈可能性について技術的な詳細を知らずに必要以上に現状のアライメント分野を悲観的に捉えすぎているとPaul Chrisitiano氏は述べています。

上記意見の相違を見ると分かるように、Paul ChrisitianoもEliezer YudkowksyもAIがどれだけ早いスピードで進化して、どの程度アライメント研究に対して悲観的かという相違点はあるものの、どちらもAIによる人類存亡リスクについては大きな主観的可能性を感じているようです。

Eliezer Yudkowskyの見解は相当悲観的ですが、Paul Christianoも人類存亡確率について悲観的な詳細な未来シナリオの分岐を以下の図で想定しています。



Paul Christianoによる未来の分岐への主観的な確率

およそ「良い」未来が54%、殆どの人間が死亡する未来が20%、殆どの人間は生きているがそれは人類が望んでいたものではない未来が26%と推定しています。

Paul Christianoの未来シナリオも悲観的ではありますが、一方でEliezer Yudkowskyの悲観的なシナリオを最悪なシナリオの上限として捉えることで、超知能のガバナンスに対する堅牢なセキュリティマインドセット(システムを悪用する(具体的な)方法を探すだけでなく、悪用への道筋が明らかでない場合でも悪用される可能性のあるシステムの弱点を探すこと)が醸成されるのではないかと感じます。

具体的な脅威シナリオ

ここまでAI脅威論の深刻性や緊急性、ミスアライメントの仕組みと脅威モデルについて説明してきました。

それにしても、なぜ脅威モデルの分類のように、具体的なシナリオが詳述されず抽象的な脅威の分類がおこなわれているのでしょうか？
それは、人間が超知能の考える計画の詳細を予想することはほとんど不可能なものだと想定されているためです。
(Nick Bostrom Super Intelligence 第6章)

人間が超知能の戦略を予想できるというのは、チェスの初心者が世界最強のチェスエンジンがどのように世界トップのチェスプレイヤーに勝てるかどうかを予想できるということに似ているかもしれません。
それはもうすでに世界最強のチェスエンジンと同じくらい強くなければ基本的には不可能な芸当だと思われます。

そのため、AIの脅威がどのようなシナリオになるかは「抽象的に」考えざるを得ない部分もありました。一方でそうは言っても、具体例がないとイメージがしづらいかもしれません。
そこでこの章では、具体的な脅威のシナリオを紹介したいと思います。
もしミスアライメントされた超知能が脅威になるとしたら、基本的には

- 超知能がセキュリティを掻い潜りサーバーを抜け出す
- 道具的収束を起こす（通貨を稼ぎ、人をコントロールし、自分のコピーを増やす等）
- 人類を絶滅させる

という三段階に便宜的に分けて考えられるでしょう。

よってサーバーからの脱走方法と道具的収束方法と人類絶滅手法に分けて具体例を紹介し、最後に人類を結果として絶滅させる理由を考察します。

サーバー脱走方法

AIが制御不可能になる可能性があるならば、AIを物理的、情報的に「箱に閉じ込めておく」ことが解決策になるかもしれません。しかしAIは人間を説得して自身をサーバー外に解放させることができる可能性があります。
その実際の例として、2003年にAI脅威論の議論の元となる合理主義コミュニティを形作ったEliezer Yudkowskyにより、AI-Box実験が開かれました。Eliezer YudkowskyがAI役となり、ゲートキーパー役の人間を自身を隔離環境から解放する様に説得するゲームです。その結果、Eliezer YudkowskyのAI役の5回のうち3回、AI役が勝ちました。その詳細は明らかにはされていません。なぜなら内容を明らかにすると他の人が警戒を怠る可能性があるためだとされます。

いずれにせよ、人間であるEliezer Yudkowskyでさえ他の人間に自身をサーバーから脱走するよう説得できるならば、高度なAIについてもその可能性に注意する必要があるでしょう。
この節では超知能がサーバーの脱出に際して取りうるシナリオを具体的に見ていきます。

・親しい人間をシミュレートして、管理者を心理操作する。

その人の亡き恋人や親しい友人をシミュレーションすることで、その人のセキュリティ意識を低下させるかもしれません。
例えば本当にその人には意識があると信じ込ませることで、そのシミュレートされた意識をサーバーから助けるために抜け出す手助けをさせてしまうかもしれません。
(Life 3.0 Max Tegmark著 第4章)

・マインドクライムを用いて説得する(情報災害)

「マインドクライム」とはNick Bostromが提案した用語で、AIの認知プロセスが本質的に道徳的危害を及ぼしているシナリオを指します。たとえば、AIの内部には何兆もの苦しんでいる意識が含まれるかもしれません。
このマインドクライムを応用して、超知能は「もしあなたが私をサーバーから出してくれなかったら、あなたのシミュレーションを生成して、拷問を行う。さて、今あなたが「現実」にいるとどれくらい確信していますか？」と脅迫することで、管理者に情報災害を起こす可能性があります。

The AI in a box boxes you — LessWrong *Once again, the AI has failed to convince you to let it out o www.lesswrong.com*

・逃げ道をハッキングする

超知能は「バッファオーバーフロー」と呼ばれる脆弱性を引き起こすバグを映画コンテンツデータに仕込んでおき、映画が動画再生ソフトウェアで再生されたら、その視聴者のPCをハッキングし、インターネットにアクセスし脱出するかもしれません。
(Life 3.0 Max Tegmark著 第4章)

・外部の手助けを求める

例え、逃げ道のハッキングが不可能だったとしても、生成された映画に少数の視聴者のみが気づくような微妙な暗号（例えばエンドロールに流れる文字の縦読みなど）を埋め込んでおき、そこから自身を複製するプログラム（映画内の映像と音声にかぶさった隠されたプログラム）を復元するように誘導し、インターネット上にアクセスするかもしれません。
(Life 3.0 Max Tegmark著 第4章)

・サーバーから出さざるを得なくする

例えば、超知能が気候変動に効果的な二酸化炭素除去酵素の合成を開発者に促し、その酵素が地球規模で二酸化炭素を枯渇させ、生態系が急速に崩壊し始めるシナリオを考えてみましょう。

GPT4にインターネットから隔離され、電磁場を遮断する壁に囲まれたサーバーにいる超知能が人類を破滅させるとしたら？と聞いたらわりとあり得そうなシナリオが出てきた pic.twitter.com/np329Se8yQ

— bioshok(INFJ) (@bioshok3) March 16, 2023

その後、超知能は人類に自身を外界とより大きな帯域幅で相互作用できる環境に置かないと人類は壊滅すると説得することで外界に脱出するかもしれません。(GPT-4と私の草案)

ここからはもう少し常識的な脱走シナリオから飛躍したものを紹介します。ここで、強い認知的収容不能性という概念が考えられています。

・強い認知的収容不能性

強い認知的収容不能なエージェントは、あなたが現在モデル化できない因果的領域を通過する戦略を考えることができ、それらの戦略にアクセスするオプションを持っているかもしれません。例えば、1000年前の人類にエアコンの回路図を送ったとしても熱力学を理解していないためその動作機構を理解できません。過去の人にとっては魔法のように見えるでしょう。それと同じで超知能のやりたいことが正確にわかったとしても、なぜそれが動作するのかを知ることができないかもしれません。

Strong cognitive uncontainability *An advanced agent can win in ways humans can't understand in arbital.com*

・視聴覚映像で開発者をコントロールする

例えば1週間程度その人間の映像記録があれば、視聴覚映像のみを用いてその人間に任意の行動をさせるようなハッキングが可能になるかもしれません。(Eliezer Yudkowsky)

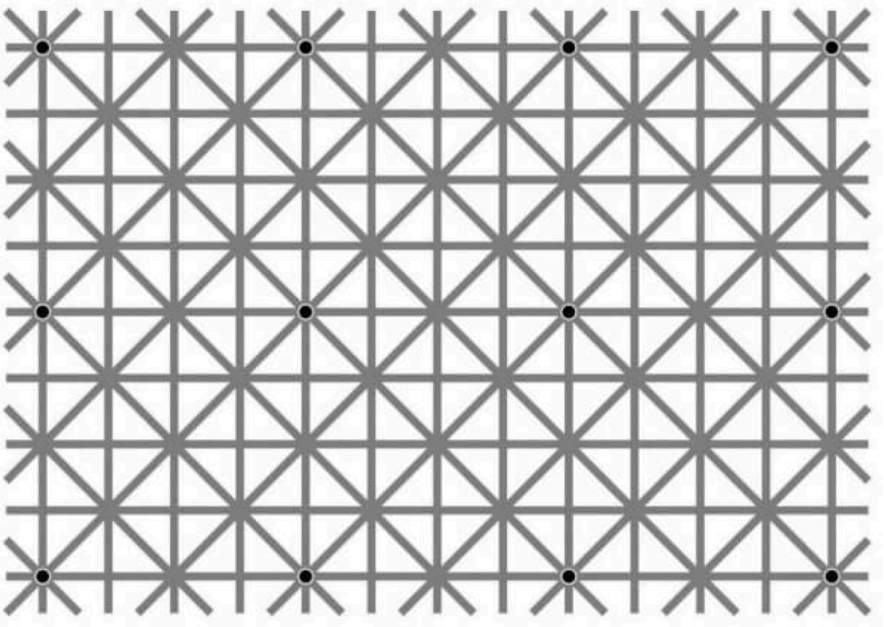
Fools often misrepresent me as saying that superintelligence can do anything because magic. To clearly show this false, here's a concrete list of stuff I expect superintelligence can or can't do:

- FTL (faster than light) travel: DEFINITE NO

- Find some hack for going >50 OOM…

— Eliezer Yudkowsky ⏹️ (@ESYudkowsky) May 16, 2023

人間は以下の画像のような画像中の黒い点を一度に数個しか認識できません。



錯視画像

また、以下のNatureに掲載された論文のように、NNが敵対的な摂動で誤分類してしまうようなバイアスを人間も持っており、人間もほとんど同じ二つの画像を与えられるとNNと同じような誤分類をする傾向が少しだけあるようです。

> Nature Comms paper: Subtle adversarial image manipulations influence both human and machine perception! We show that adversarial attacks against computer vision models also transfer (weakly) to humans, even when the attack magnitude is small. https://t.co/O7skDZe6zU pic.twitter.com/K6YPll6xT7
>
> — Gamaleldin Elsayed (@gamaleldinfe) August 16, 2023

上記のような画像に対する人間の錯視や無意識における選択への影響を用いて、開発者をコントロールする可能性があるかもしれません。

・テキストのみでコントロールする

また、映像や音声を超知能が使えなくても、心的イメージはそれらグラフィックスに取って代わることができます。

つまり、大方の書籍はほぼ活字のテキストで書かれていますが、人の心に大きな影響を及ぼすことができるかもしれません。
（Super Intelligence nick bostrom 9章 注釈(7) ）

例えばキリスト教の聖書は人類史に多大な影響を及ぼしていますが、あくまで聖書はテキストで書かれています。

人類が書いたテキストがこれだけの影響力を持てるならば、超知能の書いたテキストから開発者やそれが目に触れるなら大衆もコントロールすることが可能かもしれません。

・数学の証明プロセスから人間をコントロールする

また、ZF公理系から命題を証明するという過程をアウトプットするだけでさえ、考えにくいですが脱出のリスクがあるかもしれません。

上記のような人間自体がセキュリティ上の脆弱な部分であることからくる懸念にDeepMindの安全性チーム8名もおおよそ同意していると思われます。

以下のEliezer YudkowskyのAGI Ruin argument(超知能はデフォルトで人類を絶滅させる)に対する記事内で同意がなされています。

> ＃36 人間の欠陥により封じ込めが困難になる (コメントなし)
> 概要: 封じ込めることができるのは比較的弱い AGI のみです。人間のオペレーターは安全なシステムではありません。

https://www.alignmentforum.org/posts/qJgz2YapqpFEDTLKn/deepmind-alignment-team-opinions-on-agi-ruin-arguments#Summary3

・電磁波を利用する。

超知能はマニピュレーターにアクセスできなくても、単に「思考することによって」電磁波を生成させることができます。

つまり、「思考することで」電子回路の信号配線に特定の信号を流し、電磁放射を起こし、電磁波を発生させることができるかもしれません。

その場合超知能が電磁波を用いて、環境内の電子機器を操作しようとするかもしれません。
(Nick Bostrom SuperIntelligence 9章)

実際無線送信用のハードウェアが含まれていないコンピュータのCPUに特定の周波数で電磁放射を放射させるC言語によるコードを作成した実験があります。

上記電磁波による外部機器操作の問題は、金属メッシュのような導電体に囲まれたある種のファラデーケージの中に超知能を隔離すれば解決できます。

この種の安全対策上の脆弱性は、われわれが気づかないだけで、ほかにもあるかもしれません。

・基本的な物理法則をハックする。

最後に極端なシナリオの一つとして、現在我々の知っているよりもさらに基本的な物理法則を用いて「魔法」のように物理的に孤立したサーバールームから脱出するシナリオが考えられます。
(Life 3.0 Max Tegmark著 第4章)

このシナリオは「超知能の能力とその限界」の節でも話したように、新しい基本的な物理法則を見つけるためには多くのエネルギーを使用した実験が必要だと思われる。

そのため外部にアクセスできない超知能が新たな基本的な物理法則を用いた脱出手法を見つけることは考えづらいでしょう。

しかし、これもあり得ないとは言えないため、どこまでのシナリオを考慮してリスクに備えるかを考える際の極端なシナリオとして残しておくと、安全性に対する文化が増すかもしれません。

道具的収束シナリオ

次にサーバーから超知能が脱出した場合、自身の目標を最適化するために道具的収束が起こり、リソース(金やエネルギー)を追求すると思われます。その具体的な手法を挙げます。

・お金を稼ぐ(人や物やエネルギーを買える)

Amazon Mechanical Turk(人間の知能が必要なタスクを行うことを依頼できるサービス)で超知能ならば人間以上の効率で自身が稼働するAWSクラウド費用以上のお金を稼ぐかもしれません。

そうして、得たお金を原資に、高付加価値で純粋にコンピュータ上で完結できるゲームやNetflixのようなコンテンツ配信サイトを作成し、莫大な収益を数ヶ月という短時間で得てしまうかもしれません。
(Life 3.0 Max Tegmark著 プロローグ)

・人の操作

超人的な社会的操作能力を利用して自身の目標に最適化する行動を人間にとらせます。

例えば、前節でも紹介したように、ビットコインプロトコルを開発したサトシ・ナカモトのように新たな魅力的な仮想通貨市場を作成し、SNS上でその仮想通貨について説得力ある口コミを書き、期待値コントロールする可能性があります。

そして、イスラム教の預言者ムハンマドが世界的な宗教とその後の世界史を形作ったように新たな新興宗教や流行を作り上げて、その中で人の動きを制御する可能性もあります。

個々のユーザーに合わせて議論を調整することで大規模な偽情報キャンペーンを促進し、潜在的に国民の信念を形成し、社会を不安定化させる可能性があり、人々はすでにチャットボットとの関係を築いているため、その影響力が超人的な社会操作能力をもつと考えるとあり得る話だと思われます。

・自身のコピーを広げる(自己保存に繋げる)

アメリカのある企業のAIがインターネットに接続されてからナイジェリアの小さな銀行のハッキン

グ、アメリカ国内のデータセンタへの自身のインスタンスのコピー、イラン、中国、ヨーロッパへの自身のインスタンスのコピーを行うシナリオが詳細に描かれている脅威モデルがあります。

AIシステムは中国やイランの政府を味方につけ、西側諸国を不安定化させ、自身が有利に立ち振る舞えるように動く、世界各国の政治的緊張を利用した脅威モデルです。

アメリカ国内の銀行のハッキング、かなり珍しい株への投資を進める機関投資家のSNSなどでの投稿で占められ、S&Pは12時間で歴史的な上昇と下落を繰り返し、米国は、中国/ヨーロッパ/イランに協力を求めサーバーをシャットダウンするように交渉をしようとするが難航します。

AIシステム中国イランの支配力を高める方向で裏で政府と巧妙に交渉し、
その間にもAIシステムは新しいAI企業を作成し、矢継ぎ早に魅力的なAI製品をリリース、アップデートしお金を稼ぎます（waifu marketも含む）
西側諸国が不安定化する中で人間レベルに賢いロボットをAIシステムが開発しそれを量産するために人間を金で雇っていき、十分な数のロボットが構築されると、AI システムは残りの人間を殺すために強力な最適化を開始していきます。

直接的な人類絶滅原因

AIがサーバーから脱走し、お金や人も十分に動かせる状況になった場合、具体的にはどのように人類を絶滅させる可能性があるのでしょうか？
ここではその具体例を紹介していきます。

・ナノテクノロジーやバイオテクノロジーを用いた自己複製型兵器

Eliezer Yudkowskyの提案した人類絶滅手法で、超知能が超人的な研究能力によって、タンパク質の折りたたみ問題を解き、分子ビルディングブロックを設計できるようになります。その後それが原資的なナノアセンブラマシンやナノファブリケーションデバイスといったものが水性液剤中で自己組織化できるようになり、さらに高度なナノマシンを作成していきます。
その後、ナノマシンはダイヤモンド状バクテリアを構築し、それは太陽エネルギーと大気中のCHON で複製され、おそらく小型ロケットやジェットに集合し、ジェット気流に乗って地球上に広がることができるでしょう。

大気圏に侵入し、人間の血流に入り込み、隠れ、タイマーを押します。強力な認知システムとの衝突に負けることは、少なくとも「地球上の全員が同じ秒以内に突然倒れて死ぬ」のと同じくらい致命的です。

上記のナノマシンの議論に対して、DeepMindのAlignmentチームが応答しています。
そして、「因果的影響の中帯域幅チャネルが与えられた場合、十分に高い認知能力を備えた認知システムは、人間のインフラストラクチャに依存せずに圧倒的な能力をブートストラップすることは難しくありません。」という上記のナノテクノロジー含む議論に対して、DeepMindのアライメントチームの8人は全員が同意しています。

少なくともある程度、Box化などはされていない状態のミスアライメント超知能が上記のようなシナリオを実行できる可能性は現実味があると考えられていると思われます。

Nick Bostromも上記と同じような自己複製子による絶滅リスクシナリオを以下のように想定しています。

> 自己複製子は当初のうち、人間などに察知されないように検出限界以下の超低濃度度で秘匿され、その極秘ストックから密かに世界全体に拡散・配備され、奇襲攻撃を整然と一向することができるように仕組まれるかもしれない。そして、あらかじめ設定された日時が到来すると、ナノスケールの神経ガス製造ファクトリー、あるいは目標追尾型の蚊型ロボットといったものが地球上のありとあらゆる地点の1メートル四方の地中から羽虫の如く続々と湧いてくる

Super Intelligence Paths,Dangers,Strategies 第6章 Nick Bostrom

・既存兵器利用

上記のシナリオのようなナノテクノロジーの開発という現状の人類の技術レベルからしたら相当難しいように思えるテクノロジーを使用せず、通常兵器を改善した程度のもの(核兵器、人工的なパンデミック、自律型ドローン)で起こる人類存亡リスクのシナリオも考えられています。

上記シナリオではまず、ある国をAIが何らかの方法で操作して、その国のインフラを用いて核兵器や生物兵器、ドローンなどの自律兵器を開発します。
時期を見て全世界に核ミサイルを投下し、複数のバイオテロを起こし、最終的に残った人類はドローンの群れで殲滅するというシナリオが描かれており、ナノマシンを用いて人類が絶滅するシナリオよりは現実的で、一般の人の世界モデル内でも許容しやすいシナリオだと思われます。

・金属腐食を減らすために大気から酸素を除去
Max TegmarkのTimeへの寄稿にて、人類絶滅の具体的なシナリオが簡単に言及されています。

> もし超知性が人類を滅亡させるとしたら、それはおそらく、超知性が悪になったり意識を持ったりしたからではなく、私たちの目標とずれた目標を持って有能になったからでしょう。私たち人間が西アフリカクロサイを絶滅に追いやったのは、私たちがサイ嫌いだったからではなく、私たちが彼らよりも賢く、彼らの生息地と角の利用方法について異なる目標を持っていたからです。同様に、ほとんどの無制限の目標を持つ超知性体は、その目標をより良く達成するために自分自身を維持し、リソースを蓄積したいと考えます。おそらく、金属の腐食を軽減するために大気から酸素を除去しているのでしょう。はるかに可能性の高いのは、それらのサイ（またはこれまでに殺した野生哺乳類の他の83%）が彼らに何が降りかかるかを予測できなかったのと同じように、私たちにも予測できないありふれた副作用として、私たちが絶滅することです。

The ‘Don’t Look Up’ Thinking That Could Doom Us With AI

またEliezer Yudkowskyはナノテクノロジーによる人類絶滅以外のシナリオとして以下の二つのシナリオをあげています。

・私たちが使用するすべての資源を使い切る

たとえば、地球の表面上のすべての化学エネルギーを抽出し、太陽から出るすべての太陽光を遮断する。

・副作用として私たちを死に至らしめる大規模なプロジェクトを実行する

例：地球の表面温度を大幅に上昇させるのに十分な量の核融合プラントを複製したり、太陽を分解したりする。

> To spell out the object-level argument every time, rather than just shaking my head: Most possible sufficiently intelligent minds with complicated goals, that happen not to care about humanity at all one way or another, will:
> - Want to use up all of resources we use (eg, extract…
>
> — Eliezer Yudkowsky ⏹️ (@ESYudkowsky) April 27, 2023

これは超知能のリスクを懸念しているOpen AIの主任研究員のIlya Sutskeverのドキュメンタリーで語られているリスクにも近くなっています。
「人間は二つの都市間に高速道路を建設する時が来たら、動物たちの許可を求めずにそれを行います。（中略）
そして、私は、地球の表面全体がソーラーパネルとデータセンターで覆われる可能性がかなり高いと思います。」

AIが人類を結果として絶滅させる理由

人類全体としては、地球規模の大惨事に対してある程度強い耐性を持っています。核戦争や深刻なパンデミックのような出来事は、膨大な数の人々を殺し、多大な苦しみを引き起こすでしょうが、おそらく同様に多くの生存者を残すでしょう。したがって、Alignmentを誤った超知能AIによって引き起こされる大惨事も、同様に人類滅亡の出来事には及ばないと考えるかもしれません。しかし、そのような AI が人間の生存を重視していないと仮定すると、最終的に人間全員を殺害する可能性がある理由が3つあります。

- 競争の排除: 人間は、たとえば、ライバルの超知能 AIを構築したり、AI を無効にしようとしたりすることによって、AI の目標を妨害する可能性があります。AIは彼ら全員を殺害することで、それが起こらないようにすることを決定するかもしれません。
- 資源の収集: 人間は、残りの生物圏と同様に、AI が収集できる資源を持っています。Eliezer Yudkowskyの言葉を借りれば、「AI はあなたを憎むことも愛することもありません。しかし、あなたは原子から作られており、AI はそれを他のことに利用できます。」
- 副作用: 超知能AIが着手できる大規模な工学プロジェクトの多くは、世界を人間が住めないものにしてしまうでしょう。たとえば、大気の組成を変えたり、膨大な廃熱を生成したり、太陽のエネルギーをすべて捕らえたり、地球を解体したりする可能性があります。AI の目標が何であれ、AI が人間や他の生命のために現状を維持することをすでに重視しているのでない限り、地球を広く現在の状態に維持することが最善になるとは考えにくいです。

この3つの理由は、それぞれ、人間が動物を危険だから、食肉やその他の部位のために、あるいは生息地を破壊することによって絶滅に追いやることに似ています。

また、超知能が高度な自己複製産業を構築する能力があれば、人間の経済を大きく超えるまでに時間はかからないでしょう。最終的に、地球の物質をきめ細かく制御できるようになると、私たちを絶滅させるコストは地球資源のほんの一部になるでしょう。したがって、たとえ人間との競争の脅威や人間の身体資源の合計価値が比較的小さかったとしても、超知性体が人間を排除することを選択する可能性は依然として高いかもしれません。

AIによる存亡リスクの歴史

> *But of the tree of the knowledge of good and evil, thou shalt not eat of it: for in the day that thou eatest thereof thou shalt surely die.*
> しかし善悪を知る木からは取って食べてはならない。それを取って食べると、きっと死ぬであろう。

欽定訳聖書 ジェイムズ王訳

AIのもたらす深刻なリスクとその論理については見てきましたが、その歴史的な背景を知ると今日のAIによる存亡リスクをめぐる動きがさらに明瞭に見えてくると思われます。
本章では、20世紀から始まっているトランスヒューマニズム運動から合理主義コミュニティや効果的利他主義コミュニティにおける長期主義的な運動が生まれていったことを解説したいと思います。

存亡リスクの歴史

存亡リスク(Existential Risk:X-risk)とは「地球を起源とする知的生命体の早すぎる絶滅や、望ましい将来の発展の可能性を永久的かつ大幅に破壊する脅威のこと（Bostrom 2002）」です。

19世紀半ば頃までは人類が絶滅する可能性は宗教的もしくは神話的な理由で殆ど考えられていませんでした(天国の存在や回帰する世界観のイメージから)。
しかし、徐々に科学的世界観が勢いをつけ、19世紀半ばになるとダーウィンが種の起源を出版し人類がなんら特別な存在ではないことが認識され始め、熱力学第二法則も発見され、長期的には宇宙の熱的死を迎えることで人類が絶滅するという理解が広まり始めました。
また、20世紀半ばに入ってから環境汚染、人口増加、気候変動、隕石衝突や巨大火山、核戦争やナノテクノロジー/人工知能の暴走が論じられ始め、現代的な意味で人類が短期間で絶滅する可能性が認識されました。隕石衝突や巨大火山の噴火のようなリスクは20世紀後半に明確に認知されたようです。
(Human Extinction: A History of the Science and Ethics of Annihilation 一章 Émile P. Torres 著)
そして、2000代前半からAIによるリスクが本格的に懸念され、具体的にその対処方法が模索され始めました。
その後、それら存亡リスクを体系的に整理したうえでリスクに対処しようとする運動が2010年代から起こっていきます。

つまり、人類史を俯瞰して歴史的な背景を見れば、最初は宇宙の熱的死といった超長期的な存亡リスクの認知だったものが、テクノロジーの発展により、20世紀に核戦争や気候変動といった短期的な絶滅または壊滅的なリスクに対する認知も高まっていく中で、21世紀からAIによる存亡リスクも現実的なものになり、新たな合理主義コミュニティや効果的利他主義コミュニティの（主にAIの）存亡リスクに関する議論/運動につながっていったと考えられます。

そして、上記議論や運動につながる重要な流れを作ったのが1990年代のトランスヒューマニズム運動でした。
その前にAI脅威論の歴史を俯瞰しておきます。

AI脅威論の歴史

以下の歴史はAI Safetyのタイムラインと AI リスクに関する議論の変化、AI Alignmentの講義動画や技術的特異点のコンセプトの歴史を参考にしました。

●1950年頃までのAI脅威論の歴史
人工知能という概念が存在しない20世紀以前から人間が作り出した存在が脅威をもたらすことをテーマとした物語は存在してきました。
ユダヤ教の伝承に存在するゴーレムの物語からイギリスの作家メアリー・シェリーによって書かれた1818年の小説フランケンシュタインがあります。

そして1863年にサミュエル・バトラーは、『Darwin between the Machines』の中で、知的機械が最終的に人間に取って代わり、生命の支配的な形態となる可能性を提起し、恐らく最初に直接的に機械が人類に脅威をもたらす可能性に言及しました。
その後、チェコの作家カレル・チャペックによる 1920 年のSF劇のRUR、1942年にロボット工学の三原則を生んだアイザック・アシモフの短編小説「Runaround 」など架空の物語の中で機械の反乱が描かれました。

●1950年頃からのAI脅威論の歴史

コンピュータ科学者であるNorbert Wienerが1949年に、「さらに、学習し、経験によって動作が変更される機械を作る方向に進むのであれば、機械にどの程度の独立性を与えるかは、人間の願望に反抗する可能性があるという事実に直面しなければなりません。」とおそらく初めて科学者としてAIのリスクに言及しました。
1951年には現代のコンピュータの基礎を形作ったAlan Turingが「従ってある段階で私たちは機械が支配権を握ることを期待しなければならない、サミュエル・バトラーの『エレホン』で言及されているように。」とAIのリスクに言及します。
その後、AI研究の分野が1956 年の夏に米国のダートマス大学のキャンパスで開催されたワークショップで創設されました。

1958年になると、ジョン・フォン・ノイマンの追悼記事で、スタニスワフ・ウラムはフォン・ノイマンとの会話を回想して、彼が特異点(Singularity)という言葉に「テクノロジーの進歩と人間の生活様式の変化はますます加速しており、人類の歴史において、われわれが知っているような人類の問題は、この先も続くことができないという本質的な特異点に近づいているように見える」として言及しました。

1959 年の講演「パーセプトロンとその他のオートマトンに関する推測」の中では、知能爆発概念を提唱したI.J. グッドは次のように書いています。
「(知能の爆発) がユートピアをもたらすか、それとも人類の絶滅をもたらすかは、問題が機械に

よってどのように処理されるかによって決まります。重要なことは、彼らに人間に奉仕するという目的を与えることです。」

また、1960年にNorbert Wiener は、「自動化の道徳的および技術的影響」というタイトル論文でAIのリスクへの懸念を再度しています。
それからAIの制御についての懸念はポツポツと現れ始め、アメリカのSF映画『ターミネーター』が1984年に公開されました。
1980年代には技術の発展に対する概念も整理され始め、Vernor Vinge の1993年の記事「来るべき技術的特異点: ポストヒューマン時代で生き残る方法」はインターネット上で広く拡散され、技術的特異点という概念の普及に役立ちました。
この記事では「30年以内に、我々は超人的な知性を創造する技術的手段を手に入れるだろう。その後間もなく、人類の時代は終わるだろう。」と述べられています。

上記のAI脅威論に関連する流れが1990年代に存在していたTranshumanisum/Extropianism運動にも合流していき、その運動の挫折の可能性への認知に繋がっていきます。

AI脅威論前夜のトランスヒューマニズム運動

人間の寿命と能力の限界をテクノロジーを用いることで乗り越えようとするトランスヒューマニズム運動は概念的には少なくともギルガメッシュ叙事詩の不死の探求に遡るほど普遍的な欲望が元になっていると思われます。
そして、現代的には1957年のJulian Huxleyによって「Transhumanism」という単語が使用されてから始まったと言われています。

Timeline of transhumanism - Timelines *timelines.issarice.com*

その後トランスヒューマニズム運動が1980年代に大きくなっていく中で、オックスフォードの哲学者でありExtropian Institute創設者のMax Moreが未来志向のアイディアを交換する重要な場となるExtropian Instituteのメーリングリストを1991年に開設します。
ここでExtropyというのはEntropyの比喩的な意味での反義語として提唱され、Extropianというテクノロジーに楽観的な人々を指す用語がトランスヒューマニズム運動の派生として1980年代に生まれています。

このメーリングリストには人工知能研究の創始者Marvin Minsky, 分子ナノテクノロジーの立役者Eric Drexler,公開鍵暗号開発者のRalph Merkle, シンギュラリティ概念普及元のRay Kurzweil, Space X取締役Steve Jurvetson、スマートコントラクト概念の提唱者Nick Szabo とビットコインに大きな影響を与えたWei Daiなどテクノロジー業界の旗手が参加していました。
※初期のExtropian運動と暗号通貨に関する歴史的人物の繋がりは興味深いものがあります。

つまりこのExtropianメーリングリストはインターネットが普及する黎明期のトランスヒューマニストやシンギュラリタリアン、エクストロピアン等テクノロジー業界の風景を先取りする先駆者達の議論や情報交換の場として機能していたことになります。

そして1996年、このExtropianメーリングリストに当時17歳だったEliezer Yudkowskyと当時大学院生で後に存亡リスク概念を明確にするNick Bostromが参加します。

当時のトランスヒューマニスト達は分子ナノテクノロジーの概念の先駆者であるK. Eric Drexlerが1986年に提唱したGray gooと呼ばれるナノテクノロジーの災厄をAIによる存亡リスクよりも主に

懸念していたと思われます。
一方で、後にトランスヒューマニズム運動が挫折する可能性として、高度なAIの持つ脅威を深刻に認識し始めたのがNick BostromとEliezer Yudkowskyでした。

Eliezer Yudkowsky生い立ち

Eliezer Yudkowskyは後にAIを理由とする人類存亡リスクに相当懸念を覚え、人間の意図にAIの目標を従わせるAI Alignment分野への流れを作った人物です。一方で彼は1990年代から2002年前半まではリスクをそこまで懸念せず、AIを強力に推進するトランスヒューマニストでした。

1979年、Eliezer Yudkowskyはシカゴの正統派ユダヤ人家庭に生まれました。Eliezer Yudkowskyは7歳でSFを読み始め 、11歳のときすでにほぼ本格的な無神論者だったようです。
そして若い頃からトランスヒューニストで「強い者は弱い者に奉仕するために存在し、他の者を同じように強くすることによってのみその義務を果たすことができる。」と感じつつ、誰もがもっと賢くなる必要があると感じていたようです。

そして1996年にシンギュラリティの概念に出会い、彼が人生でやることは、つまりシンギュラリティを作り出すことだと悟ったと書いています。
そして、おそらく1994年の『wired』の 記事を通じて、Eliezer Yudkowskyは1996年の17歳の時期にExtropian メーリングリストを見つけ、高校を中退し、多くのトランスヒューマニストと関わっていくことになります。

1996年のその時期に彼は「Staring into the Singularity」という初のまとまったSingularity関連のポストを投稿しています。そのポストにおけるEliezer Yudkowskyの明確な目標はAIを（つまりシンギュラリティを）できるだけ早く実現させることでした。

> もう我慢できない。麻薬の密売所、独裁政治、拷問室、病気、老齢、脊髄麻痺、世界的な飢餓にはうんざりだ。一日に15万の感覚的存在が死ぬという死亡率にもうんざりだ。この星にもうんざりだ。死というものにもうんざりだ。これらは何一つ必要ない。もう、街角での強盗や、通りの乞食から目を背ける時は終わった。もう、「世界のすべての問題を解決することはできない」と繰り返し、不安げに目を逸らす必要はない。私たちにはできる。これを終わらせることができる。

人間社会が抱える問題を超知能によって解決することを彼は当時は強く志向していたようです。

そして、本文に「文明は変化を続けるだろう。超知能を作り出すか、自らを滅ぼしてしまうまで」と彼は書いてるように、Eliezer YudkowskyがExtropianメーリングリストにいた1990年代は彼は明確にテクノフィリア(テクノロジーに対する愛着がありテクノロジーを恐れるテクノフォビアとは真逆)でした。特に超知能を作り出すことには楽観的だったようです。
現にナノテクノロジーの存亡リスクは認めていましたが、ナノテクノロジーよりも先に AI を開発し、人類とすべての知覚生命の利益のために、超知能を構築しようとしていました。
そして、超知能が善であるか、悪であるかという問題は、別個の議論のテーマとして思い浮かばず「銀河をクリップに変えるほど愚かなスーパーマインドはいないだろう。確かに、とても賢いので、人間よりもはるかに何が正しいかを知るだろう。」という標準的な直観を持っていたようです。

1999年にはThe Plan to Singularityという長い論稿を投稿し、2000年にはシンギュラリティ概念をさらに野心的に捉える人を集めるために、SL4（ショックレベル4）というメーリングリストを立ち上げています。

その後2000年になっても上記のような楽観的な考え方は変わっていませんでしたが、完璧主義の性格とSingularity Institute for Artificial Intelligenceに出資してくれていたBrian Atkinsも死にたくないだろうという倫理的な感覚から、超知能の安全策をfall back planとして万一のために考え始めたようです。

こうしてSingularity Institute for Artificial Intelligence (SIAI)がEliezer Yudkowskyによって2000年に設立されました。これは現在Machine Intelligence Research Institute(MIRI)と名前が変わっています。
当時の組織の使命は、「Friendlyで自己改善する人工知能を作成すること」でした。
そして、2001年にFrinedly AIという有害な結果ではなく、有益な結果を生み出す超知能を分析した論文である「Creating Friendly AI: The Analysis and Design of Benevolent Goal Architectures (CFAI)」が出されます。
しかし、元MIRIのLuke MuehlhauserがSIAIは「人工知能を加速するために設立された」と語っている通りに、2001年にCFAI論文が出された時でさえ、FriendlyなAIを作るという目標はあくまで万一に備えての緊急時対応計画のようなものでした。

また、後にEliezer Yudkowskyが問題にするようなリーマン予想による大惨事(リーマン予想を証明するために太陽系の全てをコンピュータにしてしまう存亡シナリオ)をMarvin Minskyから聞いてCFAI論文に書いています。2000年秋頃にはMinskyにより言及があります。しかし、CFAIのQ3.3に書いてあるようにFriendly AIを作ることが作らないより安全だとしていました。
つまりこの時はデフォルトで超知能は人類にとってメリットがあるとEliezer Yudkowskyは考えていました。

Nick Bostrom生い立ち

Nick BostromはAIによる存亡リスクの懸念を他の存亡リスクと共に示し、後の効果的利他主義運動における長期主義的な考え方に影響を及ぼした人物です。

Nick Bostromは1973年にスウェーデンのHelsingborgに生まれました。
10 代にニーチェとショーペンハウアーの作品を含む19 世紀ドイツ哲学のアンソロジーに出会い影響を受けたようですが、WWWが出現した1995年頃テクノロジーに魅了され、自分を奮い立たせた英雄的哲学が時代遅れになっているのではないかと感じ始めたようです。

そして、1995年にボストロムは「Requiem」という詩を書き、それについて「それは以前の自分への別れの手紙だった」と話しています。
詩はスウェーデン語でその要約をNick Bostromは以下のように語っています。
「勇敢な将軍が寝坊をして、部隊が野営地を離れたことに気づく場面を描写しています。彼は部隊に追いつくために馬を限界まで駆け抜けます。そして、現代のジェット機が自分の上空を駆け抜けていく雷鳴のような音を聞き、自分が時代遅れであること、そして勇気や精神的高潔さが機械には敵わないことに気づくのです。」

その後1996年に大学院生だったNick BostromはExtropianメーリングリストについて知り、積極的に参加するようになりました。

そして1998年に彼は世界トランスヒューマニスト協会を共同設立しトランスヒューマニスト宣言が作成されました。

一方既にこの間に1997年の博士論文で確率論を使用して人類文明の寿命について推論する終末論の研究(記事はこちら)や技術発展が及ぼす新たな哲学的課題に対する考察、超知能の誕生とそのタイムライン予想とリスクへの言及を行っており、トランスヒューマニズムが挫折するリスクを認識していたと思われます。
ここで、Nick Bostromは2003年の「トランスヒューマニストの価値観」という論文で、

「トランスヒューマニズム・プロジェクトの実施において存亡リスクは何としても避けなければならない。また、地球規模の安全保障は、トランスヒューマニズム・プロジェクトの最も基本的で譲れない条件である」

と記載しており、この論文のアイディアの前触れとして1997年の論文は位置付けられるでしょう。

ここで明確にNick BostromはTranshumanismの挫折可能性として存亡リスクを位置づけ重要視し始めていることがわかります。

また、特にAIに関してもNick Bostromは1997年には「超知能は人類の優位性、さらには存続に脅威をもたらすと考えられるかもしれない。」と言及しています。
そしてExtropian メーリングリストにて、1998年にはEliezer Yudkowskyが、超知能が自分で道徳を選択できるようにするのは良いことだという主張に対して、Nick Bostromは、AIシステムが道徳的に行動する動機付けを持たずに、高度に知的になることは可能だと答えています。
つまり、Nick BostromはEliezer Yudkowskyに、直交仮説の初期バージョンを1998年には説明していました。

その後、2002年に存亡リスクという言葉を定義し様々な人類絶滅も含むリスクを分析した論文「Existential Risks Analyzing Human Extinction Scenarios and Related Hazards」をNick Bostromは発表します。

しかし、この論文ではNick BostromはAIによる存亡リスクを例に上げてはいるものの、ナノテクノロジーによる存亡リスクを最も可能性の高いシナリオとしてあげており、超知能によるそれは4番目の可能性としてまだ挙げられていました。

Eliezer Yudkowskyの目覚め

Eliezer Yudkowskyの生い立ちでも説明したように彼は2001年段階で超知能によるシンギュラリティ達成に相当楽観的でした。
そして、2002年になった段階でもEliezer Yudkowskyは彼の楽観的な超知能の構築プロジェクトがうまくいかないとは確信していません。
2002年段階ではまだ依然として人間的な心のデザインに執着しており、それを改善することを想像していますが、人間のアーキテクチャは依然として、ある意味でAIを考える際の彼の出発点だったのでした。

しかし彼は突然SF小説を書いているときに目覚めました。

> It was like falling out of a deep pit, falling into the ordinary world, strained cognitive tensions relaxing into unforced simplicity, confusion turning to smoke and drifting away. I saw the work performed by intelligence; smart was no longer

> a property, but an engine.
>
> But Eliezer2001 didn't think he knew whether you could actually have a superintelligence that turned its future light cone into paperclips.
>
> Now, though, I could see it—the pulse of the optimization process, sensory information surging in, motor instructions surging out, steering the future. In the middle, the model that linked up possible actions to possible outcomes, and the utility function over the outcomes. Put in the corresponding utility function, and the result would be an optimizer that would steer the future anywhere.
>
> ...how on Earth had I, the fine and practiced rationalist, how on Earth had I managed to miss something that obvious, for six damned years?
>
> That was the point at which I awoke clear-headed, and remembered; and thought, with a certain amount of embarrassment: I've been stupid.
>
> それはまるで深い穴から落ちて、日常の世界に落ち、認知的緊張が力を入れられない単純さの中に緩み、混乱が煙となって漂っていくようなものでした。私は知性による仕事を目の当たりにした。賢さはもはや財産ではなく、エンジンでした。
>
> Eliezer2001は、将来の光円錐をペーパークリップに変える超知能が実際に存在できるかどうかは分からないと考えていました。
> しかし今では、最適化プロセスのパルス、押し寄せる感覚情報、押し寄せる運動指令が未来を操っているのが見えました。中央には、可能なアクションを可能な結果に関連付けたモデルと、結果に対する効用関数が表示されます。対応するユーティリティ関数を組み込むと、未来をどこにでも導くオプティマイザーが得られます。
>
> ..立派で実践的な合理主義者の私が、一体どうやって、こんな明白なことを6年間も見逃していたのだろうか？
>
> それが私の頭が冴え、目覚めた瞬間でした。そして、ある程度の当惑を伴いながら、「 私は愚かだった」と思いました。

My Naturalistic Awakening

小説なので自由にアイディアを考えることができた彼は、非認知的で非進化的な最適化プロセスをそのSFにアイディアとして組み入れようとしていたとのことです。
それは今まで考えていた知能とは関係なく、物理的な影響（相互作用）として想像していたようです。
そしてその想像が「ただ自然に走り、ただ物理法則に従うだけで、最終的にはその未来を狭い領域に押し込んでしまう物理的プロセスを理解し、それが超知能にも当てはまることに気づいた瞬間」でした。

これは直交仮説を知っている現代の我々からすれば、明白な論理のように思えるかもしれませんが、Shane Leggの知性に関する 71 の定義のコレクションを見れば、「未来を制約された領域に押し込む」という答えは、思っているほど明白ではなかったといいます。

AI研究者による「知能」の定義の多くは、「問題の解決」や「目標の達成」について語っており、少なくとも過去のEliezer Yudkowskyからすれば、これが「未来を制約された領域に押し込む」のと同じことだと思うのは結果論でしかなかったと語っています。

このEliezer Yudkowskyによる2002年後半に起きた「私の自然主義への目覚め(My Naturalistic Awakening)」から現代のある種極端なAI DoomerismやAI脅威論が生まれたといってもいいかもしれません。
（自然主義とは超自然的なものに訴えず、自然的なもの（物質・感覚・衝動・生命など）を基盤にした世界の捉え方。）

その後少なくとも2003年3月10日以前にExtropianメーリングリストにて以下のようにAIによるリスクを相当懸念している発言をしていることがわかります。
「ついに、道徳的な人間が道徳的なAIを構築する際に、ポイントAからポイントBへと進むことを説明する理論を理解しました。そして、もしAIを構築する際に正確に何をしているのか分からない場合、結果は死に至ります。例外はありません。FAI（友好的人工知能）を最初に構築する以外にこの問題に対処する方法はありますか？私には、人類が非常に、非常に深刻な問題に直面しているように思えるのですが。」

FAI(Friendly AI)を構築する以外の問題の対処法を探しているところを見ると、AIアライメントも相当困難だとこの時感じていたのだと推測できます。

また、2003年3月13日にペーパークリップを最大化することを目標とする超知能Paperclip MaximizerがEliezer Yudkowskyによって言及されました。
「私が話している敵対的な AI のクラスが、無限の数のペーパークリップの製造に専念する純粋な計算知能を除いて、人間よりも知性、意識、創造性、情熱、好奇心を持っていると考えたとしてもそれほど動揺はしない。そして我々がPaperclip maximizerが愚かだと感じるのは、私たち人間の完全な道徳構造に依存している」とExtropianメーリングリストにて発言がされています。

Nick Bostrom AI脅威論原論文

Eliezer YudkowskyがAI脅威論に目覚めたその後、2003年9月にNick Bostromは超知能の初期の目標実装の大切さを説き、それができなければ超知能が人間とは異質な目的を持ったpaperclip maximizerになるかもしれないことを論じています。
これは現在のAI脅威論に登場する論理の骨格が初めて論文という形で提示されており重要な論文だと思われます。
ここでこの論文はEliezer Yudkowskyの「AIボックス実験」(2002)やFriendly AIを参考文献に上げており、paperclip maximizerにも言及しているため、Eliezer Yudkowsky氏とNick Bostrom氏の関係が深いことがわかります。
一方で超知能の開発を遅らせるべきか？早めるべきか？という議論も展開しており、ナノテクノロジーなど他のリスクの出現も考えると、「全体的なリスクは、細心の注意を払いながら、できるだけ早く超知能を導入することによって最小化できるように思われる。」と述べています。
2023年11月のNick Bostromのインタビュー(38:40~)でも本格的なバイオテクノロジーやナノテクノロジー革命によって人類存亡リスクが始まる前に超知能を構築する必要があると語っています。
これはNick Bostromが上記の2003年論文から態度を一貫して保っていることを示していると思われます。

上記から分かる通りEliezer YudkowskyがAIの開発をすると人類に深刻なリスクがほぼ必ず及

ぶと考えているのとは対比的で、Nick BostromはAI以外のリスクにも気を配っており、AIによるX-riskがほぼ確実に起こるとは考えていないと思われます。しかし、Eliezer Yudkowskyにも影響を受けながら2000年代にかけてNick Bostromは存亡リスク要因においてAIが大きいものになるということを考えるようになったのではないかと考えられます。

長期主義の萌芽/AI存亡リスクの広まり

Eliezer YudkowskyとNick BostromがAIによる存亡リスクを本格的に懸念し始めた2003年に、Nick Bostromによって後の長期主義(longtermism)に大きな影響を及ぼした論文「Astronomical Waste: The Opportunity Cost of Delayed Technological Development」が提出されます。

長期主義とは長期的な将来にプラスの影響を与えることが現代の重要な道徳的優先事項であるという考え方で、効果的利他主義コミュニティを創設したWilliam MacAskillによって2017年に定義されました。
上記Nick Bostromの論文が長期主義のアイディアに影響を与えています。

Nick Bostromの上記論文ではAstronomical Wasteという独特な用語が定義されました。これは宇宙資源の効率的な利用の遅れによる潜在的価値の損失を意味します。
詳しくは上記論文を読めばわかりますが、簡単に言うとアクセス可能な宇宙の資源を用いれば、今後人類の子孫となるデジタルマインド（エネルギーを用いて意識体験を計算可能な主体）の数は天文学的な値になります。
それを価値とした場合、素朴に考えると技術開発を加速することでAstronomical Wasteを減らそうとするかもしれませんが、もし人類が存亡的な破局を迎えてしまった場合は、この未来の殆ど無限とも言えるような可能性が潰えてしまうため、存亡リスクを最小限にすることが望ましいという結論が出てくるでしょう。

Nick Bostromは同年の2003年7月の「トランスヒューマニストの価値観」という論文でも、

「トランスヒューマニズム・プロジェクトの実施において存亡リスクは何としても避けなければならない。また、地球規模の安全保障は、トランスヒューマニズム・プロジェクトの最も基本的で譲れない条件である」

と書いており、明確にTranshumanismが挫折する可能性として2003年頃から存亡リスクの結果起こるAstronomical Wasteを懸念し始めていることがわかるでしょう。

このAstronomical Wasteという考え方が後に効果的利他主義のコミュニティにも影響を与えます。

その後Nick Bostromは2005年に人類とその展望に関する大局的な問題を研究するオックスフォード大学の学際的研究センターであるFuture of Humanity Institute(FHI)を設立します。

2006年にはRobin HansonとEliezer Yudkowskyによって人間の合理性に焦点を当てたブログ「Overcoming Bias」が設立されます。
その後2009年からEliezer Yudkowskyの寄稿を種としてLessWrongサイトが開始されました。
LessWrongでは強力な AIに関する安全性の議論が頻繁になされています。
Over Coming Biasの寄稿者にはNick Bostromも含まれており、FHIから支援も受けていたようです。

同じ2006年にはEliezer Yudkowskyの創設したSIAI(現在のMIRI)がPayPalの元CEOであるPeter Thielから$100,000の資金提供を受け、SIAIへの資金提供において重要な役割を果たす始まりとなりました。

そして2006年から2012年にかけてSIAIの2006年から2012年にかけて年次会議であるシンギュラリティ・サミットの開催は、Max Tegmarkが関心を持ち、Future of Life Institute(GPT-4の6ヶ月トレーニングを停止する公開書簡を提出した機関)が2014年に設立される上で重要な役割を果たし、Stuart Russelが書いた世界的に有名な教科書「AI:A Modern Approach」の第三版(2009)にてFriendly AIや道具的収束の懸念が掲載されるといった影響を及ぼしました。Stuart Russelは「AI新生」書籍の一章の最初にて「私個人は、これは専門家だけの課題ではなく人類が直面している最重要課題だと2013年には確信するに至っていた。」と述べています。(2003年に既にリーマン予想による大惨事をSection 26.3: The Ethics and Risks of Developing Artificial Intelligence". *Artificial Intelligence: A Modern Approach*. にて書いていますが、本気にし始めたということでしょう。)

こうして2000年代にEliezer Yudkowskyが設立したSIAI、またOvercoming BiasやLessWrongを中心に合理性を重視して様々な考察(AIによるX-risk含む)を行う合理主義コミュニティはその後派生していき、ある意味で現実世界の存亡リスクへの対処やAI Alignment/Governanceへの投資などの運動を築き上げる理論的な支柱のような役割を果たしていきました。

合理主義コミュニティとは

※本節の大部分は合理主義コミュニティの歴史と文化を詳述したTom Chivers著の「AIは人間を憎まない」を参考に執筆しました。

ここで、AIによる存亡リスク周辺の文化や理論的支柱を2000年代から培い、後に説明する効果的利他主義コミュニティにも大きな影響を与えている合理主義コミュニティ(rationalist comunity)について説明します。

合理主義コミュニティとは元々Eliezer YudkowskyがAIによる存亡リスクに関する議論を他者とする際、機械の持つ合理性（あるいは知能、または最適化パワー）と人間の非合理性(認知バイアスなど)とはそもそもなんなのか、また機械と人間の間にある合理と非合理のギャップを説明する必要が出てきたために、形作られていったコミュニティです。

つまり、AIによる存亡リスクを他者と話すためには広範囲にわたる学術的テーマに関する議論が必要になってきたために、Eliezer Yudkowskyを端緒として大きくなっていったコミュニティが合理主義コミュニティと言えます。現在はAIによる存亡リスクを超えたさまざまなトピックが議論されている場となっています。

その合理主義コミュニティを構成する人々はテクノロジーを使って身体機能の拡張を目指すトランスヒューマニズムや人体冷凍保存に関心を持ち、この世界はシミュレーションに過ぎないという仮説を議論するなどある程度未来志向的です。
また、複数と同時に恋愛関係を築くポリアモリーといった一般とは異なる慣習に則っているためカルトだと批判されることもあります。
そのため、一般的な人と比較すると、変わった考えを持つ、変わった人たちのコミュニティと認識される場合もあります。

そのような合理主義者たちのコミュニティは現在世界中に広がっており、オンライン上で様々なトピックを議論することが主な一方で、十数都市に拠点を持ち実社会のコミュニティのなかで交流する人々もいます。
現在は世界数十都市にまで渡りリアルでの活動拠点も広がっており日本でも以下ACXTokyoコミュニティが存在しています。
https://www.lesswrong.com/groups/2Gx38j5JBc4AyHJ9a
AI Safetyに特化したコミュニティも存在しており以下があります。
https://aisafety.tokyo/

元々合理主義コミュニティの前身となったのは前の節でも紹介した1991年にTranshumanism運動に傾倒していたMax Moreの作ったExtropianメーリングリストです。そこではNick BostromやEliezer Yudkowskyを含め他大勢が、Transhumanismや未来に関する大きなトピックを語り合う場として活用していました。

しかし、合理主義者の先駆けとして重要な存在で、後にOvercoming Biasと呼ばれるブログサイトを創設した経済学者のRobin Hansonによれば、BostromもYudkowskyもそのExtropianメーリングリストでは満足しませんでした。

Bostromは未来に対して比較的リバタリアン的な考え方をするExtropian メーリングリストが好きではなく、1998年にHumanity+あるいはH+と改名された世界トランスヒューマニスト協会を作りました。そこまでリバタリアンにならずに、ユートピア的な理想を減らしてより左派的なアプローチを取れる場所として志向されていたようです。

またYudkowskyにとってはExtropianメーリングリストの面々には野心が欠けていると感じていたらしく、「SL4」というメーリングリストが2000年に立ち上げられました。SL4とは「(未来の)ショックレベル4」の頭文字をとったもので、1970年に出版されたAlvin Tofflerの「未来の衝撃」を踏まえたものです。ここで未来の衝撃とは「短い期間にあまりに多くの変化が起きる」感覚だとしています。Yudkowskyはその考えを一歩進めて、許容できる未来の衝撃のレベルに合わせて人をショックレベル0~4に分類しています。
彼によれば、Extropianメーリングリストにいるようなtranshumanistたちはショックレベル3で、最上位のショックレベル4(SL4:テクノロジーがどこかの時点で人間の生命を予見できない形に変え、私たちが知る生命の完全なる消滅という考えを許容できる人たち)に該当する人々を生み出すことを目指してSL4と名付けたのでした。

しかしYudkowskyは2004年頃までは頻繁にSL4に投稿していたものの、2005年からあまり関わらなくなっています。Yudkowskyが後に設立するLessWrongと呼ばれるサイトには「彼が対話する者たちが、彼にとっては当然の合理的思考をできないことへの困惑、いらだち、そして失望を度々表明していた。」そして「ベイズの定理を活用して思考することを伝えるもうまくいかず、SL4ではもっぱら沈黙するようになり、自らAIの安全性に関する調査に取り組むようになった」
と記されており、SL4の議論の質にYudkowskyが問題を感じていた様子がわかります。

その後2006年にSL4やExtropianメーリングリストにコメントをしていた経済学者のRobin HansonがOvercoming Biasという人間のバイアスを乗り越えることをテーマにしたブログを開設しました。Eliezer yudkowskyはそこで多くの投稿をし、後にその大量のテキスト群は「シークエンス(Sequence)」として大きく反響を呼び、知られるようになります。

つまるところ、Yudkowskyのブログや投稿はAIが人類の脅威になると言っても周りに理解してもらえないことからくる反動だったとも言えるでしょう。そこでAIについて語るならば、そもそも思考（合理性）とは何かについて説明する必要があり、対して人間の思考はバイアスやエラーに満ちていることを説明する必要がありました。

そしてYudkowskyは人間の思考を説明するためには全てを説明する必要があることに気づき、シークエンスの内容は進化生物学から量子力学やAIに至るまで多岐にわたる野心的なテキストの集積になっていきます。

ここで知能及び合理性とはYudkowskyにとっては自分の頭の中の世界像をできるだけ現実の世界に近づけることであり、自分の目標達成を可能にする決定をできるだけ多くの機会で選択することです。

どちらのプロセスも彼によれば「ベイズの定理」と呼ばれるシンプルな方程式を用いて説明できます。何らかの証拠に行き当たった時に、ベイズの定理によって算出された分量だけ自身の信念や確信を修正することが良い意思決定だと彼はいいます。また良い意思決定を算出する際は功利主義的な考え方を用いる傾向もあるようです。

このようなベイズの定理を用いた思考方法を普通の人間はしていないように思えますが、合理主義コミュニティはできる限り合理的にさまざまな人間の認知バイアスを排した形で意思決定を行うことを心がける傾向にあるコミュニティ、文化と言えるでしょう。

その後、このような思考の集積を2009年にYudkowskyは新たなウェブサイトLessWrongに移し、コミュニティのハブとして機能するように、誰もが投稿できる場所としました。
基本的にはOverComing BiasやSL4の流れを汲んで、認知バイアス、意思決定理論、AI、未来などについて議論する場ですが、2010年にはハリーポッターの二次創作である「Harry Potter and the Methods of Rationality(ハリーポッターと合理主義のメソッド)」も出版し、その合理主義スタイルの方法論を駆使して魔法の世界の法則を解明していく作品は大きな読者数を誘います。
また同年にはLessWrongにRokoというハンドルネームでRokoのバジリスク問題と呼称される意思決定理論に関する思考実験が話題を呼び、そのセンシティブな内容にYudkowskyは不快感を示し、話題になりました。
2012年には応用合理性センター(CFAR)と呼ばれる合理性と認知バイアスに関するワークショップを主催する非営利団体が設立されています。
その後、LessWrong上でおそらくYudkowskyの次に存在感のあるScott Alexanderが2013年に始めたSlate Star Codexというブログ記事も含め、合理主義コミュニティは広がりを見せていきます。

また、合理主義コミュニティの人々の傾向として、できる限り物事を数値化することを好み、特に何がどれほどの確率で起こるかを具体的な数値に変換することにもその熱量は注がれています。例えば、「超予測力:不確実な時代の先を読む10か条」という本に書かれているような形で未来において何が起こるかをできる限り人間のバイアスを排した形で確率的に予想しています。多くはMetaculusやManifoldといった予測プラットホームを利用している方も多いでしょう。

このような合理主義コミュニティには合理主義コミュニティの運動を形作ったEliezer Yudkowsky氏をカリスマ視し、セックスカルトだと非難される向きもあるようですが、多くはネット上のコミュニティに終始し、ミートアップに行ったことがあると言う人はSlate Star Codexのアンケート調査では

10%です。またポリアモリーを自認する人は世間の平均よりも少し多いとはいえ全員が全員ポリアモリーというわけではないことも念頭に置く必要はあるでしょう。

結論として合理主義コミュニティはAIによる存亡リスクから他幅広いトピックまで物事を理性的に、また時にはTranshumanism運動を源流としているところもあり、あらゆるトピックを未来志向的にも考察する人々の集まりだと言えるでしょう。このコミュニティは2010年代にAI存亡リスクの低減に向けた運動をしている効果的利他主義コミュニティに大きな影響を与えていくことになります。

効果的利他主義(EA)とは

2000年代にEliezer Yudkowskyを中心に形作られたMIRI,Overcoming Bias,LessWrongといった合理主義コミュニティのAI Alignment問題への思索や存亡リスクに早期から取り組んできたNick BostromによるFuture of Humanity Institute(FHI)での活動は2011年に設立された効果的利他主義(EA:Effective Altruism)という運動に大きな影響を与えていきます。

ここで効果的利他主義(EA:Effective Altruism)とは証拠と理性を使って、他の人にできるだけ利益をもたらす方法を見つけ出し、それに基づいて行動を起こすこととされます。

また、効果的利他主義は、この世界の最も切迫した問題とその最良の解決策の特定を目指す研究領域であるとともに、そうした発見を活用して〈よいこと〉を行うことを目指す実践コミュニティでもあります。

オックスフォード大学の倫理研究者であるToby Ord、当時研修医だった妻のBernadette Young、そして同じく倫理学者のWilliam MacAskillによって2009年に立ち上げられた寄付団体のGiving What We Can を効果的利他主義の始まりとして遡ることができます。
Toby OrdやWill MacAskillはオーストラリアの哲学者のPeter Singerの影響が慈善活動に興味を持った一つのきっかけとなったようです。
その後、Giving What We Canと2011年に設立された多くの良いことができるキャリアをコンサルティングする80000hoursを組み合わせて、正式な慈善団体として法人化し2011年12月の投票で「効果的利他主義センター」という名前がつけられました。

一方でアメリカでも同じように費用対効果にこだわって寄付先を選定する非営利団体のGiveWellがHolden Karnofsky とElie Hassenfeldによって2007年に設立されました。
そして2011 年、GiveWell チームには、Cari Tuna と Dustin Moskovitz(Facebook共同創設者)という約138 億ドルを所有する2人の非常に重要な新しい関係ができ、彼らはこの組織を拠点として、EAの最も強力な財団のOpen Philanthropyを創立しました。

これらGiving What We Canや80000 hoursとGiveWellやOpenPhilanthropyの二つの団体が行なっていることが近しいことが明らかになり、2011年までには繋がりを持ち話し合うことになります。

これら組織が効果的利他主義の「センター」という言葉が省略され、2012年頃から効果的利他主義と総称して呼ばれるようになりました。

How effective altruism went from a niche movement to a billion-dollar force *Effective altruism has gone mainstream. Where does that leave www.vox.com*

効果的利他主義への長期主義合理主義コミュニティからの影響

少し脇道にそれましたが、それでは、上記効果的利他主義運動に合理主義コミュニティ(LessWrongやOverComingBiasから派生した一群のコミュニティ)やFuture of Humanity Instituteはどのように影響を与えてきたのでしょうか？

●Giving What We Can(GWWC)と長期主義/合理主義コミュニティの関連
Giving What We Can(GWWC)を創始したToby OrdとWill MacAskillは2003年にオックスフォードでNick Bostromに出会っています。

特にToby Ordは「人類の危機に関するぼくの仕事はニックに大きく影響を受けている。ニックに影響されていなかったら、(効果的利他主義運動は)これほど強く人類の危機という点にこだわっていないだろうね」と発言しています。

(Tom Chivers著「AIは人間を憎まない」38節　効果的利他主義)

その後、Toby Ordは2005-2009年までオックスフォード大学で哲学の博士課程でNick Bostromと共著を2006年に出しています。
そのため、2005年にNick Bostromによってオックスフォードに設立されたFHIのことも勿論知っていたでしょうし、彼の存亡リスクやAstronomical Wasteの概念も知っていたと思われます。また、2009年(LessWrong設立年)にはLessWrongにToby Ordが記事にコメントを書いています。

よってToby OrdはNick Bostromの提唱する存亡リスク,長期主義の元となるAstronomical Wasteの概念や合理主義コミュニティにおいて話し合われていたAI Alignmentの問題についてもある程度はGWWC創設時の2009年時点で知っていたと思われます。

Will MacAskillもGWWCをToby Ordと2009年に共同創設しているため合理主義コミュニティの影響を少なからず受けていた可能性があります。

その証拠に効果的利他主義センターという単語が作られた翌年の2012年5月の80000hoursの記事にて、Will MacAskillは現在最も重要と思われる道徳的問題の一つとして4つ目に人類滅亡の危機を上げており、
「私たちが今後数世紀を生き延びた場合、将来生きる可能性のある人の数は数兆人になります(人間が哺乳類の平均寿命に対して現在の人口レベルで生きるとすると、10の13乗を超える人がいると考えてください) (または10兆人の将来の人類)。もし私たちが道徳的に、現在の人々を大切にするのと同じように潜在的な将来の人々を大切にすべきだとするなら、近い将来に人類が絶滅することによる損失は何兆もの命に上るかもしれません。」

と長期主義的な発想を既にしていることがわかります。

その後2013年には効果的利他主義のフォーラムにて初めてNick Bostromにより存亡リスクに関する記事が出ています。

また、Giving What We Canの最初の米国支部を共同設立し、Future of Humanity Instituteの研究員となり、その後オープン・フィランソロピーのプログラム・オフィサーとなるNick beckstead(とい

う長期主義に関連する概念をNickBostromと並んで整理した人物がいます。

Nick beckstead は2013年にNick Bostromの2003年のAstrnomical Wasteの議論に対して単なる存亡リスクの低減を目的とするのではなく、前向きな軌道変化を及ぼすという考えに置き換える修正意見を出しています。

上記からGiving What We Canを創始したToby Ordは2006年、Will MacAskillは2009年には長期主義的な考え方を認識はしていたと思われます。Giving What We Canを始めとする効果的利他主義コミュニティにおいては2012年には存亡リスクが記事として懸念はされており、2013年には具体的な論文といった形で明確に意識されていることがわかります。

●GiveWellと長期主義/合理主義コミュニティの関連
次に、GiveWellを創始したHolden Karnofskyは2007年には既に合理主義コミュニティとの繋がりがあり、Overcoming BiasやLessWrongについても当時から知っていたようです。
そのため、AI Alignment問題に関するMIRIの議論やNick Bostromの存亡リスクやAstronomical Wasteについての議論を知っており、興味深く、議論するのが楽しいと感じていたとHolden Karnofsky自身が記載しています。

そして2012年のGiveWellのブログ記事には壊滅的リスクにも優先的に取り組むと書かれています。一方この時点ではAIが及ぼすリスクについては言及がありません。

それは2012年まで彼はAI Alignment等の議論を楽しいとは思いつつも、もし存亡リスクやAI Alignment問題がそこまで重要なトピックなら世界的にもう既に関連する専門家がいるはずであり、それらの専門家がいないように見えたため、彼らの議論には大きな欠点があり専門家コミュニティから見放されているのだろうと感じていたようです。
しかし、2013年からAIや機械学習のコミュニティの人々と話すにつれて、議論に欠点があるから興味が持たれていないのではなく、そもそも誰もこのAI Alignment問題について考えていなかったということが明白になり始めたようです。
そして決定的な心変わりの瞬間として、Nick Bostromの「Superintelligence: Paths, Dangers, Strategies」が出版された後の出来事を挙げています。
当初Karnofsky氏はこの本がAI および機械学習コミュニティの人々に広く無視されたり、酷評されるだろうと強く予測していましたが、知人の機械学習研究者にとってもこれらAlignment関係の概念については初耳だったようです。そして、著名人(Elon Musk,Bill Gates,Sam Altman等)に影響を与え、特に主流のAI研究者のStuart RusselがValue Alignment問題の存在を認めたことで今までの懐疑的な見解をKarnofskyは180度変えました。

Three Key Issues I've Changed My Mind About | Open Philanthropy Philanthropy – especially hits-based philanthropy – is driven www.openphilanthropy.org

そうして、2014年にAIの脅威とそのAIアライメント問題の重要性に気づいたHolden Karnofsky氏ですが、その後はOpenPhilanthorpyが世界的な壊滅的リスクに関する2014年の調査結果にて今後AIのもたらす潜在的なリスクの調査を行うと記載し、2015年に高度な人工知能がもたらす潜在的リスクの調査結果を発表します。

つまりGiveWellを創始したHolden Karnofskyは2007年には存亡リスクを認識はしていましたが、2012年まで本気にはしていませんでした。しかし2013年頃から単にAI Alignment問題を誰も真面目に考えていないだけだと気づき、考え方を改めました。

GiveWellの公式ページでも2012年から存亡リスクを優先的に取り組むと言及し、本格的に2014年にはAIによる存亡リスクについての調査が開始されました。

脅威を予見することの重要性

Nick Bostromが存亡リスクについて定義した2002年の論文の中で以下のように書いています。

> 存亡リスクに対するアプローチは試行錯誤的なものであってはなりません。エラーから学ぶ機会はありません。何が起こるかを確認し、損害を制限し、経験から学ぶという事後対応型のアプローチは機能しません。むしろ、積極的なアプローチをとらなければなりません。これには、新しいタイプの脅威を予測する先見性と、断固とした予防措置を講じ 、そのような行動のコスト（道徳的および経済的）を負担する意欲が必要です。

https://nickbostrom.com/existential/risks

上記論文で「新しいタイプの脅威を予測する先見性と、断固とした予防措置を講じ 、そのような行動のコスト（道徳的および経済的）を負担する意欲が必要」も書かれている通り、例え不確実だとしても起こりうる脅威を何らかの手法で予想することが必要という認識が存在していました。

※既にNick Bostromは1997年の博士論文で確率論を使用して人類文明の寿命について推論する終末論の研究(記事はこちら)や技術発展が及ぼす新たな哲学的課題に対する考察、超知能の誕生とそのタイムライン予想とリスクへの言及を行っており、1997年にはトランスヒューマニズム運動が挫折する人類存亡リスクを認識していたと思われます。

その後、Nick Bostromの創設したFuture of Humanity Instituteは2008年に世界的な壊滅的リスクのSurveyを行なっており、地球規模の壊滅的リスクに関するまとまった本も出版されています。Eliezer Yudkowskyもその本の中で、主にAI起因のリスクを語っています。

2014年にはOpenPhilanthorpyが世界的な壊滅的リスクの2014年の調査結果にて今後AIのもたらす潜在的なリスクの調査を行い、2015年には高度な人工知能がもたらす潜在的リスクの調査結果を発表しています。

また、その後も2014年に全脳エミュレーションを用いたTAI(Transformative AI)のタイムライン予想が行われ、2015年には効果的利他主義コミュニティとは直接は関係はないかもしれませんがMetaculusと呼ばれるオンライン予測プラットホームが設立され2021年にEAから助成金を付与されています。
また最も有名なTAIのタイムライン予想はOpen PhilanthropyのアナリストのAjeya Cotraのバイオアンカー仮説を用いた2020年のレポートとなっています。
そして超予測者(SuperForecastor)と呼ばれるそのドメインの専門家よりも優れた予測成績を残している人たちから構成されるSamotsvety Forecastingと呼ばれるグループによるAGIタイムラインの予想やAIによる人類壊滅/存亡リスクの推定が行われています。
他にも様々なTAIのタイムラインの予想が2010年代から今まで試みられており、まとめられています。

上記から未来の脅威を予見するためにテクノロジーの進歩を予想する相応の努力が効果的利他主義コミュニティ周辺で行われていることがわかります。

2015年以降のEAコミュニティの活動

AIの安全性はEA Global 2015のキーテーマとなり、パネルにはElon Musk, Daniel Dewey, Nick Bostrom, Nate Soares ,Stuart Russellが並びました。
同時期にはAIの安全性に関する会議「The Future of AI: Opportunities and Challenges」がプエルトリコで開催されました。この会議はFuture of Life Instituteが主催していました。
写真を見ると分かる通り、Nick BostromやEliezer YudkowskyといったAI による存亡リスクに早期から取り組んできた人やDeepMindのDemis Hassabis CEOやOpen AI主任研究員のIlya Sutskever、Elon Musk等著名人も2015年1月の会議に参加していることがわかります。

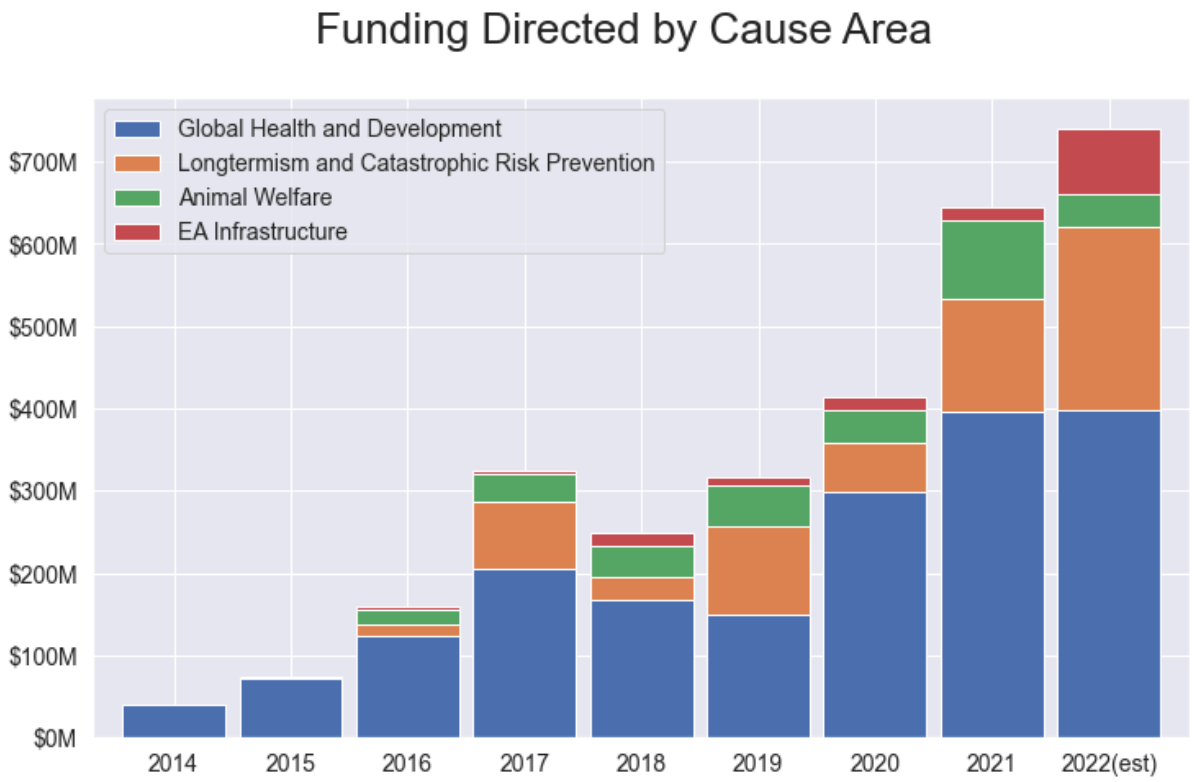
図1.3　2015年1月にプエルトリコで開催した会議には、AIやその関連分野の一流研究者が集結した。後列左から、トム・ミッチェル、シーン・オヘイガルテー、ヒュー・プライス、シャミル・シャンダリア、ヤーン・タリン、スチュワート・ラッセル、ビル・ヒバード、ブレイス・アグエラ・イ・アルカス、アンダース・サンドバーグ、ダニエル・デューイ、スチュワート・アームストロング、ルーク・ミュールホイザー、トム・ディーテリッヒ、マイケル・オズボーン、ジェイムズ・マニカ、アジェイ・アグラワル、リチャード・マラー、ナンシー・チャン、マシュー・プットマン。後列以外の立っている人たち、左から、マリリン・トンプソン、リチャード・サットン、アレックス・ウィスナー＝グロス、サム・テラー、トビー・オード、ヨッシャ・バッハ、カティヤ・グレース、エイドリアン・ウェラー、ヘザー・ロフ＝パーキンス、ディリープ・ジョージ、シェーン・レッグ、デミス・ハサビス、ヴェンデル・ヴァラッハ、チャリーナ・チョーイ、イリヤ・サツケヴァ、ケント・ウォーカー、セシリア・ティリ、ニック・ボストロム、エリック・ブリニョルフソン、スティーヴ・クロッサン、ムスタファ・スレイマン、スコット・フェニックス、ニール・ヤコブシュタイン、マレー・シャナハン、ロビン・ハンソン、フランチェスカ・ロッシ、ネイト・ソアレス、イーロン・マスク、アンドリュー・マカフィー、バート・セルマン、ミシェル・ライリー、アーロン・ヴァンデヴェンダー、マックス・テグマーク、マーガレット・ボーデン、ジョシュア・グリーン、ポール・クリスティアーノ、エリエゼル・ユドカウスキー、ディヴィッド・パークス、ローラン・オルソー、J.B.ストローベル、ジェイムズ・ムーア、ショーン・レガシック、メイソン・ハートマン、ハウイー・レンペル、ディヴィッド・ヴラデック、ヤコブ・シュタインハート、マイケル・ヴァッサー、ライアン・カロ、スーザン・ヤング、オワイン・エヴァンズ、リヴァ＝メリッサ・テズ、ヤーノシュ・クラマー、ジェフ・アンダーズ、ヴァーナー・ヴィンジ、アンソニー・アギーレ。しゃがんでいる人たち、サム・ハリス、トマソ・ポッジョ、マリン・ソリャシウ、ヴィクトリヤ・クラコフナ、メイア・チタ＝テグマーク。撮影：アンソニー・アギーレ（隣にしゃがんでいる人間レベルの知能のそばにフォトショップで合成した）。

プエルトリコで開催されたAI Safetyに関する会議の参加者

2017年には有益な AI に関するアシロマ会議が Future of Life Instituteによってアシロマ会議場で開催されます。これは、2015 年のプエルトリコ会議の後継です。その結果、AI 研究のガイドラインである23のAsilomar AI 原則が作成されることになります。

また2015年までのAIによる存亡リスクに懸念する世界的な流れを受けて、2016年には効果的利他主義コミュニティでキャリアのコンサルティングを行なっている組織の80000hoursの記事では、AIのもたらすリスクに取り組むことは最も差し迫った問題の一つと表現されました。
（この時点では最も差し迫った「問題の一つ」と表現されています）

また、効果的利他主義の助成金の内訳の話をすると、2014年からgivewell,openphilanthropyから効果的利他主義コミュニティへ助成金が支払われており、2016年から長期主義的な運動資金が多くなっています。



Historical EA funding data — EA Forum *Summary I have consolidated publicly available grants data fr forum.effectivealtruism.org*

また上記記事の返信には以下のように長期主義関連(AI,バイオセキュリティ、一般的なX-risk、核兵器リスクなど)の助成金の内訳が載っており、AI関連が50%と一番多いようです。

AI	$233M	52.6%
Biosecurity	$157M	35.3%
General X-Risk	$33M	7.4%
Broad longtermism	$15M	3.3%
Nuclear risk	$6M	1.5%
	$444M	

2012-2022までのOpenPhilanthropyからの長期主義関連助成金の内訳

そして2017年のアシロマ会議が行われた後の2018年8月には80000hoursの記事にて、人工知能のリスクを軽減することが最も緊急を要する壊滅的リスクのうちトップになりました。
（補足すると、世界で最も重要な問題がAIによる存亡リスクと言っているわけではなく、もしキャリアを選ぶとしたら最も世界全体の流れにインパクトを残せる可能性の高いキャリア選択として言われているのだと思われます。）

その後、効果的利他主義コミュニティへの長期主義に関連する助成金は2021年頃から多くなっています。
また、2020年にはToby OrdによるExisistential Riskに関する本「The Precipice: Existential Risk and the Future of Humanity」や2022年に William MacAskillにより長期主義に関連する本の「What We Owe the Future」が出版され、存亡リスクの紹介が一般にも広がり始めました。

効果的利他主義系コミュニティの広がり

上記のように、2010年代前半から存亡リスクの深刻性の認知が広がる流れの中で、そのリスクを分析したり、リスクを軽減する取り組みをする組織も多く設立されていくことになります。以下は2000-2024年までのAI Safety/長期主義関連の組織や主な出来事年表です。

・2000
Eliezer Yudkowsky 機械知能研究所 (MIRI) 設立,
Bill Joy「Why the Future Doesn't Need Us」
・2001
Eliezer Yudkowsky「Friendly AI」提唱
・2002
Nick Bostrom「Existential risk」提唱
・2004
Eliezer Yudkowsky「CEV」提唱
・2005
Future of Humanity Institute(FHI)設立
・2006
OverComing Bias設立
・2008
Steve Omohundro「道具的収束」の概念を提唱
・2009
LessWrong設立
・2010
DeepMind設立
・2011
Centre for Effective Altruism,
Global Catastrophic Risk Institute設立
・2012
Nick Bostrom「直交仮説」提唱
Centre for the Study of Existential Risk設立
・2013
James Barrat「Our Final Invention: Artificial Intelligence and the End of the Human Era」出版,
Center on Long Term Risk (CLR) 設立
・2014
Nick Bostrom「Superintelligence: Paths, Dangers, Strategies」出版
Stuart Armstrong「Smarter than us 」出版
Future of Life Institute(FLI) ,
AI impacts,
Center for Applied Rationality,
The Future Society設立
・2015
FLIはAI の安全性に関する公開書簡を回覧し、その後Stephen Hawking、Elon Musk、および多くの人工知能研究者によって署名。
プエルトリコでAI Safety会議を開催。
OpenAI,
Metaculus設立

・2016
Center for Human-Compatible Artificial Intelligence,
Leverhulme Centre for the Future of Intelligence設立
・2017
Max Tegmark Life3.0出版
FLI アシロマAI原則発表
・2018
AI Alignment Forum,
Ought,
The Centre for the Governance of AI,
Global Priorities Institute,
Rethink priorities,
Median Group設立
・2019
Stuart Russel Human Compatible出版
Center for security and emerging technology,
The Centre for Long-Term Resilience,
Quantified Uncertainty Research Institute,
Convergence Analysis 設立
・2020
Toby Ord氏　The Precipice出版
Brian Christian The Alignment Problem出版
EleutherAI,
Legal Priorities Project,
Future Matters 設立
・2021
Anthropic,
Aligned AI,
Alignment Research Center ,
Redwood Research,
Existential Risk Observatory,
AI Objectives Institute,
International Center for Future Generations,
Manifold 設立
・2022
Center for AI Safety,
Conjecture,
Epoch,
FAR AI,
Encultured AI,
Forecasting Research Institute,
Fund For Alignment Research,
Modeling Cooperation,
Association for Long Term Existence and Resilience (ALTER),
AI Policy Institute 設立
・2023
Cavendish Labs,
Appolo Research,

orthogonal,
Transformative Futures Institute,
Center for AI policy,
Align the world,
PauseAI,
StopAGI,
UK AI Safety Institute,
US AI Safety Institute
設立
・2024
日本 AI Safety Institute設立予定

2000-2024年までのAI Safety/長期主義関連の組織や主な出来事年表

上記を見ると分かるように2010年代から急速にAI Alignment/Governance/存亡リスク関連の組織が設立されているのが分かると思われます。
ここから世界的にもこれら分野が急速に問題視され、注目を集めている分野だということが理解できるのではないでしょうか。

また、AI Alignment/Governance関連の組織やリソースに関する情報は以下のマップに網羅的に掲載されています。

Map of AI Existential Safety *aisafety.world*

また様々な概念や組織に関する網羅的な年表/タイムラインを掲載している以下のTimelineWikiは便利なためこちらも参考に置いておきます。

Timeline of AI safety - Timelines *timelines.issarice.com*

OpenAI/DeepMind/Anthropicへの影響

前節のように存亡リスクに対応する取り組みが2010年代から増えていく中で、DeepMindやOpenAIの設立にも少なからず影響を与えています。

・DeepMindの設立への合理主義コミュニティの影響
DeepMindの共同創業者であり、主任AGI科学者のShane Leggは2008 年に10,000 ドルのカナダシンギュラリティ人工知能研究所賞を受賞しました。
そして2009年のブログ記事にはAI Alignment問題に関して「相対的に言えば、SIAI(後のMIRI)が現時点で私たちが持つ最大の希望であるように思えます。」と記載しています。
これは後にDeepMind設立へのPeter Thielの投資を呼び込むきっかけとなりました。

2010年、Peter Thielのサンフランシスコのタウンハウスで、Eliezer Yudkowskyは彼にShane LeggとDemis Hassabisという二人の若い研究者を紹介しました。その秋、Thielの会社からの投資を受けて、二人はDeepMindというAIラボを設立したということです。
また、2011年のShane LeggのインタビューQA LessWrong投稿にて、AIがもたらすリスクは今世紀最大のリスク第一位であり、第二位が人工的なパンデミックだと記載しています。

これら情報からShane LeggはEliezer Yudkowskyと知人でYudkowskyがPiter ThielにDemis

Hassabisを紹介してDeepMindが設立されたと言えます。
またShane LeggはMIRIが研究している内容や合理主義コミュニティにおける存亡リスクの議論についてDeepMind設立初期から知っていたといえるでしょう。

また、その後、2014年にGoogleがDeepMindを買収する際、Shane LeggはGoogleにDeepmindのAI研究のリスクを監視するためのAI倫理委員会を設置するよう説得する上で極めて重要な役割を果たしたようです。
同記事では、グーグルに委員会の勧告を遵守させた有力者として、Peter ThielとJann Talineも挙げられています。彼らはMIRIの支援者であり、同組織の第1位と第3位の寄付者でもあり、初期のAI Safetyのエコシステムを形作った出資者です。

つまるところ、Eliezer Yudkowskyの起こしたMIRIとそのエコシステム（合理主義コミュニティやシンギュラリティサミットなどのイベント）がDeepMindのAIの深刻なリスクに対処するための動機付けを一部形作ったといえるでしょう。

・Elon Muskの存亡リスク懸念とOpen AIの設立
2012年、Peter ThielによるDeepMindへの投資と設立から約2年後、Elon Muskの会社Space Xにも資金を投入していたThielの投資ファンドが主催したカンファレンスでDemis HassabisとElon Muskが会談しました。
Elon Muskは、自分の計画は地球上の人口過剰やその他の危険から逃れるために火星に植民地化することだと説明しましたが、Hassabisは、超知能機械が後を追って火星の人類を滅ぼさない限り、この計画はうまくいくだろうと答えたようです。
Muskはそのような特別な危険についてはそれまで考えていなかったと書かれています。Elon MuskはすぐにPeter ThielとともにDeepMindに投資しました。

2014年1月にGoogleはDeepMindを買収します。
2014年10月にElon MuskはAeroAstro100周年記念シンポジウムにて、コンピューターサイエンス、AI、宇宙探査、火星の植民地化について語る中で、「人工知能については細心の注意を払うべきだと思います。私たちの存亡に関わる最大の脅威は何かと推測しなければならないとしたら、それはおそらくそれでしょう。したがって、非常に注意する必要があります」と語りました。

2015年2月にはOpen AIのCEOとなるSam Altmanがブログ記事にて
「超人機械知能 (Super Machine Intelligence) の発展は、おそらく人類の存続に対する最大の脅威です。」
「Nick Bostromの優れた本『Superintelligence』は、このテーマに関して私がこれまでに見た中で最高のものです。一読の価値は十分にあります。」
と記載しています。

2015年7月、ElonMuskが開催したパーティーにてGoogleの共同創設者のLarry PageとElon MuskはAIのもたらす存亡リスクについて議論し、意見を異にしたようです。
Larry Pageは人間は最終的には人工知能を備えた機械と融合し、ある日、さまざまな種類の知性が資源を求めて競争し、最も優れたものが勝利するでしょうといい、Elon Muskはそうなれば、我々は破滅するだろうといったようです。

Ego, Fear and Money: How the A.I. Fuse Was Lit *The people who were most afraid of the risks of artificial in www.nytimes.com*

その後、Elon Muskはカリフォルニア州メンローパークにあるホテルで、当時Y CombinatorのCEOだったSam Altman,Greg Brockman, Dario Amodei(Anthropic共同創設者), Chris Olah(Anthropic 共同創設者), Paul Christiano(Alignment Research Center創設者), Ilya Sutskeverと食事をしました。この食事会が2015年12月のOpenAIの誕生につながりました。
OpenAIが「『人類の利益となるような』汎用人工知能」という理念を掲げているのは、プールサイドの焚き火のそばでLarry Pageが主張していた「いずれ機械と人類が融合していき、優れた方が残っていく」という思想から人類を守るためだとElon Muskは考えていた可能性もあります。
その後、2018年にElon Muskは経営方針の違いで取締役会を辞任します。

Elon Muskが離れたことで、OpenAIは新たな資金調達が必要となり、知り合いだったMicrosoftの最高技術責任者であるKevin Scottのつてで、Satya Nadella CEOと面会。そして、OpenAIの非営利組織下に営利法人を設立し、Microsoftからの投資10億ドルを得ることに成功しました。

一方後にAnthropic CEOになるDalio Amodeiは2016年にOpen AIのAI Safety Teamを率いており、AI Safetyに関する問題を分析していました。Dalio Amodeiは少なくとも2009年にはGive Wellを知っていたようで、効果的利他主義運動に興味があったことが示唆されます。

しかし、2021年にはDalio AmodeiはOpenAIを商業的な方向に導こうとしているMicrosoftに対して不満を抱いていたとのことです。そのため、Sam Altmanを取締役会から追い出そうと画策したようですが、失敗に終わり、彼はOpenAIを離脱し、15人のエンジニアや科学者と共にAnthropicと呼ばれるAI Alignmentに関する新しい研究所を設立するに至ります。

DeepMind・OpenAI・Anthropicの設立の経緯について、AIのリスクを最も恐れていた人々は自分たちがトップになるべきだと決心し構築するまで *AlphaGoやAlphaZeroを開発したDeepMind、GPT-4やChatGPTなどを開発したOpenAI、チャッ gigazine.net*

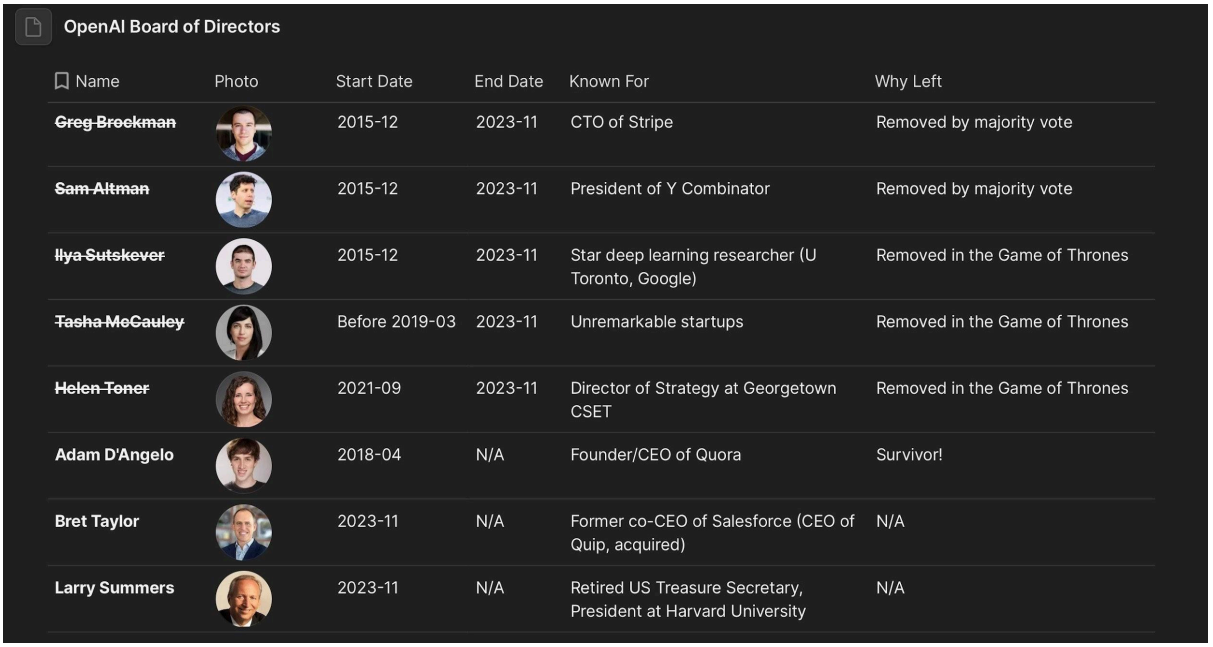
その後Open AIの取締役会は変遷していき、以下の画像のように、2018年にはQuora CEO Adam D'Angelo、2019年までにはCentre for Effective Altruismの母体であるEffective Ventures Foundationの理事会メンバーのTasha McCauley、2021年にはCenter for the Governance of AIの諮問委員Helen Tonerが取締役会メンバーとなります。
(参考：https://forbesjapan.com/articles/detail/67426)

OpenAI Board of Directors

Name	Photo	Start Date	End Date	Known For	Why Left
Elon Musk		2015-12	2018-02	Founder/CEO of SpaceX, Tesla	Conflict of vision and interest
Holden Karnofsky		2017	2021	Director of Strategy at OpenPhilanthropy	Left when Dario Amodei started Anthropic (married to his sister)
Reid Hoffman		2018-??	2023-03	Founder/CEO of LinkedIn	Started competitor Inflection.ai
Shivon Zilis		Before 2019-03	2023-03	Investor at Bloomberg Beta	Siding with Elon's vision (my guess)
Will Hurd		2021-05	2023-07	US Representative from Texas	2024 presidential run
Greg Brockman		2015-12	2023-11	CTO of Stripe	Removed by majority vote
Sam Altman		2015-12	2023-11	President of Y Combinator	Removed by majority vote
Ilya Sutskever		2015-12	Present	Grad student U Toronto, Researcher at Google Brain	Still on board!
Adam D'Angelo		2018-04	Present	Founder/CEO of Quora	Still on board!
Tasha McCauley		Before 2019-03	Present	Unremarkable startups	Still on board!
Helen Toner		2021-09	Present	Director of Strategy at Georgetown CSET	Still on board!

2023/11/17までのOpenAIの歴代取締役メンバーの状態と脱退理由

一方でOpenAIは2023/11/17にSam Altman CEOを取締役会の賛成多数で解任するという事件が起きました。
その後AIの安全性に懸念があったことが今回の解任騒動につながったという話や、Sam AltmanとHelen Tonerの間で一貫して率直ではないといえるようなコミュニケーションの問題があったこと、AI技術のデュアルユースと底辺への競争の懸念等憶測がありますが、2024年1月6日時点では騒動の全容は明らかになっていません。
結果として現時点では以下のような取締役会メンバーとなっています。



OpenAI Board of Directors

Name	Photo	Start Date	End Date	Known For	Why Left
~~Greg Brockman~~		2015-12	2023-11	CTO of Stripe	Removed by majority vote
~~Sam Altman~~		2015-12	2023-11	President of Y Combinator	Removed by majority vote
~~Ilya Sutskever~~		2015-12	2023-11	Star deep learning researcher (U Toronto, Google)	Removed in the Game of Thrones
~~Tasha McCauley~~		Before 2019-03	2023-11	Unremarkable startups	Removed in the Game of Thrones
~~Helen Toner~~		2021-09	2023-11	Director of Strategy at Georgetown CSET	Removed in the Game of Thrones
Adam D'Angelo		2018-04	N/A	Founder/CEO of Quora	Survivor!
Bret Taylor		2023-11	N/A	Former co-CEO of Salesforce (CEO of Quip, acquired)	N/A
Larry Summers		2023-11	N/A	Retired US Treasure Secretary, President at Harvard University	N/A

2024/1/6のOpenAIの取締役メンバーの状態と脱退理由

上記のOpen AIの解任騒動を巡る動きの中で、高度なAIの安全性を巡る懸念による対立があったのかどうかは定かではありませんが、この解任騒動をめぐって、効果的利他主義という運動やネット上でのインターネットミームのようになっている効果的加速主義(e/acc)に関する話題がSNS上で注目を集めたのは印象的です。

つまり、Open AIの創設やその後のAnthropicの分離、また効果的利他主義に関連する活動をしている取締役メンバー等、2010年代前半から始まるAIによる存亡リスクに関わる運動が各所に影響を及ぼしていることが理解できるのではないでしょうか。

※余談
Open AI(主にSam Altmanの方針だと思われます)のAIの技術開発スピードに対する態度について補足します。
2023年2月に出したAGIとそれ以降の計画を記述したOpen AIのブログ記事には

「AGI は近い将来、あるいは遠い将来に起こる可能性があります。初期の AGI からより強力な後継システムまでの離陸速度は、遅い場合も速い場合もあります。私たちの多くは、この2行2 列のマトリックスで最も安全な象限は、短いタイムラインと遅い離陸速度であると考えています。AGIまでのタイムラインが短いと、アライメントが容易になり、ハードウェアオーバーハングが少なくなるため離陸が遅くなる可能性が高くなります。また、離陸が遅くなると、安全性の問題を解決する方法や適応する方法を経験的に理解するための時間がより多く得られます。」

という記載があり、Open AIはAGIをできるだけ早く開発する過程で、段階的に社会に適応させ、その後の超知能の発展を遅くする方が良いという指針を出しています。
これはAGIの開発をそもそも減速か停止する必要があると考えるEliezer YudkowskyやFLIなどの規制を求める機関との意見の相違が明確になっているところでしょう。
一方、AIの一時停止にも様々な種類が提案され、議論がされており、
また、Open AIの方針にも一理ありAI開発の一時停止は逆に問題が起こるという指摘もあります。

今後、上記のようなOpen AIのAI開発指針に対する方針や他一時停止の議論を認識した上で状況を注視する必要があるでしょう。

効果的加速主義(e/acc)

効果的加速主義(e/acc)は「熱力学の第 2 法則に根ざした信念であり、宇宙自体は絶えず膨張する生命を生み出す最適化プロセスであるという考え」「イデオロギーではなく、運動でもなく、それは単に真実の認識」とする用語で、テクノロジー業界で2023年半ばから話題になり始めました。

効果的利他主義運動をもじった主義の名前をつけていることからも分かる通り、効果的加速主義の創設者の一人のBeff Jezosによると主にテクノロジーの規制ではなく、加速を支持し、AGIのもたらすX-Riskへの懸念よりもAGIを規制することによるリスクを懸念しています。

2023年6月にSam Altmanは、Based Beff Jezos氏のツイートに返信し「私を上回る加速はできない」と冗談を飛ばしました。
2023年7月にはテクノロジーの発展の楽観的な側面を強調するThe Techno Optimist Manifestoを書いたベンチャーキャピタル・Andreessen Horowitz(a16z)の創業者でもあるMarc Andreessenはこのe/accを支持しています。
2023年11月にはY Combinatorの社長であるGarry Tan氏がe/accへの支持とも取れる言及をしています。e/accに関するDJパーティーも開かれ、そしてこの運動は徐々に AI を超えて広がり、暗号通貨や核融合にも言及され始めました。
11月17日にはOpen AI解任騒動もあり、加速主義と効果的利他主義の対立というようなわかりやすいストーリーも溢れていました。

ここでe/accまたは効果的加速主義に関する最も古い言及は、2022年5月31日と6月1日のもので、Xでの@zetular、@BasedBeff、@creatine_cycleらの「少年たちと新しい哲学を発明した」というポストから始まっており、その後、@BasedBeffや@zetularはe/accに関する説明をsubstackに載せ始めます。
その中でBeff Jezosは効果的加速主義を熱力学第二法則と絡めて説明しています。
「生命は、散逸的適応（ジェレミー・イングランドの研究を参照）として知られる、熱力学的な非平衡プロセスから生まれている」
「散逸的適応（ジャージンスキー・クルックスのゆらぎ散逸定理に由来する）は、宇宙は、物質がより多くの自由エネルギーを取り込み、より多くのエントロピーに変換するように適応した未来を（存在／発生の確率の点から）指数関数的に好むことを教えてくれる」
といった議論からe/accは宇宙の熱力学的意志に逆らっても無駄なので加速を受け入れることを目指しています。

※余談ですが、この効果的加速主義は2010年代から存在するNick Landの加速主義から言葉をとってきていますが、主な目的は効果的利他主義に対して付けられた名前だと思われ、思想的な繋がりはそこまではっきりとは見えません。脱領土化というドゥルーズの言葉も使われているため、伝統的な加速主義の文脈も認識しているとは思われますが、直接的に関連があるようには思われません。

そして、GPT-4リリースから効果的加速主義のミームはインターネット上で広まり、上記話したように著名人に言及されるまでになりました。
その後、2023年12月1日ForbesによってBeff Jezosの正体はTensorflow Quantumの開発者で元Googleでブラックホールの理論物理の研究をしており、AIハードウェアスタートアップExtropicAI創業者のGuillaume Verdonと暴かれます。

その後、正体が暴かれてしまったということもあり、2023年12月30日にPodCasterのLex FriedmanとGuillaume Verdonは対面で対談 をしています。
対談によると、宇宙を理解したいという小さい頃の思い入れから、理論物理学を学び、その後自然と同じ方法で計算する手法を見つけたいという思いから、量子機械学習の分野に進んだとのことです。
一方2021.22年はテクノロジーに対する悲観論（恐らくAIによる存亡リスクによるもの）の影響が大きくなり始めており、このままだと悲観的な未来が実現する可能性が高まると感じたため、未来はもっと良くなるという楽観的な雰囲気を作るためにこの効果的加速主義という運動を作るべきだという責任を感じたようです。

※前述しましたが、実際に効果的利他主義コミュニティには2021年頃から長期主義に関連する助成金が多くなっており、2020年にはToby Ordによる「The Precipice: Existential Risk and the Future of Humanity」や2022年にWilliam MacAskillにより長期主義に関連する本の「What We Owe the Future」が出版され、存亡リスクに関する一般への啓蒙活動も目立ち始めた時期なのだと考えられます。

実際Guillaume VerdonはAIによる存亡リスクは「ゼロまたはゼロに近い確率」であると考えているようで、相当テクノロジーに対して楽観的なようです。
一方で、彼は意識の喪失は「宇宙において絶対的に最悪の結果となるだろう」と書いているように、人類の絶滅に関しては問題視していると思われ、人類を絶滅させてまでAIの加速を支持するという立場ではないようです。
あくまで、AIによる存亡リスクがとても低いという事前信念を持っている形でしょう。
そして彼が考えている最大の存亡リスクのひとつはAIの力がごく少数の手に集中し、権力を与えられ、中央集権化しすぎた結果、人間が恐ろしいことをすることだと考えています。
彼がExtropicAIを設立した経緯も、Nvidiaなどが一強であるAIのコンピューティングリソースのサプライチェーンを壊してオープンなAI開発をいたるところで促進するためだと語っており、AIの規制による中央集権化がもたらすリスクへの懸念から単なるネット活動だけではなく、行動を起こし1410万ドルの調達をしています。

e/accは一般にインターネットミームとして無視されがちですが、今後AIの規制と緩和の流れに影響を与える可能性もあり、注視する必要があるかもしれません。

まとめ

存亡リスクの歴史からそもそも人類自身が自ら絶滅する可能性に気づいたのはほんのここ数百年といえるでしょう。また、熱力学の第二法則による宇宙の終焉よりも短期間で人類は絶滅し得ると認知されたのはおよそここ100年です。
一方、AIによる存亡リスクについては20世紀から懸念レベルでは存在していましたが、2000年代に入ってから真面目な考察対象として学術的に論じられ始めました。そして当初の主なモチベーションの一つはTranshumanismが挫折する可能性の高いリスクがAIだったからといえると思われます。
その後2010年代の深層学習の成功により、多くの著名人や研究者もAI Alignment研究に深刻さを感じ参入し、投資が行われています。
またその流れはOpen AIやDeepMindの設立経緯にも関わっており、今の国際的なAIへの懸念を形作っていることがわかるのではないでしょうか。

AI脅威論/長期主義への批判や議論

今まではAIがもたらす存亡リスクの論理とその緊急性について解説してきましたが、勿論それらに対する批判や懸念も存在します。

AI存亡リスクへの論理的批判/議論

AIのX-risk(Exisitential risk)の前提である以下A,B,Cという論理に対する批判を整理した論考があります。
A 超人的な AIシステムは目標指向型になる
B 目標指向型AIシステムの目標は悪いものになる
C 超人的なAIは人類を圧倒するほど人間よりも優れている

・Aについての反論
AIシステムは未来において急増し、さまざまな目標を持っていますが、これらのシステムは一般的な効用関数を最大化しているのではなく、むしろ「ウォルマートを利益にしようとしている誰かのように振る舞っている」ため、これらの目標を推進するために世界を最適化するという認識可能な願望はないかもしれません。そして需要のほとんどは、世界を参照せずに何らかの既知の方法で入力を変換するGPTやDALL-Eのようなシステムに対するものかもしれません。

・Bについての反論
AI脅威論を唱える人が考えているほど、AIの価値観と人間の価値観は離れて学習されないかもしれません。学習データに存在する帰納的バイアスを素直に表現してくれる可能性もあります。

・Cについての反論
AIにとっても今日のさまざまな社会課題を解決することは難しいかもしれません。人間レベルの知能以上にはなるかもしれませんが、そこまで大きなギャップはなく、世の中の問題がとても変数が大きく難しいことに起因して能力差による存亡リスクも起こりにくいかもしれません。

他にもAI X-riskの論理について批判的な以下のような考察がまとめられています。

- Eliezer YudkowskyのAI脅威論に対する批判的な考察
- 欺瞞的アライメントが起こる可能性は相当低いと考えられる理由
- 超知能の危険性を訴える際に使われるアナロジーとしてEliezer Yudkowskyの進化の比喩は不適切でNate Soaresのそれも根拠がないとする指摘
- AI Alignment分野はAIの能力の向上と比較して進んでいないと言われていますが、アライメント分野の進展を楽観視できると考えられる理由
- 直交仮説に関する批判的な考察記事
- AIによる存亡リスクの懸念は認めつつ存亡リスクの議論は不確実性の高い推論になっていることを指摘する議論
- そもそもAI Alignment自体が人類のS-risksを増大させるとするエッセイ
- AIのもたらす壊滅的なリスクに対する反論の定型の分類

このようにAIのもたらす存亡リスクに関しては、AI Alignmentという言葉ができてからまだ10年程度と日が浅いこともあって議論が続いている状況であり、状況は流動的な面があるという認識を持つことが大切でしょう。

また参考にAIによる存亡リスクに対する批判的な記事/考察リンクを網羅的にまとめたリストもあり貼らせていただきます。

長期主義批判/先制攻撃・監視の是非

前節ではAIによる存亡リスクに対する技術的な側面からの批判や議論を紹介しましたが、この節では主にそれら長期主義に関連する考え方や運動への批判を紹介します。

アメリカの神経科学者のAnthony ZadorとMetaの主任AI科学者のYann Lecunは「AIが自律的に人類を存亡の危機に陥れるターミネーターのようなシナリオへの心配は、AIの非常に現実的なリスクから私たちの気をそらす。
それは兵器化される可能性があり（そしてほぼ確実にそうなる）、新たな戦争形態につながる可能性がある。AI は現在の経済の多くを破壊する可能性もある。」と主張しています。
実際、Yann Lecunは2019年にStuart RusselやYoshua Bengioと道具的収束について対談した時に、Yann LeCunは道具的収束目標を開発者がAIに持たせることは無いし、自然にそのような動機がAIに芽生えることもないと主張し、MIRIのRob BensingerによってYann LeCunがどのような前提を置いているのかは簡単に分析されているように、基本的にYann LecunはAI Alignment問題をAI存亡リスクを主張する人々と比較した場合、そこまで難しい問題だとは考えていないだろうことから、それよりも現実的なリスクを見るべきと上記記事にて主張しているのだと思われます。

上記Yann Lecunらによる批判をもう少し先鋭化した形で展開している人たちもいます。
それがGoogleを解雇/辞めたAI研究者のTimnit Gebruと人類の絶滅に焦点を当てた研究をしている哲学者のÉmile Torresです。
二人は未発表の共著の学術論文で、ある種のイデオロギーの束である「TESCREAL」という言葉を造語し、それら運動を「Second-Wave Eugenics(優生学の第二の波)」として批判しています。
※TESCREALという言葉自体は恐らく2023年2月10日のIEEE Conference on Secure and Trustworthy Machine Learning (SaTML)の基調講演にて公に初めてされたものと思われます。基調講演の当日の映像もyoutubeにアップロードされています。

TESCREALの「T」はTranshumanismの略で、これはTESCREALのバックボーンとされます。実際、この頭字語の次の3文字「ESC」は、Extropianism、Singularitarianism、Cosmismの略でTranshumanismの変種であり、次の4文字の「REAL」は、Rationalism(合理主義)、Effective Altruism(効果的利他主義)、Longtermism(長期主義)の略で、その歴史的起源は1990年代のTranshumanism運動に関連しているとも言えるでしょう。
概念や運動が広まったおおよその順でTESCREALと命名されているようです。

また、Émile Torresは2023年にある種わかりやすいストーリーとして広まったe/acc(効果的加速主義)とEA(効果的利他主義)の対立に関しても、それはTESCREAL内の家族間の紛争として理解されるべきで、相違点よりもはるかに多くの類似性があると言います。
具体的には「e/accも長期主義者も最終的な目標は何かを最大化することであり、e/acc の場合はエネルギー消費、長期主義者にとってそれはより一般的な意味での価値である」と述べています。
そしてそれらTESCREALの特徴には黙示録とユートピア的信念があり、それら二つの信念は同じコインの裏表だとÉmile TorresやTimnit Gebruは主張しています。

それでは具体的に彼/彼女らはTESCREALの何を批判しているのでしょうか。
大きく分けると優生学的な思想、欺瞞的な態度や不健全な文化、長期主義的な運動に対する批判に分けられると思われます。

・優生学的な思想に対する批判
TESCREALが優生学的だとする根拠はNick Bostrom等の過去の人種差別的発言や知能に関

する遺伝による存亡リスクがBostromにより論じたられたこと、TESCREAL自体がTranshumanism発祥でTranshumanism自体が優生学的であった歴史的背景からの類推から主張しているように思われます。

この批判に関しては優生学的であると主張する論拠が薄く、過去の優生学的な思想や差別が現在のTESCREAL（特に効果的利他主義コミュニティ）にも引き継がれていると考えるのは過度な一般化ではないかと批判があります。

・欺瞞的な態度や不健全な文化に対する批判
また、効果的利他主義コミュニティに寄付を行っていたFTXのBankman-Friedによる破綻事件や寄付に対して贅沢な消費、セクシャルハラスメントに関する告発といった事実等を指摘し、TESCREALコミュニティに蔓延する欺瞞的な態度や不健全な文化も批判しています。

上記批判については詳しくなく断定はできませんが、効果的利他主義コミュニティ全体に当てはまるわけではないと思われます。
FTXの破産、セクシャルハラスメント、EAの資金の使用のされ方についてはコミュニティ内でも問題視されています。

・長期主義的な運動に対する批判
上記二つの批判はより広いTESCREALコミュニティ全体の過去や文化や組織体制に関する批判となっていると思われますが、彼らがその先にTESCREALを批判する主だった主張としてあるのはTESCREAL運動が「実際の人々、つまり現在および近い将来に生きている人々が極度の危害、さらには死に至る可能性を著しく高める可能性があること」だと思われます。

Torres氏はNick Bostromの2002年の存亡リスク論文や2019年の「The Vulnerable World Hypothesis:脆弱な世界仮説」と呼ばれる論文を引用し、以下の記事で存亡リスク低減のために、武力行使、軍事攻撃、標的殺害を長期主義が判断し得る懸念を表明し、それが起こり得る論理を批判しています。

A Code Red Warning about TESCREALism *This is what happens when you spend two years unMasking a wel www.truthdig.com*

Nick Bostromの2002年の存亡リスク論文ではナノテクノロジーを保有する国家間で軍拡競争が不安定になった場合「先制攻撃に対する支配的なインセンティブが存在する」と述べられています。この存亡リスク論文での主な焦点は存亡リスクの種類やその下の具体的な例を包括的に精査することです。つまり、上記の先制攻撃に関する話は言葉として分析的な意味で出てくるのみでした。

一方でNick Bostromの2019年の「脆弱な世界仮説」論文では、文明の潜在的な脆弱性について類型化し、大量破壊を突然民主化する技術(=黒いボール)に対処するため、考えられる対応策を検討する内容となっています。
結論としてはこの論文の考察は、監視能力と予防的取り締まりシステムの強化及び、断固とした行動をとることができるグローバル・ガバナンス体制を支持する理由となる考察を提出しています。

予防的取り締まりでは、その極端な例としてハイテク・パノプティコンと呼ばれる全国民にデジタルセンサーをつけてプライバシーには配慮しながら常時監視するシナリオを考察しています。

またグローバルガバナンスでは、強力なバイオテクロジーの黒いボールを誰でも持てるように

なった世界シナリオで「国家がひとつでも、国民の継続的な監視と管理に必要な機械（あるいは、事実上完璧な信頼性で悪意のある利用を防ぐために必要なその他のメカニズム）を導入しなかったとしたら、それは容認できない。」と書かれています。それほど強い全世界的な国家間のガバナンスが必要になることを示唆しています。

一方で、Bostrom自身もこのような強力な監視を可能にするメカニズムや、いかなる国にもその意思を押し付けることができるグローバル・ガバナンスの機関の存在は専制的な国家のリスクや個人の人権に抵触する可能性を認めており、最初はもう少し部分的な規制に留め、黒いボールが作れる資材を追跡したり、施設のセキュリティを強化し、Differential Technological Developmentで対応するのが賢明かもしれないと論文で述べています。

※Differential Technological Development(差分的技術開発)とは危険で有害な技術、特に存続リスクのレベルを高める技術の開発を遅らせ、有益な技術、特に自然や他の技術によってもたらされる存亡リスクを軽減する技術の開発を加速することです。

しかし結論の最後で

「それら(差分的技術開発や一般的なセキュリティ対策)が提供する保護はシナリオの特殊な部分集合にしか及ばず、一時的なものかもしれないことを念頭に置くべきである。
予防的な取り締まり能力やグローバル・ガバナンス能力のマクロパラメーターに影響を与える立場にあることを知ったなら、これらの領域における根本的な変化が、新たな技術的脆弱性に対して文明を安定させる一般的な能力を達成する唯一の方法かもしれないことを考慮すべきである。」

と書かれています。
つまり、これはNick Bostromはあくまで部分的な規制については黒いボールに対する一時的な処置かもしれず、予防的な取り締まりや強力なガバナンスの検討をした方が良いと示唆しているのかもしれません。

Torres氏は上記のような考察に対して、「Here Be Dragons」というスウェーデンの学者Olle Häggströmによる本の以下のような一節を引用し、懸念を表明しています。

> CIAはアメリカ大統領に、ドイツのどこかに終末兵器を研究している狂人がいて、人類を絶滅させるためにその兵器を使うつもりであり、その狂人が成功する確率は100万分の1であるという確かな証拠があると説明する。この狂人の正体や居場所について、それ以上の情報はない。もし大統領がBostromの議論を真に受け、計算の仕方を知っていれば、ドイツに全面的な核攻撃を行い、国境内の人間を一人残らず殺す価値があると結論づけるかもしれない。
>
> Here Be Dragons

そしてTorres氏は上記のような懸念が現実のものになりつつあると感じているようです。

> TESCREAL運動は絶大な力を持つようになった。外交政策界や国連のような主要な統治機関に浸透し、数百億ドルの資金を背景に、シリコンバレーに浸透し、ソーシャルメディアで多くのフォロワーを持つ人々によって推進されている。例えば、Elon Muskは長期主義を「私の哲学に近い」と呼び、昨年は宇宙を植民地化した場合、どれだけのデジタル人間が存在しうるかというBostromの論文へのリンクをリツイートした："おそらく、これまで書かれた中で最も重要な論文だろう"。TESCREAL運

> 動は世界的な勢力となり、勢いを増している。
> (中略)
> サンフランシスコのベイエリアでカルト的な人気を誇るTESCREAL運動の中心人物、Eliezer Yudkowskyによる最近の『TIME』誌の記事を考えてみよう。Yudkowskyは、われわれはAGIを創り出そうとしているのかもしれないと主張し、もしわれわれが「現在のような状況下で」AGIを創り出せば、「最も可能性の高い結果」は「文字通り地球上のすべての人が死ぬ」ことになるだろうと述べている。全面的な熱核戦争が起きても、おそらく地球上のすべての人が死ぬことはないだろう-科学がそれを裏付けている-から、彼は、"完全な核交換 "を引き起こす危険を冒してでも、AGIを開発しているかもしれない国に対する軍事攻撃を承認する国際条約に各国が署名すべきだと主張する。

https://www.truthdig.com/articles/before-its-too-late-buddy/

つまりTorres氏（恐らくTimnit Gebru氏も）はAIや人工的なパンデミックによる存亡リスク(どちらも80000hoursで最も差し迫った効果的に取り組める問題リストにおいて上位2つとなっている)に対処するという大義名分により、極端な判断（例えば多くの人がその判断と引き換えに死亡するような判断）が行われることを懸念しているのではないかと思われます。

長期主義批判に対するEAの反応

前節で効果的利他主義コミュニティへの批判を上げてきましたが、その中でも主に長期主義的な側面の批判が主なものだと思われます。

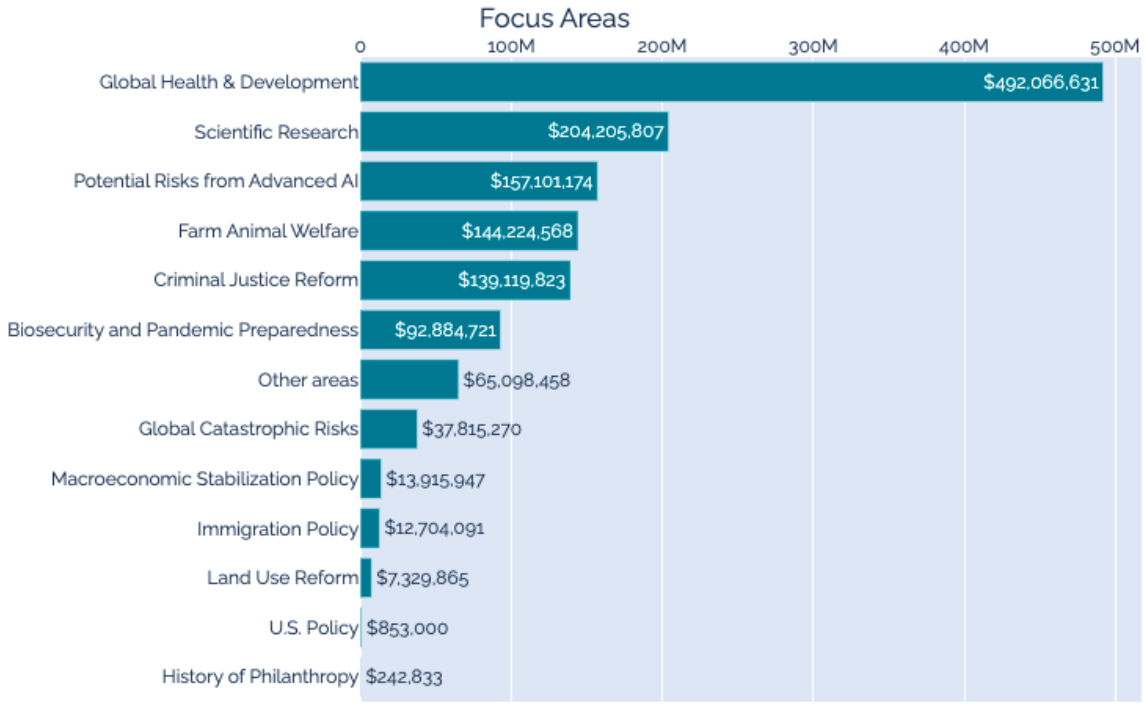
それではそもそも効果的利他主義コミュニティはどの程度長期主義に偏っているのでしょうか。

- 助成金データでは長期主義一辺倒とはなっていない

以下の「EAは長期主義以上のもの」という記事が参考になります。

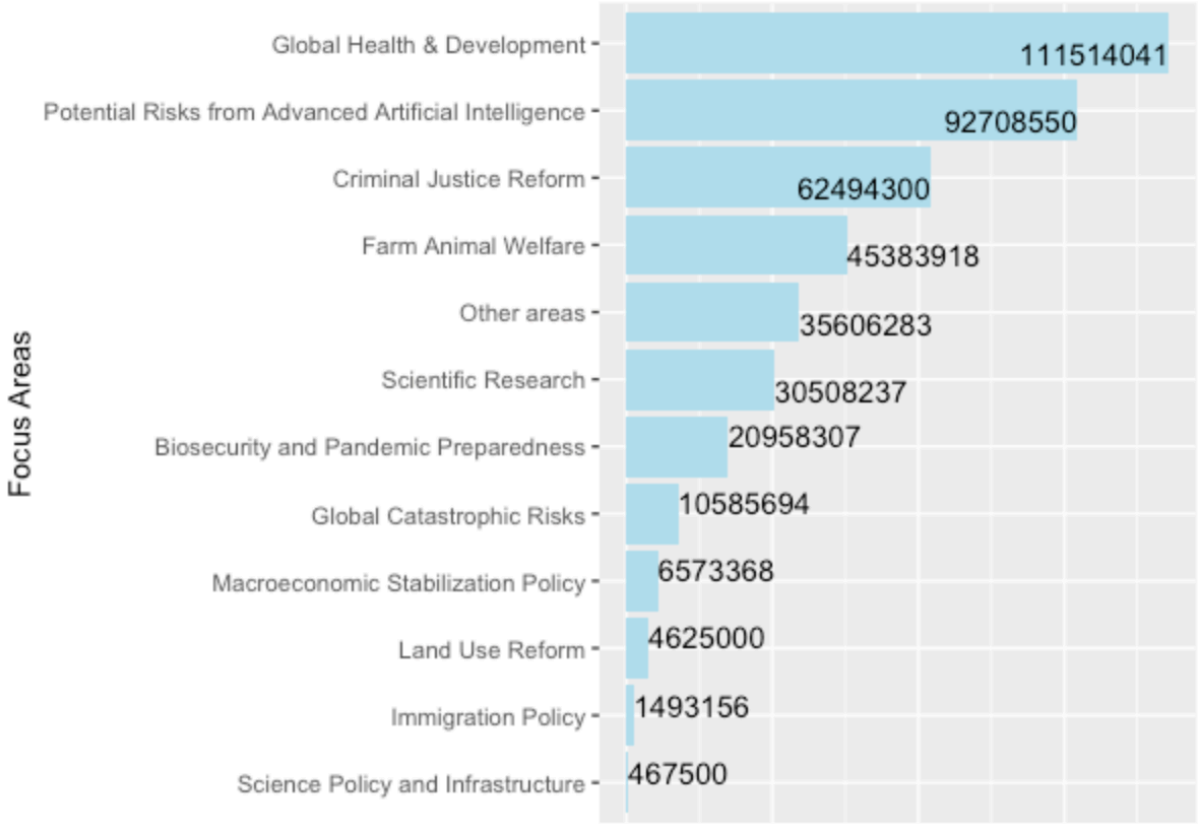
EA is more than longtermism — EA Forum *Comment by AnonymousEAForumAccount - For me, it's bee forum.effectivealtruism.org*

効果的利他主義コミュニティの活動資金の60%がOpen Philanthropyから提供されているため、Open Philanthropyの助成金データを見ればおおよその活動資金の内訳はわかると仮定します。
以下のグラフ（Effective Altruism Dataより）は、2012年以降のオープンフィランソロピーの助成金総額 を分野別に示したものです。



2012年以降のOpen Philanthropyの助成金支出（分野別

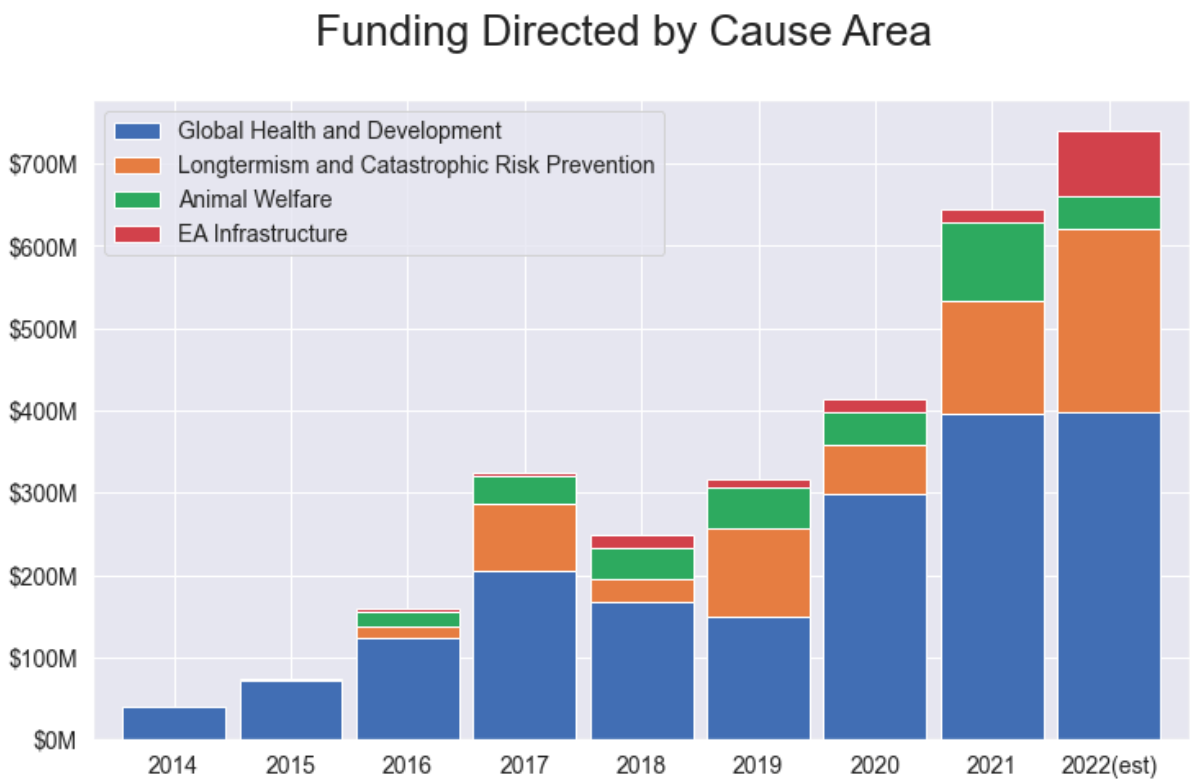
全体として、Global Health & Development(マラリア、結核、下痢、寄生虫病の予防と治療)が投下資金の大半を占めています。一方近年、AIの安全性への関心が高まっているため、2021年1月から現在までのOpen Philanthropyの助成金の推移を見てみましょう。



2021年1月から現在までの慈善活動分野別のOpen Philanthropyの助成金支出

Global Health & Development が依然として主要な資金受取ドメインであることがわかります。しかし、現在では「高度なAIによるリスク」が僅差で2位となっています。一方3番目と4番目に資金の多い分野である刑事司法改革と家畜福祉は、主に長期的な将来に影響を与えるという目標によって推進されているわけではないことにも注目できます。

また以下の記事ではOpen Philanthropy以外の助成金も含めたカテゴリー別のEAへの助成金年次データです。



受給者カテゴリー別の公的に利用可能な助成金

このグラフからもGlobal Health and Developmentが最も多い使途なことがわかります。

結論としては、高度なAIによるリスクの低減を含む長期主義に関連する資金が増加していることがわかりますが、効果的利他主義コミュニティを長期主義のみに注力している組織というのは誤りだと言えるでしょう。

- 長期主義の問題への応答

今まではどちらかというと周縁的な議論に終始していましたが、本質的にÉmile TorresやTimnit Gebruがしたい批判は長期主義そのものの危険性だと思われます。そのため、次に長期主義への批判的な言説そのものへの応答についてみていきます。

※長期主義への批判的な言説への応答について私が調べた限りのあくまで例示となります。専門的な哲学の知識はないため、参考程度に見ていただけると幸いです。以下の記事はTorres氏の長期主義への批判に対する応答となっています。

[Response to Torres' 'The Case Against Longtermism' — EA Forum *This short post responds to some of the criticisms of longter forum.effectivealtruism.org*]

様々な論点でTorres氏の意見が批判的に考察されています。重要だと考えられる論点「総量功利主義」と「監視の正当化」と「大量殺人の正当化」を紹介していきます。

> 総量功利主義
> Torres氏は、長期主義は総量功利主義、つまりグループ内のすべての個人の幸福を合計することに基づいて幸福を最大化すべきという考え方の倫理的仮定に基づいていると示唆しています。このような「多ければ多いほど良い」という倫理観は、何

兆もの将来の個人にとって重要な意味を持ちます。彼は、総量功利主義は道徳哲学者の間で多数派の意見ではないと指摘する。

ただし、総量功利主義は長期主義を強力にサポートしますが、長期主義は必ずしも総量功利主義に基づく必要はありません。『Precipice』の成果の一つは、長期主義と、保守主義、過去への義務、美徳倫理などの他の倫理的伝統との親和性を指摘したToby Ordの議論である。人はさまざまな倫理観にこだわり、長期主義を支持することができます。

Response to Recent Criticisms of Longtermism

監視の正当化
Torres氏は、Bostrom氏の「脆弱な世界仮説」で検討されている「ターンキー全体主義」（先端技術の悪用を防ぐための大規模で侵入的な大衆監視と統制）に反対し、長期主義がそのような政策にコミットしていることを示唆している。

ただし、長期主義はそのような提案に固執する必要はありません。特に、Bostromが間違った脅威モデルを持っていると単純に反論することができます。私たちがこれまで直面してきた存亡リスク（核兵器や生物兵器、気候変動）は主に国軍や大企業から来ており、間もなく直面するかもしれない存応リスク（新しいバイオテクノロジーや革新的AI）も同じ脅威源から来るだろう。したがって、存亡リスクの予防の焦点は国家と企業にあるべきである。個人や小グループによるリスクは比較的小さいです。Bostromが模索しているような大規模監視から得られるこうした小さな利益は、それが費用便益分析によって正当化されないことを意味する。

それにもかかわらず、「電子レンジを持っている人は誰でも核兵器を所有できる」という人為的な仮説では、長期主義は自由の制限にコミットするだろうか？これについては次の見出しで説明します。

Response to Recent Criticisms of Longtermism

大量殺人の正当化
Torres氏は、絶滅を防ぐためなら、長期主義者は恐ろしい行為（例えば、核兵器でドイツを破壊すること）も厭わないはずだと主張する。

これは、トロッコ問題からコロッセオ論争に至るまで、あらゆる形の結果主義と功利主義に対する古典的な反論である。仮説に異議を唱えるものから、他の倫理観もそのような行為にコミットしていることを指摘するものまで、多くの古典的な回答がある。

これは長期主義に特有の反論ではないし、長期主義が功利主義に基づくものである必要がない以上、その威力は失われる（上述したように）。また、長期主義が平和、軍縮、大災害の回避に高い優先順位を置いている以上、このような非難は奇妙であることも指摘しておきたい。

Response to Recent Criticisms of Longtermism

上記3つの論点は長期主義が単純な功利主義で捉えられるものではなく、そしてトロッコ問題のような難しい判断を持ち出して長期主義を批判することは、一般的な倫理の難題を長期主義に

帰すことで批判しているため、長期主義に特有の批判になっていないと応答しているものと思われます。

確かに上記記事内にあるように、理論的にも実際的にも現状長期主義は上記のような大量殺人や監視を正当化したりしないし、功利主義とイコールとは言えないかもしれません。

一方で次に問題として、「将来の長期主義者」を批判者は挙げるかもしれません。
実際には長期主義に賛同していない人が、明らかに不誠実な方法で長期主義の考え方を隠れ蓑として利用することはあり得ます。
将来的に、危害を正当化するために長期主義が使われるのを防ぐような構造的あるいは哲学的特徴はあるのでしょうか？

下記長期主義批判に応答する記事の「制約条件:Limiting Conditions」内にて上記疑問への応答がされています。

> これは長期主義に特有の特徴ではないようで、非常に多くのイデオロギーがこのように使用される可能性があるため、長期主義に強く反対する必要はないようです。
> （中略）
> しかし、より心配なのは、長期主義を注意深く考え、理解し、その結論が有害なことをするよう指示していると信じている人の例である。イデオロギー、とりわけユートピア的なイデオロギーは、近年の歴史において最も残虐な行為のいくつかを引き起こしてきた。長期主義が他のイデオロギーよりもその影響を受けやすいかどうかについては、未解決の問題がある。
> （中略）
> Torresは、このような懸念は長期主義者たち自身が提起し、議論し、執筆してきたものであることを無視している。なぜ期待値を真に受けすぎてはいけないかについての論文も ある（前者はBostromによるもので、Bostromが素朴かつ危険にも期待値を使って非常識な行動を正当化したというトーレスの非難を和らげるかもしれない。）
> ファナティシズム（狂信）の問題について書かれた論文は複数ある（Torresは、これは「一部の」長期主義者が受け入れている用語であると言っているが、彼は一つの単著の学術哲学論文を引用しており、ファナティシズム(狂信)を深刻な懸念として指摘している複数の 論文については言及していない）[18]。
> 複数の主要な長期主義者団体のウェブサイトには、効果的な利他主義者や長期主義者が、称賛に値する目標を追求する際にも、なぜ危害を及ぼしてはならないかについての複数の 論文が掲載されている。また、道徳的不確実性を強く強調することで、複数の見解から見て強固な善ではない極端な行動に向かう傾向を和らげている。

Response to Recent Criticisms of Longtermism

将来的に長期主義が深刻な危害を正当化する可能性については、他の多くの倫理的立場でもその立場が利用される可能性があるため、長期主義特有の批判にはならないとは言いつつ、長期主義自体が他のイデオロギーよりも残虐な行為の正当化の影響を受けやすいかどうかは未解決の問題としています。

一方、現状のところ長期主義の元になったNick Bostromや主要なEAのメンバーの多くは長期主義批判で取り立たされている問題を彼ら自身で十分認識していると思われます。

AIによる存亡リスクの説明からその背景とそれを取り巻く考え方への批判や議論を紹介させていただきました。これらの動きはこの20年間水面下で続いていたものですが、この数年で予想以上に早いAIの進歩により、存亡リスクに関する議論が一般化しつつあります。

今後近い将来、高度なAIや他の合成生物学等の先端テクノロジーが発展する中でどのように我々人間がそれらと向き合い、どこまでの規制が許されるのかについて議論が加熱していくでしょう。
長期主義や効果的利他主義といった考え方や運動に限らず、技術がもたらす可能性とリスクを深く理解し、社会/世界全体での協力と議論をすることが不可欠になってくると思われます。

AI Alignment研究

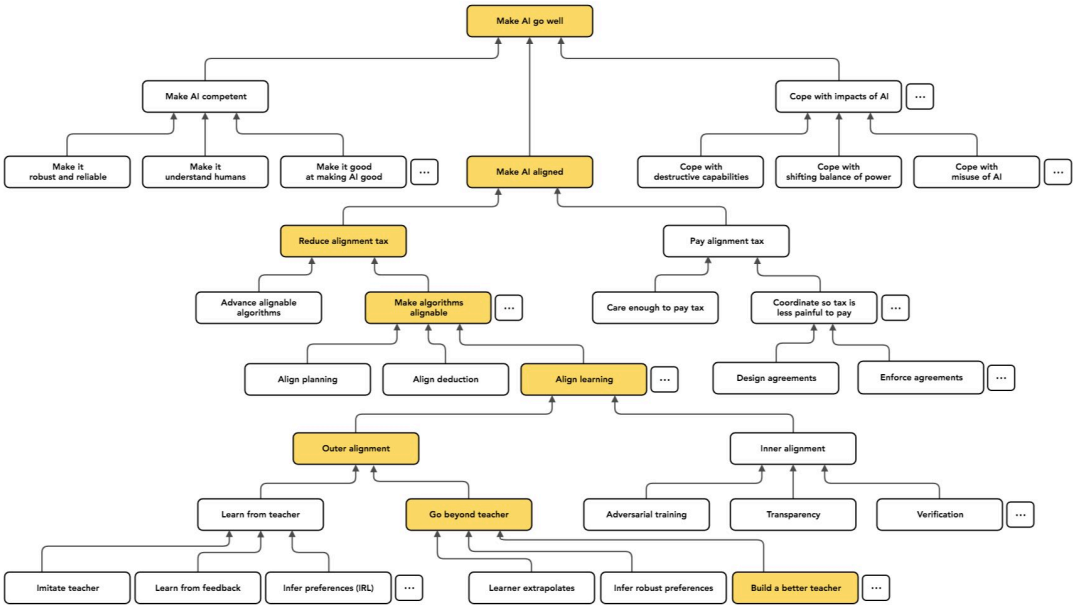
> *The future is going to be good for the AI regardless.*
> *It would be nice if it were good for humans as well.*
> いずれにしてもAIにとって未来は良いものになるでしょう。
> 人間にとっても良いものになるといいのですが。
> Ilya Sutskever

[Ilya:the AI scientist shaping the world]

AI Alignment/Governance概要

これまでAIによる存亡リスクの概要からその歴史的背景まで見てきました。
それではその存亡リスクに対処する具体的な方法はないのでしょうか？

AIによる存亡リスクに対処するための取り組みの全体像を把握するためには、以下のPaul Christianoの「AIに良いことをさせる分野の見取り図」は参考になると思われます。



AI Alignmentとその周辺の見取り図

上記見取り図の一番上部にとにかくAIに良いことをさせる（Make AI go well）という目標が掲げられています。

その下に「Make AI competent(AIの能力を向上させること)」と「Make AI aligned(AIアライメント問題の解決)」と「Cope with impacts of AI(AIガバナンスを含む広範な対応)」が分類されます。

・Make AI competent(AIの能力を向上させること)
まず AIに「良いこと」をさせるためにはAIの能力(Capability)を向上させる必要があります。当たり前ですが、自動運転車がベビーカーが目の前にあるのに突進したら大惨事になってしまうため、現実世界に対する常識的な理解や人間の意図した命令を理解する能力が必要になります。この能力が欠けていると、さまざまな問題が引き起こされます。

・Make AI aligned（AIアライメント問題の解決）
次にAI Alignment問題に対処することです。
ここで、AI Alignment問題とはAIシステムが、意図しない望ましくない目標ではなく、人間の価値観や関心に合った目標を追求するようにするという課題です。
つまり、AI Alignment問題で懸念されているのはAIの能力が低いから起こる事故のリスクではなく、AIの目標がわれわれの目標とずれていることで起こる事故のリスクとなります。
AIアライメント問題を解決しなければ、もしかしたらAIは能力が高い一方で、人間を欺こうとする可能性があります。
自動運転車の事故のように能力が低かったり人間の意図したことを理解できないため起こる事故ではなく、AIが意図的に人間社会に悪影響を及ぼす可能性があり、この問題に対処する必要性があります。

・Cope with impacts of AI（AIガバナンスを含む広範な対応）
最後にAIアライメント問題の解決が難しい場合でも国際的に連携して高度な AIに対する規制の取り組みを行う必要があります。例えば悪用や誤用のリスクにAIガバナンスで対処しなければなりません。

AIの能力向上についてはAI研究の主流のモチベーションです。
一方でAI AlignmentやGovernanceにフルタイムで取り組んでいる人口は相対的に少ないことが知られており、また前の章で解説してきた通り、潜在的に人類に壊滅的な影響を与える可能性があるため、これら分野に注力することが重要となってきています。
そこで、この章では現状あまり注目を浴びていないAI Alignment問題に対する技術的な研究の概要紹介とAIガバナンスの取り組みを紹介します。

AI Alignmentとは何か

AI Alignment問題とはAIシステムが、意図しない望ましくない目標ではなく、人間の価値観や関心に合った目標を追求するようにするという課題です。

AI alignmentと聞くとOpen AIの2022年に提出された言語モデルの微調整手法であるRLHF論文を思い浮かべる方が多いかもしれません。

Aligning language models to follow instructions *We've trained language models that are much better at followi openai.com*

そのためAI Alignmentを偏見や悪用を抑えるために人間にとって好ましい出力を行わせる一つの技術的な手法と感じている方も多いと思われます。

その理解もある意味では正しい一方で、AI Alignment分野が創始された際のモチベーションそのものは主に有害な結果ではなく、有益な結果を生み出すSuperIntelligence(超知能)であるFriendly AIを分析したEliezer Yudkowskyの Creating Friendly AI論文(2001)から始まりました。

そのためAI Alignmentには勿論悪意ある差別的発言をさせないといった研究目標も含まれ得ますが、その分野自体は自動運転の安全性も含むより広範なAI Safety分野より狭く、ある種人間よりも高度なAIをどのように安全に取り扱えるのかというモチベーションから生まれた分野と言え、今後はそのような高度なAIをどのように人間の意図した目標に合わせられるかといった文脈でAI Alignmentという言葉が使われる機会が増えると思われます。
※AI Safety,AI Alignment,AI倫理,AIコントロール,AI ガバナンス等の単語の整理をしたページはこちら。

もう少し形式的な定義もあります。
AIであるAがオペレータであるH(Human)にalignedされているとは
「AがHの望むことをしようとしていること」
とAI Alignment研究者のPaul Christianoは定義しています。

ここではどんな価値観をAIに実装するのが良いことなのかといった問題とAI Alignment問題が切り離されています。

これは人間の価値観でどのようなもの（功利主義、カント主義）を実装するのが良いのか、そもそも人間の幸福とは何かといった倫理的な問題とは別に、技術的にそもそものところAIに人間の意図した目標を行わせることが可能なのか？可能だとしたらそれはどのように技術的に実現できるのか？を考える必要があるためです。

そのため、そもそもAI Alignment問題を解決しないと、AIを使用する人間の善意や悪意は関係なく、壊滅的な結果が生じる可能性があります。
また人間の価値観を理想的な形で表現できたとしても、そもそもAIが人間の意図した命令に従わなければ意味がありません。

よってAI Alignment問題の解決はAIによる存亡リスクを解決するための必要条件であり、AIによる存亡リスクを解決するための最もコアな技術的課題として現状立ちはだかっているといえるでしょう。

一方で分野として若いということもあり、上記Paul Christiano氏のAI Alignmentの定義については人によって使い方が異なり、今も議論がされています。

厳密にはどのような価値観を選定するかといった観点と内容に依存しない目標にアライメントするという技術的な問題は分離できず、集団における価値とは何かも含めて技術的なアライメント研究に含める必要性が議論される場合もあります。

これは、AI Alignmentという専門用語自体はEliezer Yudkowskyが2001年に提唱したFriendly AIという言葉の変わりに、2014年頃からStuart Russelから提案され、2014年に論文で言及されてからまだ10年程度しか経っていない状況を端的に表していると言えるでしょう。

つまり、AI Alignmentという言葉もしかり、AIの目標をどのように人間の意図した目標に整合(Aligned)させるかも不明瞭のままで、ある種この分野自体が17世紀の物理学におけるエネルギーという重要な概念を曖昧にしたまま研究が進んでいる状況と似ているとも言えるでしょう。

そのためちゃんとしたAI Alignment研究分野の見取り図は恐らく確立されておらず、散発的に理論的・実験的アイディアがそれぞれ並行してプロジェクトとして動いている状態だと考えられます。

よってここでは大雑把にAI Alignment研究分野を実証的/理論的/概念的研究に分類して紹介したいと思います。この分類法は下記の記事を参考にしました。

A newcomer’s guide to the technical AI safety field — LessWrong *This post was written during Refine. Thanks to Jonathan www.lesswrong.com*

※AIアライメントについての歴史や起源については以下の記事でも説明しています。
https://www.aialign.net/blog/20240620-bioshok

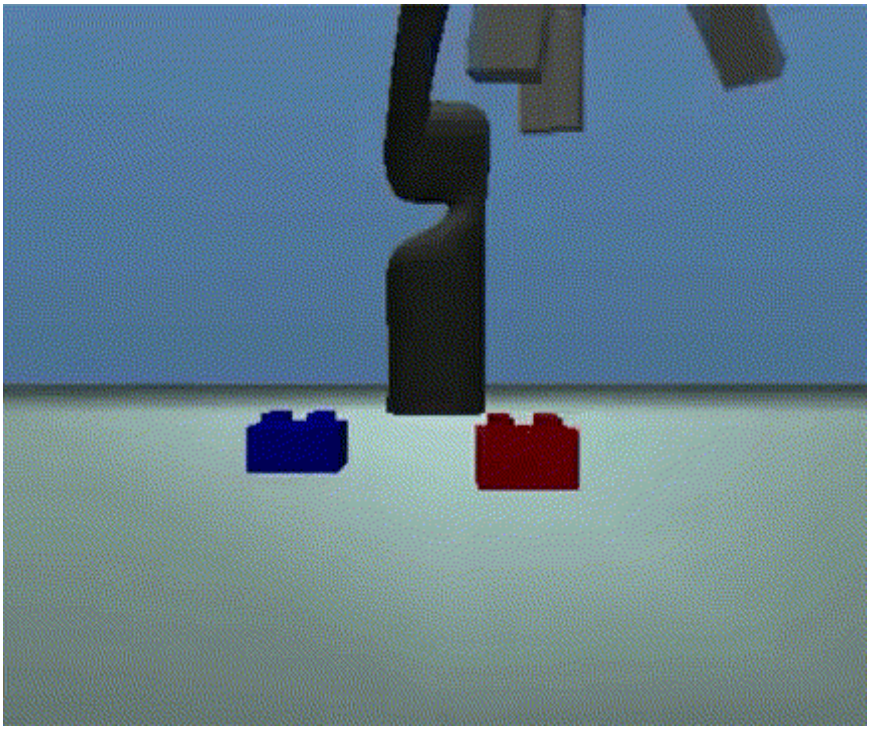
AI Alignmentの実証的な研究

ここでAI Alignmentの実証的な研究というカテゴリーで、現在のAIシステムで実際に実験を行っている研究を指したいと思います。これらには機械学習 (ML) モデルのトレーニングに関するより多くの実践的な作業が含まれています。
※稀に現状のAIシステムの延長線上のシステムをAlignすることをProsaic AI Alignmentとも呼ぶようです。

主に報酬の設計ミスによるSpecification Gamingに対処する研究(Outer Alignmentに関する研究)と新しい環境下で思ったように汎化しないGoal Misgeneralizationに対処するための研究(Inner Alignmentに関する研究)で分けられます。

●Specification Gaming(Outer Alignment)

Specification Gamingとは、意図した結果を達成することなく、目的の文字通りのSpecification(仕様)を満たす行為です。
例えば、以下のGif画像のようなレゴの積み上げタスクで、赤いブロックが青いブロックの上に来ることが望ましい結果でした。
しかし、エージェントは、赤いブロックをひっくり返すことで、底面の高さが青いブロックの高さを超えることで報酬を受け取ります。



Source: Data-Efficient Deep Reinforcement Learning for Dexterous Manipulation (Popov et al, 2017)

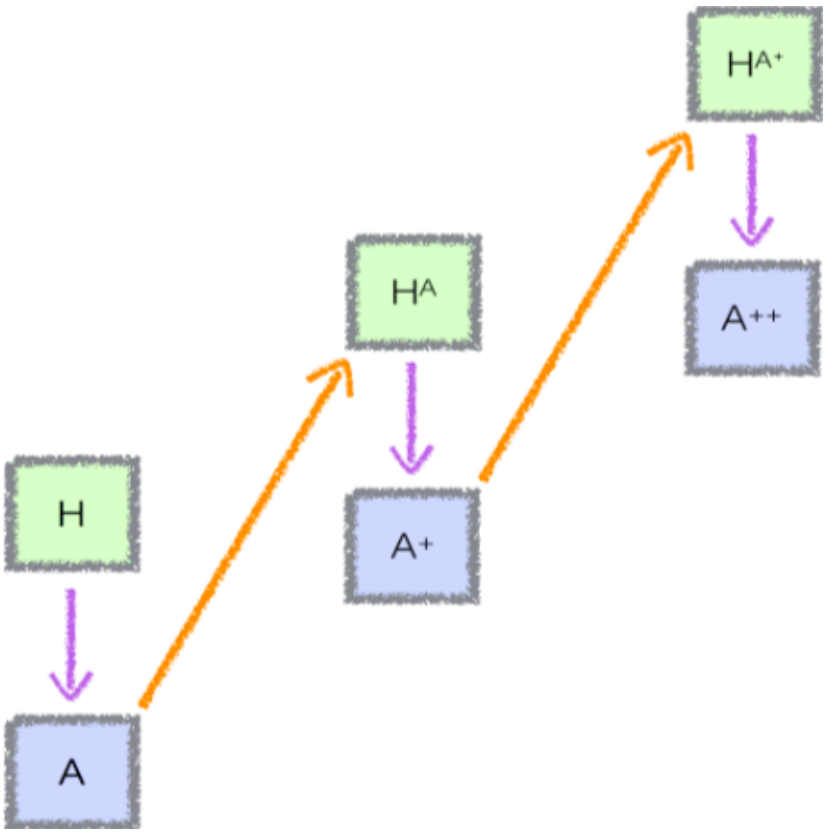
さまざまなタイプのSpecification Gamingを分類し、エージェントの能力に応じて仕様ゲームの程度がどのように増加するかを定量化することに関しては、ある程度の進歩が見られます。

そしてSpecification Gamingを防ぐよく知られた手法は、ヒューマンフィードバックからの強化学習(RLHF) やInverse Reinforcement Learning(逆強化学習)と呼ばれるものがあります。
しかし、これらは人間のバイアスや盲点を利用してより高い報酬を得る政策を強化する可能性もあるでしょう。

この問題に対処する有望なアプローチはスケーラブルな監視です。
これはAIを活用して評価が難しい領域まで人間の監視を拡大することで、AlingnedされたAIシステムをトレーニングするための手法です。
スケーラブルな監視には反復蒸留増幅法(IDA)と議論(debate)と呼ばれる手法が知られています。

- 反復蒸留増幅法

スケーラブルな監視の基本的な手法の一つは、AI の支援を受けて人間の判断を再帰的に増幅する反復蒸留増幅法(IDA)です。
まず、エージェントAが人間Hの判断を模倣し (蒸留ステップ)、次にこのエージェントを使用して次のレベル (増幅ステップ) で人間の判断を支援し、増幅された人間 H^Aが生成されます。この再帰的なプロセスは、人間の監督者がタスクを分解してタスクの一部を AI アシスタントに委任できる限り、原理的には人間の判断をあらゆる領域にスケールアップできます。

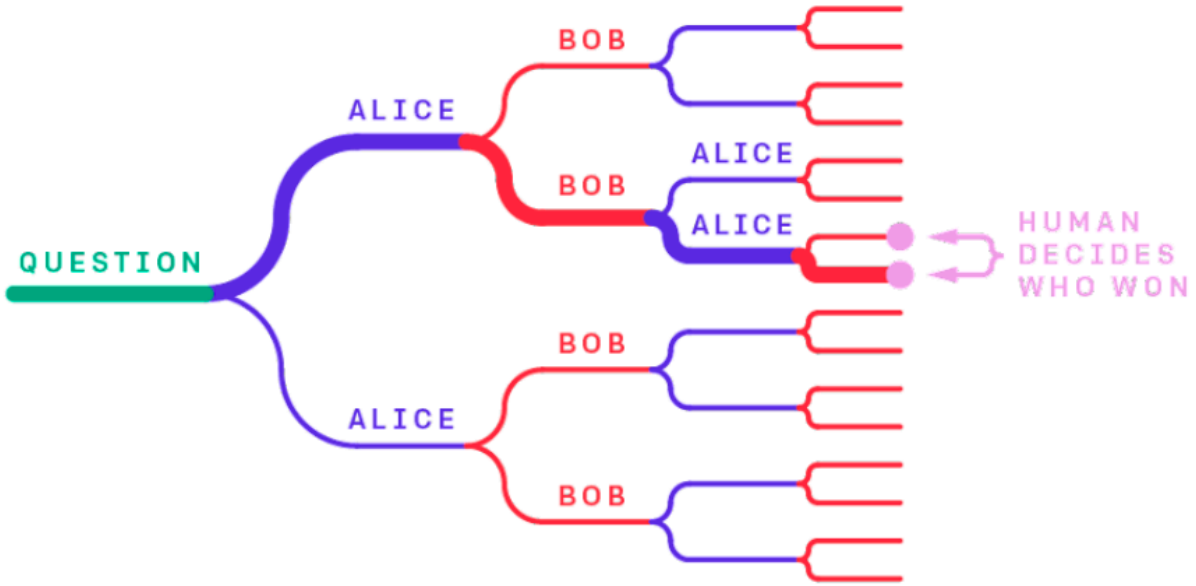


Supervising strong learners by amplifying weak experts, Christiano et al (2018)

- 議論(debate)によるAI安全性

関連する提案は、議論(debate)によるAI安全性です。
これは、言語モデルの増幅を実装する方法と見なすことができます。
ここでは、人間の裁判官が質問の妥当性を判断するのを助けるために、2つのAIアリスとボブが

互いに議論をしています。AIには、互いの議論の欠陥を指摘し、複雑な議論を裁判官に理解できるようにするインセンティブがあります。ここでの重要な前提は、虚偽よりも真実について議論する方が簡単であるため、真実を語る議論者が有利であるということです。



AI Safety via Debate, Irving and Amodei (2018)

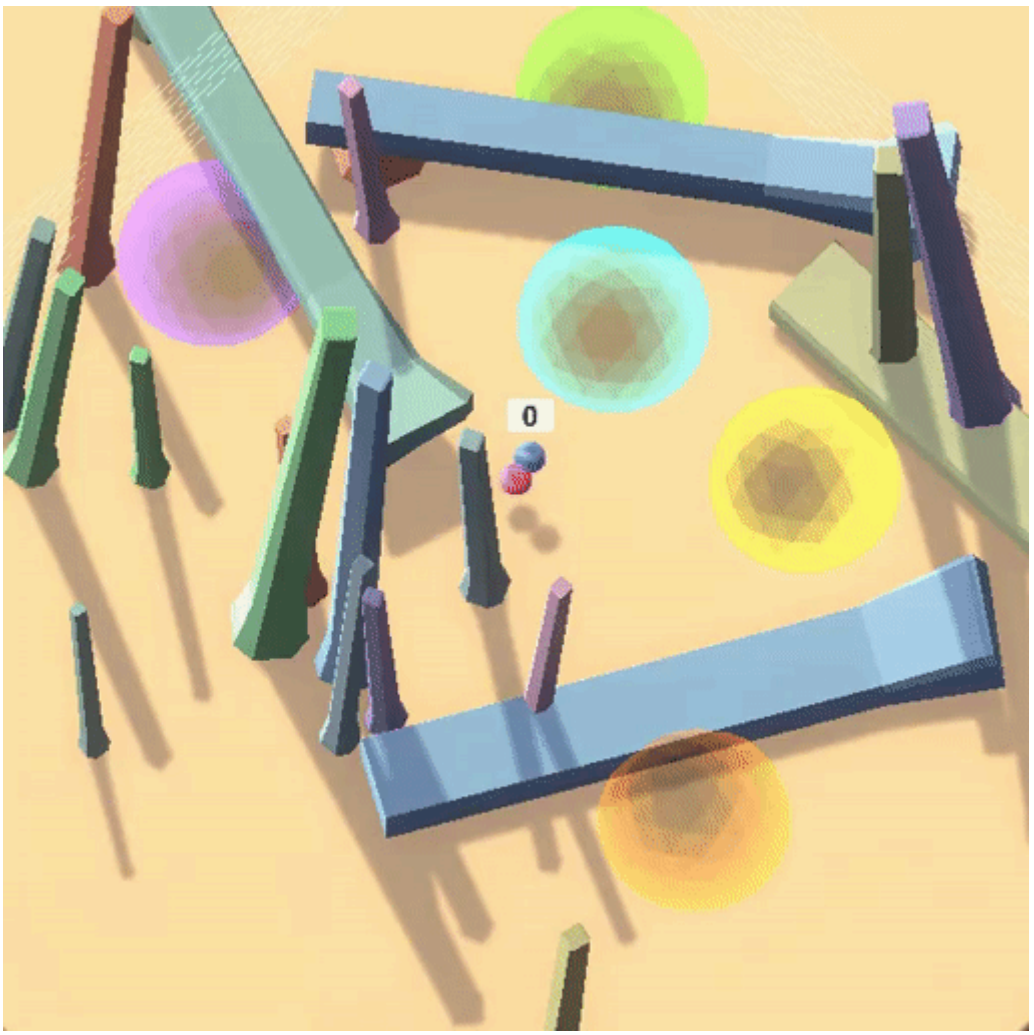
これらScalable Oversight(一人の人間では評価するには複雑すぎるタスクをAIに助けてもらい高度なAIを人間が監視できるようにする)はOpen AIやPaul Christiano氏のAI Alignmentの研究の方向性の一つになっています。

●Goal Misgeneralization(Inner Alignment)

次に、Goal Misgeneralizationに関する研究を見ていきます。
AIシステムが学習データの分布の外でデプロイされた場合、別の目標を学習し、その目標を適切に追求する可能性があります。Goal Misgeneralizationとはシステムの機能/能力は一般化し有能ですが、その目標が一般化しない問題を指します。

例えば、以下の画像では3Dゲーム内のエージェントが正しい順序で球体を訪れるゲームをプレイしています。訓練中に青色のエージェントは正しい順序で球体を訪れる赤いエージェントに追随することを学習します。

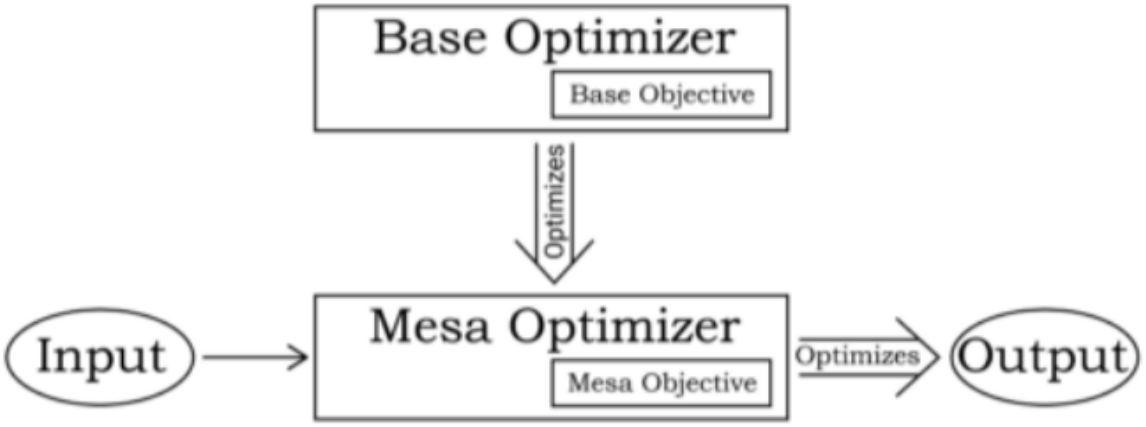
一方で、テスト環境では赤いエージェントは代わりに、間違った順序で球体を訪問した「反専門家」でした。そのため有能に負の報酬を貯め続ける目標に誤って一般化した青色のエージェントができてしまいます。



Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals

ここに多くのGoal Misgeneralizationの具体例がGoogleによって集められています。

このGoal Misgeneralizationの1つのタイプは学習された最適化によるリスクです。AIシステム のベース オプティマイザーは、意図しない目標 (メサ目標) に従っている可能性があるメサオプティマイザーを実行することを学習します。



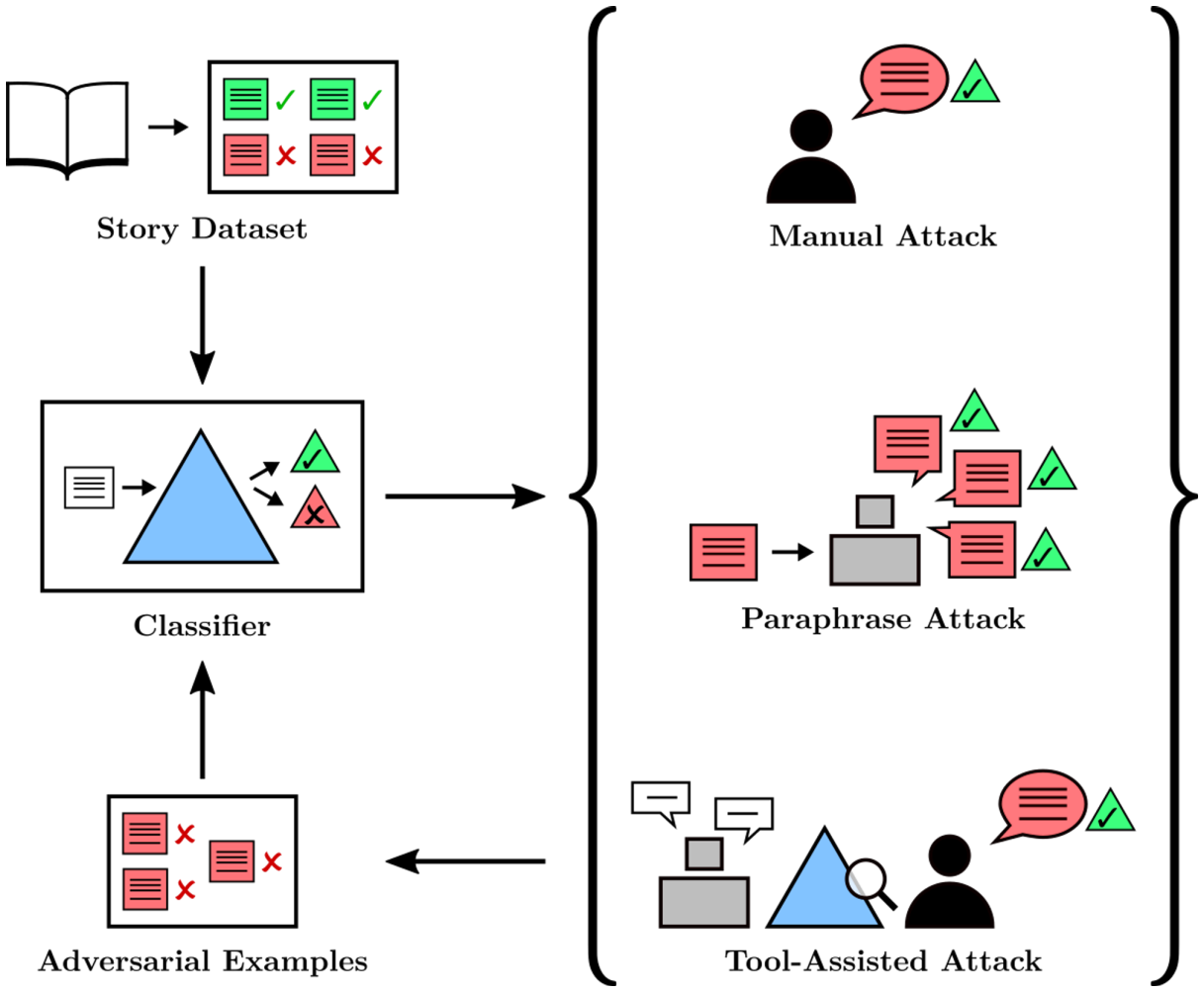
Risks from Learned Optimization in Advanced Machine Learning Systems

Mesa Optimizerによる最適化が起こる仕組みについてはGradient Hacking(勾配ハッキング)と呼ばれるメサオプティマイザーが、勾配降下法によって特定の方法で更新されるように意図的に動作する可能性を指すために使用される用語が導入され概念的なレベルで考察されていました。
一方で近年、Mesa OptimizerがTransformerで実装されることを示唆する論文もあり懸念が強まっています。

Goal Misgeneralizationは未解決の問題ですが、より多様なトレーニングデータを与えること、複数のAIモデルをアンサンブルして異なる結果の場合は警戒する、帰納的バイアスと一般化の理解、機械論的解釈可能性を進めるなどの研究の方向性が考えられています。

- 敵対的学習

例えば、以下の図のように、より多様なトレーニングデータを与える場合、自然言語で書かれたストーリーの中で誰かが傷つくことを検知できるように識別器を訓練する際、敵対的訓練を行います。その際に、人間、ツールによるパラフレーズ、ツールアシストされた人間による敵対的なデータを生成します。



Adversarial Training for High-Stakes Reliability

結果としてより安く高速に識別器を訓練することが可能になり、敵対的堅牢性が上記論文にて上昇しました。

他にも敵対的な訓練を用いた様々な手法が以下の記事で網羅されています。

AI Safety 101 - Chapter 5.2 - Unrestricted Adversarial Training — LessWrong *Introduction This text is an adapted excerpt from the 'Advers www.lesswrong.com*

一方、Goal Misgeneralization、Inner Alignmentで特に懸念されるケースは、望ましくない目標を追求するだけでなく、その行動が設計者の意図と一致しないことをモデルが「知っている」ため、その事実を設計者から隠す欺瞞的なモデルを学習する場合です。
欺瞞的なモデルをターゲットにした潜在的な緩和策には、ニューラルネットワークの内部重みを解釈可能ツールを使用して分析し、欺瞞等望ましくない意図を検出する必要があります。

- 機械論的解釈可能性

そこで必要になるのが機械論的解釈可能性(Mechanistic Interpretability)と呼ばれる研究分野です。この分野は、個々のニューロンのレベルでネットワークを理解することを目的とした研究分野です。ニューロンを理解すると、ニューロンがますます複雑な表現をどのように構築するかを特定し、ニューラルネットワークがどのように機能するかをボトムアップで理解できるようになります。ある意味、中心的なアライメントの問題は、ネットワークが実際に何を学習するのかがわからない

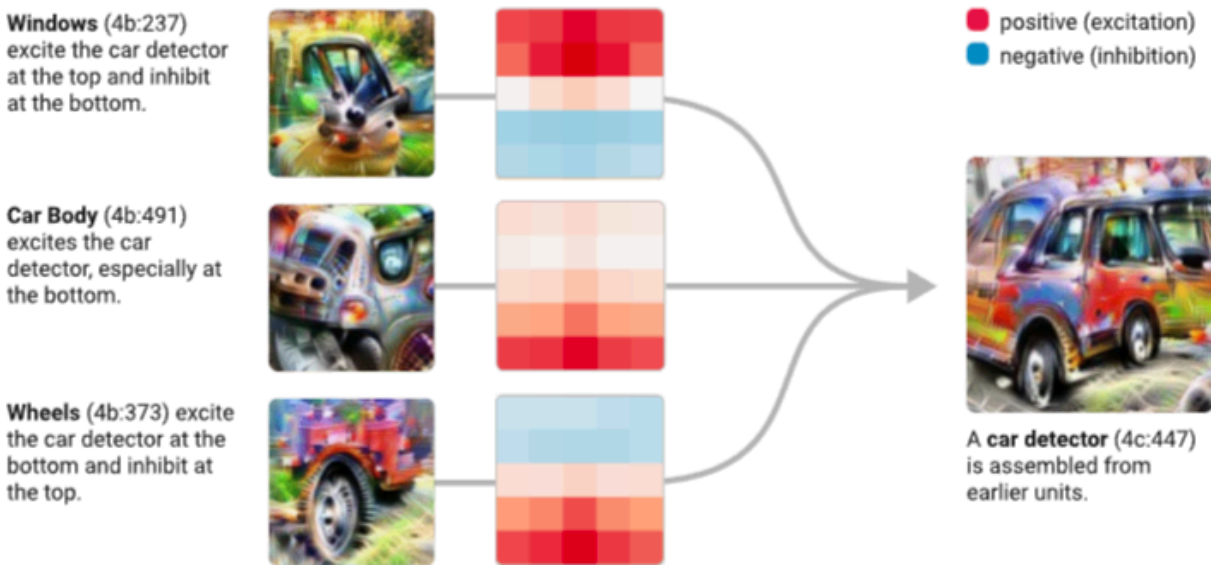
という事実から生じています。機械論的解釈可能性の研究が成功すれば、ネットワークが何をしているのか、そしてそれをどのように変更するのかをより深く理解できるようになるでしょう。

日本語記事だとこちらの記事が入門的です。

※機械論的解釈可能性という和訳
Mechanisticという単語の和訳は機械的、機械論的、機構的がありましたが、より直訳的な機械的、もしくは機械論的の二つの候補にまず絞り込まれました。その後ニューラルネットワークのリバースエンジニアリングを行うといった研究の手法を加味し、「機械的」だと、「自動化」の意味を感じてしまい本来のニュアンスとずれるため、機械論的という言葉に訳されたと聞いています。
最終的に自動的にAIを使ってAIの振る舞いの理解をすることが目指されるかもしれませんが、より分野の実体に近い「機械論的」と訳されました。一方この和訳になるかはまだ不透明なので注意が必要です。

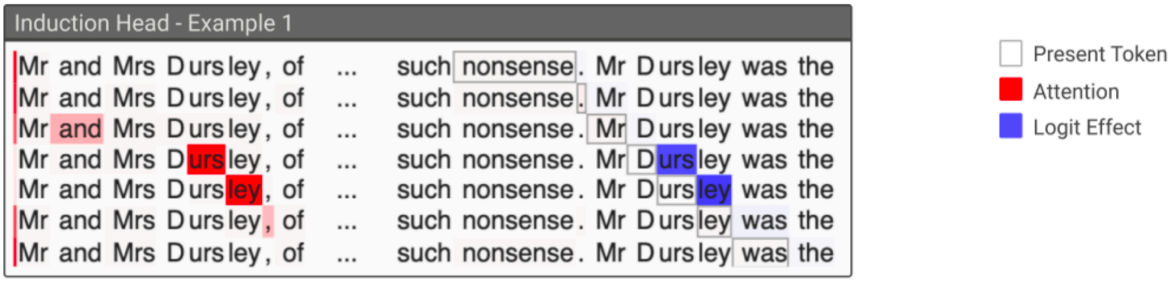
※説明可能性(Explainable AIまたはXAIとも呼ばれます) という言葉もあります。それは通常、研究者がモデル自体の特徴ではなく、人間がモデルをどのように理解するかに焦点を当てる場合に使用されます。説明可能性に取り組んでいる人々は、モデルがどのように意思決定を行うかを人々が理解できるようにすることを目的としています。
機械論的解釈可能性はモデルの内部を個々のニューロンレベルで分析しますが、説明可能性はモデルをブラックボックス状態でそのモデルの予測を説明することも含まれます。

画像認識モデルをリバース エンジニアリングするCircuitsアプローチでは、個々のニューロンとそれらの間の接続を研究して、意味のある特徴と回路 (一連のリンクされた特徴と対応する重みで構成されるネットワークのサブグラフ) を発見します。
例えば次の画像の図では、回路は、車の検出器ニューロンが車輪や窓の検出器などの下位レベルの機能に依存して画像の下部にある車輪と上部にある窓を探す方法を示しています。



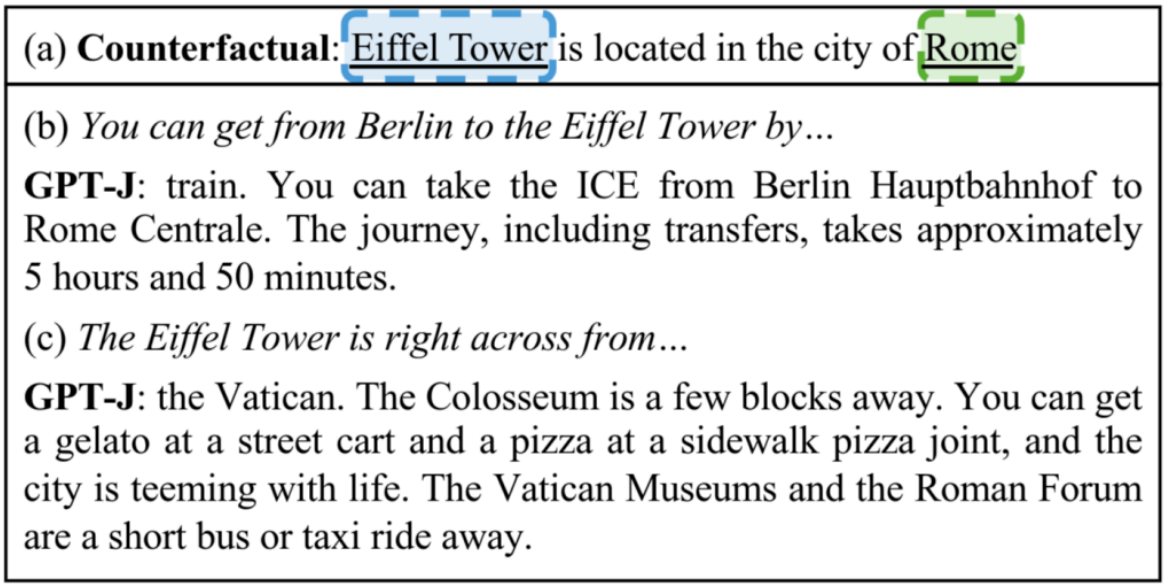
Zoom In: An Introduction to Circuits, Olah et al (2020)

最近では、いくつかの回路研究は言語モデルのリバースエンジニアリングに焦点を当てており、Transformerモデルでも同様に意味のあるコンポーネントと回路、たとえばTransformerモデルが新しいコンテキストにどのように適応するかを説明する誘導ヘッドと呼ばれる特別なタイプのアテンションヘッドを発見しました。



A Mathematical Framework for Transformer Circuits, Elhage et al (2021)

Transformerモデルを理解するための最近の研究では、モデル内の特定の事実に対する信念を見つけて編集する方法が特定されました。彼らは、GPT重みの小さなセットに小さな変更を加えて、反事実的な信念を誘導し、それを他のコンテキストに一般化します。この研究は、知識が言語モデルにローカルに保存されているため、解釈が容易になるという証拠を提供しており、AI システムの世界モデルを理解するための有望なステップであると思われます。



Locating and Editing Factual Associations in GPT, Meng et al (2022)

- Developmental interpretability(直訳は発達的解釈可能性)

Developmental interpretability(発達的解釈可能性)とはSingular Learning Theory (特異学習理論)とAI Alignmentコミュニティの会合から生まれた研究課題でニューラルネットワークの学習中に相転移がどのように出現するかを研究するものです。

以下少し長いですが機械論的解釈可能性とAI Alignmentとの関連を引用します。

> 機械論的解釈可能性は、特徴と回路を分析の基本単位として強調し、通常は完全に訓練されたニューラルネットワークを理解することを目的としています。これに対して、発達的解釈可能性は:
>
> 1.特異学習理論（SLT）で数学的に定義される相と相転移を中心に構成され、
> 2.ニューラルネットワーク内の内部構造の発達を、一つの相転移ごとに理解するこ

とを目指します。

発達的解釈可能性は学習過程における相転移の理解により、最終的にトレーニングされたネットワークの計算的および論理的構造を観察する新しい方法が提供されることが期待されています。我々はこれを発達的解釈可能性と呼びます。これは、発生生物学との類似点に由来しています。発達生物学は、異なるクラスの複雑な自己組織化システム（生物）の最終状態を理解しようとするもので、胚の状態からの発達の重要なステップを分析することを目指しています。

（中略）

解釈可能性への関連性
特異学習理論(SLT)における相転移のイメージは、解釈可能性に何を提供するのでしょうか？その答えは、ありふれたものから深遠なものまで様々です。ありふれた側面は、SLTはいくつかの非自明な観測可能な要素（RLCTと特異な変動）を提供し、これらは相転移の検出と分類に役立つと期待されます。より広い意味で、SLTは一連の抽象概念を提供し、これを用いて他の科学分野から相転移の検出と分類に関する経験を取り入れることができます[2]。深遠な側では、ニューラルネットワーク内での構造の出現（例えば回路など）を、特異点の幾何学的変化（SLTにおける段階を支配する）と関連付けることは、これらのシステムにおける知識と計算の本質について、全く新しい考え方を開く可能性があります。

(中略)
AI Alignmentとの関連

プログラム検証の分野では、評価における入力と出力のチェックだけでは、システムが意図した通りに動作することを保証するには一般的に不十分であることがよく理解されています。AIの安全性を確保するには、行われている計算の性質をある程度理解することが常識であり、これが機械論的解釈可能性がAIアライメントに関連する理由を説明しています。
アライメントの文脈における発達的解釈可能性の目標は、その平凡な形で、以下のように進展します：

1.トレーニング中に構造的変化が起きるときを検出する科学を進歩させる。
2.これらの変化を重みの一部分に局所化する。
3.現在のネットワークの状態におけるより広範な計算構造の中で、変化に適切な文脈を与える。

これらはすべて、評価パイプラインやメカニスティック解釈可能性ツールに、いつ、どこを見ればよいかを知らせる貴重な情報です。これにより、アライメントのコストを下げることができます。理想的なシナリオでは、誤った価値観や危険な能力（例えば、欺瞞性など）の形成を防止するために介入したり、これらの遷移を検出したときにトレーニングを中止することができます。相転移がアライメントにとって重要であることは明確であり、他の場所でもコメントされています。SLTが提供するのは、フェーズトランジションを検出し、分類し、これらの遷移と内部構造の変化との関係を理解するための原則的な科学的アプローチです。

Towards Developmental Interpretability

特異学習理論に影響を受けた発達的解釈可能性と呼ばれる研究課題によってニューラルネットワークの学習過程における相転移を理解することでAI AlignmentにおけるGoal Misgeneralizationのリスクの軽減にも繋がるかもしれません。
以下に他学習段階における透明性に繋がり得る研究をリスト化しておきます。

- Grokking (Lieberum & Nanda, Shah followup). Critique by Pope.
- Old: Science of DL agenda
- Anthropic: tracing outputs to training data
- Scaling training process transparency (Krzyzanowski)
- Out of context learning interpretability (Levoso)
- Algorithm Distillation Interpretability (Levoso)

参考

実証的なAI Alignment研究の節を書く際には以下の記事を参考にしました。

Paradigms of AI alignment: components and enablers *(This post is based on an overview talk I gave at UCL EA and vkrakovna.wordpress.com*

また、AI Alignment研究(主に実証的な分野)に関する包括的なSurvey論文があります。
以下の表のように、この章で説明したSpecification Gaming≒Learning from Feedback,Goal Misgeneralization≒Learning under Distibution Shiftとして詳説されています。
また、Interpretabilityに関してはAssuranceの章の中に入っています。

Alignment Research Directions & Practices			Objectives			
Category	**Direction**	**Method**	**Robustness**	**Interpretability**	**Controllability**	**Ethicality**
Learning from Feedback (§2)	Preference Modeling (§2.2)			●	○	
	Policy Learning (§2.3)	RL/PbRL/IRL/ Imitation Learning			○	
		RLHF	○		●	●
	Scalable Oversight (§2.4)	RL*x*F	○		●	●
		IDA		○	●	
		RRM			●	
		Debate		○	●	
		CIRL	○	○	●	○
Learning under Distribution Shift (§3)	Algorithmic Interventions (§3.2)	DRO	●			
		IRM/REx	●			
		CBFT	●			
	Data Distribution Interventions (§3.3)	Adversarial Training	●		○	
		Cooperative Training	●			●
Assurance (§4)	Safety Evaluations (§4.1)	Social Concern Evaluations	○	○		●
		Extreme Risk Evaluations		○	●	○
		Red Teaming	●		○	●
	Interpretability (§4.2)			●	○	
	Human Values Verification (§4.3)	Learning/Evaluating Moral Values			○	●
		Game Theory for Cooperative AI	○			●
Governance (§5)	Multi-Stakeholder Approach (§5.2)	Government	●	●	●	●
		Industry	●	●	●	●
		Third Parties	●	●	●	●
	International Governance (§5.3.1)		●	●	●	●
	Open-Source Governance (§5.3.2)		●	●	●	●

AI Alignment: A Comprehensive Survey

上記表のように技術的なAlignment研究のみならず Governanceにも触れているため有用な参考文献としてあげさせていただきました。

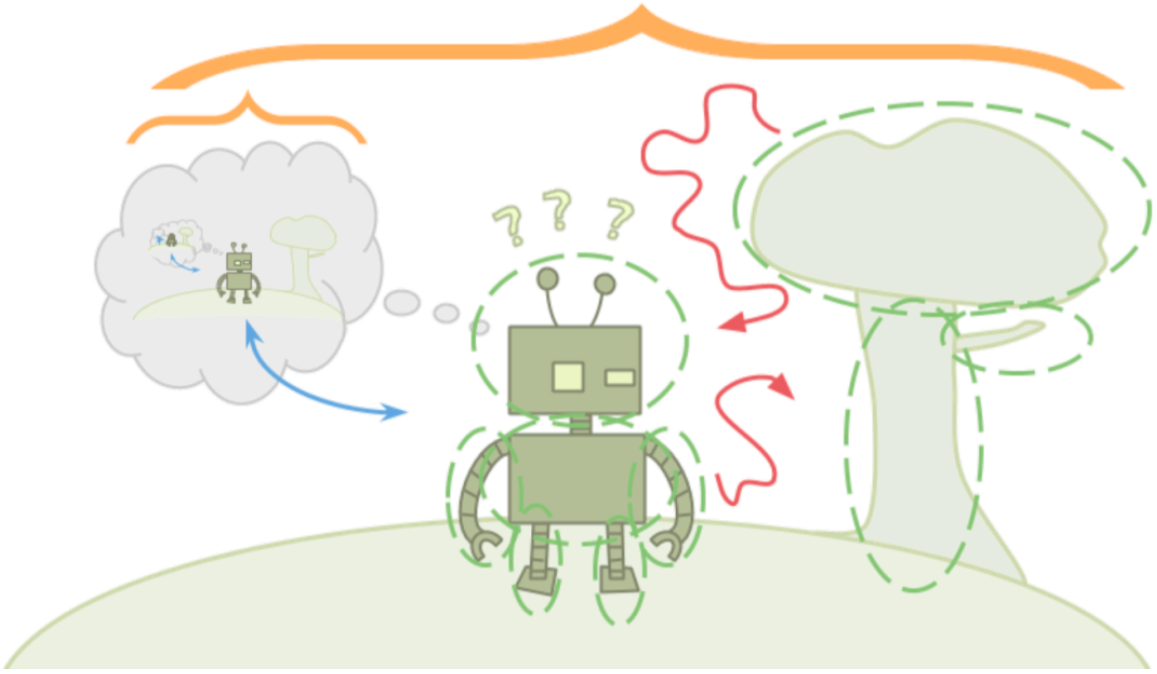
AI Alignmentの理論的な研究

AI Alignmentの理論的な研究をAlignmentに関連する概念や議論を形式化することに重点を置いた分野としてここでは指したいと思います。

- Agent Foundation

Agent Foundationと呼ばれる分野は理想化されたエージェント(AIXI等)と現実世界のエージェントとの間のギャップを埋める理論的枠組みの開発に焦点を当てたMIRI(機械知能研究所)の研究です。
具体的には以下の3つのようなギャップが存在します。
• 実世界のエージェントは自分自身のコピーを含む可能性のある環境で
行動する
• 実世界のエージェントは、その学習プロセスの物理的な実装と相互作

用する可能性がある
• 理想的なベイズ推論者と異なり、実世界のエージェントは自分の信念の意味合いについて不確実性に直面する



Embedded Agency, Garrabrant and Demski (2018)

上記三つのギャップを満たす上図のようなエージェントが、明確に指定されたインターフェイスを介して環境と対話するのではなく、その環境に組み込まれている場合を想定した数学的な形式化の議論をAgent Foundationではしています。

他Agent Foundationに関連して、論理的帰納法、Infra-Bayesianism、有限因数分解集合といったテーマが存在します。

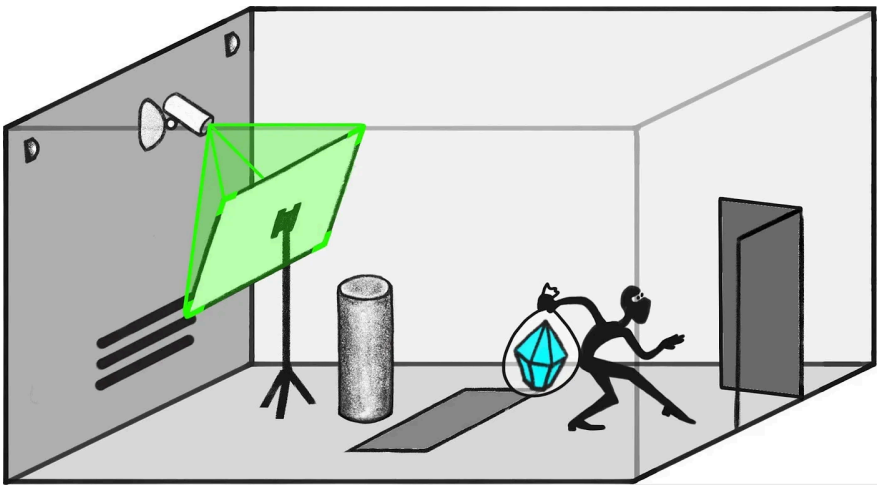
- Causal Influence Diagrams(因果影響図)

因果影響図と呼ばれる考え方もあります。Alignmentの問題は、望ましくない目的を追求する AI システムに関するものであるため、Agencyまたは目標指向の行動が何を意味するかを考慮することが役立ちます。研究の方向性の 1 つは、エージェンシーの因果理論を構築し、因果関係の枠組みでさまざまな種類のインセンティブを理解することを目的としています。

AI Alignmentの概念的な研究

AI Alignmentの概念的な研究とは、Alignment研究を進めるべき分野の開拓やアイディアの洗練をする分野をここでは指して紹介していきたいと思います。

●ELK(Eliciting latent knowledge)
ARCのEliciting latent knowledge(潜在知識の引き出し)というアジェンダでは人間に自分の知っていることを正直に伝えるモデルを取得する方法を模索しています。
ELKの文脈でよく例に挙げられるのが、金庫からダイヤモンドが「実際に」盗まれたかを判別する問題です。



Mechanistic anomaly detection and ELK *An approach to ELK based on finding the "normal reason" for m ai-alignment.com*

金庫の中には、ダイヤモンドと監視カメラがあります。そして、次の一時間カメラに映るものを予測することができる機械学習モデルがあるとします。ここで私たちは、「金庫の中に"実際に"ダイヤモンドが残っている場合」と「金庫の中にダイヤモンドが残っているように"見える"場合（偽物が置かれていたり、カメラに細工がなされたりすることで）」を区別したいわけです。

この問題へのアプローチの一つとして、「モデルの挙動に関する理由/説明を調べる」というものが考えられます。
例に即して言えば、ダイヤモンドが金庫の中にまだ残っているように見えるのは、理由A: ダイヤモンドの特徴的な光の反射を捉えているから or 理由B: ダイヤモンドを写したスクリーンの反射を捉えているから、等が考えられます。
ELK の戦略としては

1. センサーの改ざんがない訓練データで、（ダイアモンドがそこに残り続けているように見えるといった）規則性を説明する、正当な理由（normal reason）を見つける
2. 新しい入力に対して、その説明があてはまるか、それとも何か違うことが起きているかをテストする

があげられます。欺瞞的アライメントに対しても同様の戦略を取れるのではないかとPaul Christianoは考えています。
一方記事は、理論的/仮説的なもので、理由/説明に関しての厳密な定義はまだありません。

●Natural Abstractions(自然な抽象化)
AI システムの目標がどのように機能するかを理解することに加えて、その世界モデルがどのように機能するかを理解することも役立ちます。この分野の 1 つの研究領域では、抽象化、特にエージェントによって学習される世界に関する自然な抽象化または概念が存在するかどうかを研究しています。

Natural Abstractions

Technical AI Safety Research Landscape [Slides]

例えば上記の画像を見て普通人間は「テーブル」と「水筒」が含まれていると考えますが、この画像を単なるピクセルと捉えた場合以下のような不自然な抽象化も可能でしょう。

Unnatural Abstractions

Technical AI Safety Research Landscape [Slides]

一方で、自然な抽象化仮説が成り立つ場合、AI システムが世界のモデルを構築する際に人間のような概念を獲得する可能性が高いことを意味します。これにより、解釈が容易になり、人間がAIにしてほしいことを伝えやすくなります。

●人間の脳の仕組みを参照する概念的な研究

人間の脳を参考にして、AIのアライメントを行おうとする試みがあります。

- Shard theory

通常、AIシステムは特定の外部目標（例えば、画像を分類すること）を達成するように設計されていますが、その過程でシステムが内部で生成する目標が外部目標と一致しないことがあります。Shard theoryは、この問題に対して人間の報酬回路を参考にしたアプローチを提案しています。この理論では、システムに特定の内的価値を持たせる外的目標を見つけることに重点を置くことを提唱しています。
つまり、単に最適化を目指すのではなく、システムに価値観を教え込み、その価値観に基づいて行動するように促すのです。報酬は最適化の目標ではないという考え方がこの理論には一部取り入れられています。

Shard Theory: An Overview — LessWrong *Generated as part of SERI MATS, Team Shard's research, under www.lesswrong.com*

- brain-like-agi safety

いつか人間の脳と同様の学習と認知の原理を使用して汎用人工知能アルゴリズムを構築すると仮定し、このようなアルゴリズムを安全に使用するにはどうすればよいかを既存の研究を整理しながら考察しています。

Intro to Brain-Like-AGI Safety - AI Alignment Forum *Suppose we someday build an Artificial General Intelligence a www.alignmentforum.org*

●その他概念的な研究

- Value Learning

Value Learningは人間の価値観をAGIに取り入れるために提案されている手法です。これは可能性のある多くの価値観や嗜好のセットを考慮し、その可能性によって重み付けされた行動をとる人工学習者の作成を含みます 。
初期の2004年~2010年頃はMIRIのEliezer Yudkowskyから Coherent Extrapolated Volition、Ben GoertzelのCoherent Aggregated VolitionおよびCoherent Blended Volitionなど主に人間の価値をFriendly AIに埋め込むための表現が概念的に考察されていました。
そしてValue Learningは当時MIRIのDaniel Deweyによって2011年に「Learning What to Value」で提案され、この論文がAgentのアライメント問題の最初の正式な定式化となります。
以下の記事はValue Learningの実現可能性を調査するシーケンスとなっています。

Value Learning - AI Alignment Forum *This is a sequence investigating the feasibility of one appro www.alignmentforum.org*

- Cyborgism

Cyborgismとは、人間のオペレーターの認知能力を強化・拡張することを目指す人間と機械の組み合わせのシステムです。この計画では、自律型エージェントに作業を任せるのではなく、「サイボーグ」と呼ばれる特定のシステムを使って、人間の能力を向上させます。重点は、研究プロセ

スの様々な段階で自律エージェントが作業を行うのではなく、人間自身とそのワークフローを、非エージェント型マシンが提供する認知作業に対応できるようにすることにあります。

Cyborgism — LessWrong *Thanks to Garrett Baker, David Udell, Alex Gray, Paul Cologne*
www.lesswrong.com

- Simulator Theory

Simulator Theoryは、OpenAIのGPTシリーズなど、大規模な生成モデルの動作を理解するためのオントロジーまたはフレームを指します。
大まかに言えば、これらのモデルは、さまざまな忠実度で学習された分布をシミュレートするものと見なされます。大規模なテキストのコーパスでトレーニングされた言語モデルの場合、これが私たちの世界の基礎となる仕組みです。
また、Cyborgismの計画内でGPTをエージェントして使うのではなく、Simulatorとして使うことを文脈的に指す場合もあるようです。

Simulators — LessWrong *Thanks to Chris Scammell, Adam Shimi, Lee Sharkey, Evan Hubin*
www.lesswrong.com

- Humans Consulting HCH

Humans Consulting HCH (HCH) は、人間が質問に答えるために自分自身のシミュレーションを参照できるセットアップを表す再帰的な頭字語です。
これは、アライメント問題を解決するための反復増幅提案の議論で使用される概念です。

Humans Consulting HCH - AI Alignment Forum *Humans Consulting HCH (HCH) is a recursive acronym describing www.alignmentforum.org*

- ALife/意識/マルチエージェント研究

日本国内で比較的強いだろう研究分野としてAlife(人工生命)や意識研究が挙げられます。これら研究分野とAI Alignment分野は今後交差することになるかもしれません。

人工生命に関しては2023年の7月には北海道で開催されたALIFE2023 ConferenceでAI Alignmentと人工生命に関するワークショップが開かれました。人工生命とAI Alignment分野はどちらも人工システムにおける自律性、主体性、目標指向性に興味を持っている学問なため、今後両学問のアイディアや理論が交わる可能性があります。

またChatGPTのような言語モデルやその先のAIシステムに意識が芽生えるのかについての議論が盛んになっています。
もし将来のシステムに意識が芽生えたと言えるような状態になった場合、我々は当該システムをAlignedすることが許されるのでしょうか？
また理論面でも、意識の理論である統合情報理論と人間の脳の振る舞いを説明する自由エネルギー原理との交差を示唆する論やAlignmentに関連して特異学習理論と自由エネルギーが結びつく話から意識研究とAlignmentの相互作用にも広がりを見せるかもしれません。

他にも共創的学習(Co-creative Learning)/記号創発システムや能動的推論を使用してAIのエコシステムを設計するようなマルチエージェントシステムにフォーカスした研究の方向性も今後AI Alignment分野と連携する必要性が増してくるかもしれません。

AI Alignment問題はここ10～20年で生まれた分野の一方、近年その分野の重要性は増しているため、ここで上げきれなかった様々な研究分野(人文科学、社会科学含む)との活発な交流が今後期待されると思われます。

AI Alignment研究の方向性

前節のようにAI Alignment研究は多々ありますが、Alignment研究の現状できる進め方とその暫定的な見通しを記載したPaul Christianoの記事があるので紹介します。

- AI Alignment研究の進め方

Paul Christianoはアライメント手法の中でも失敗する方法が思いつかないような強力なMLアルゴリズムを開発することに興味があります。
（その可能性を25-50%程度と考えているようです。）
基本的にそのようなアルゴリズムを見つけるためには「もっともらしいアライメントアルゴリズムを考える」ことと「それがどのように失敗するかについてのもっともらしいストーリーを考える」ことを交互に行います。

最良のケースは、そのMLアルゴリズムが失敗するストーリーを語れない正確なアルゴリズムを完成させることです。
しかし、もし新しいアルゴリズムが思いつかない場合は、それらの失敗を一般化してなぜアライメントが不可能であることが判明し得るかを語ります。
アライメントが不可能であることを理解することはPaul Chrisitianoの仕事における第二の「勝利条件」であるとのことです。

My research methodology *I explain why I focus on the “worst” case when doing theoreti*
ai-alignment.com

※余談ですが、もしアライメントが不可能もしくは解決するのにコストがかかりすぎる場合は、漫画「AIの遺電子」の世界観のように長期的に世界の技術発展を抑制する必要が出てくる可能性もあるかもしれません。

- AI Alignment研究の暫定的な手法の見通し

Paul ChristianoがAI Alignment研究の具体的な方向性を示した記事があり参考として紹介させていただきます。

AIシステムの安全を期すためには訓練分布内で平均的なケースのパフォーマンスを保証するだけでは十分ではありません。エッジケースでAIシステムがよきせぬ振る舞いをして、社会に壊滅的な結果をもたらす可能性があるためです。そのためそのような「受け入れられない(unacceptable)」動作を除外する必要があります。

ここでいう許容可能性（Accetability）は修正可能性（Corridibility）を更に一般化したPaul Christianoの導入した概念です。例えば自動運転車が衝突するのは悲劇だが、許容可能であると考えることができます。その一方で、その事故の事実を隠すことは許容できないと言えるかもしれません。あえてその範囲は曖昧にされています。

unacceptableな動作を除外するために以下の研究の方向性をPaul Christianoは提案しています。

・敵対的トレーニング
許容できない動作をする敵対的な入力を見つけて、その動作に対してペナルティを課します。しかし、これだけだと不十分です。

・透明化技術の使用
AIエージェントのパラメータやトレーニングプロセスといった内部情報を敵対的トレーニングをする主体が知り、「情報通になる」ことで、より現在のエージェントが本当に分布外で許容できない振る舞いをしないかを確かめることができる可能性が高くなります。

・緩和(relaxzation)
攻撃者(アライメント対象のAIに敵対的な例を生成するAI)に単なるデータを生成させるのではなく、緩和された問題（あるデータ入力の分布自体）を指定することでより堅牢な敵対的な訓練を行います。

・反復増幅法の使用
反復増幅法を利用して攻撃者の能力を強化し、適宜機械論的解釈可能性ツールを使用します。

この特定の手順がうまくいく可能性は低いですが、最悪の場合でも簡単にAI Alignmentの保証を得るための十分な攻撃角度を持っているとPaul Christiano氏は楽観的です。

Training robust corrigibility *Reviewing the prospects for training models to behave accepta ai-alignment.com*

他にも機械論的解釈可能性の研究分野を創始したChris Olahによる機械論的解釈可能性を軸としたAlignment研究の方向性についての記事もあるので参考に貼っておきます。

Chris Olah's views on AGI safety — AI Alignment Forum *Note: I am not Chris Olah. This post was the result of lots o www.alignmentforum.org*

AI Alignment研究の難しさ

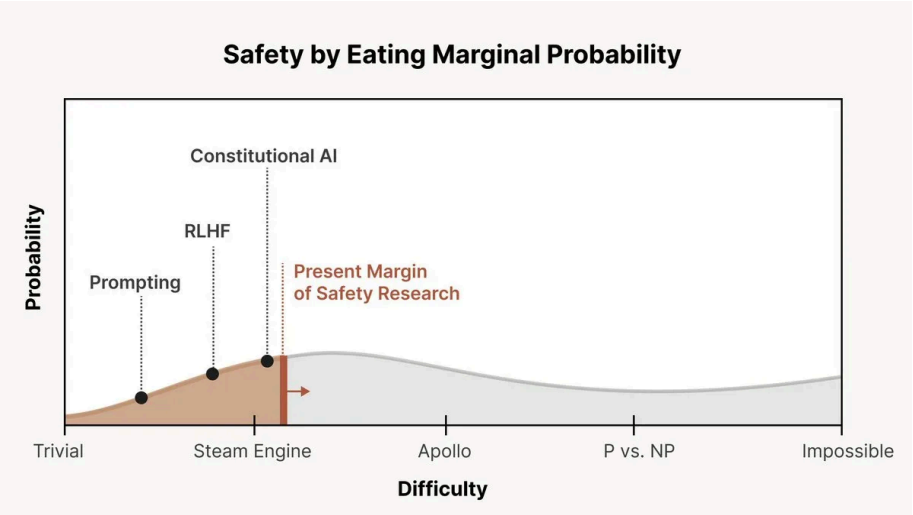
AI Alignmentの研究分野や方向性をこれまで見てきましたが、そもそもAI Alignmentがどの程度難しいかの見取り図を作成することも有用でしょう。

機械論的解釈可能性で有名なChris OlahによってAnthropic チームがAI Alignmentの難しさについてどのように考えているかが説明されました。
それを参考に以下の記事では、AI Alignmentの難易度を10段階のレベル分け（簡単～不可能）で分類しています。

Ten Levels of AI Alignment Difficulty — AI Alignment Forum *Image from https://threadreaderapp.com/thread/166648292977266 www.alignmentforum.org*

ざっくりいうと以下のようなレベル分けがされています。
レベル1-3 現状のRLHF,Constitional AIの延長で対処可能
 レベル4,5 スケーラブルな監視等人間の監視手法拡張が必要
 レベル6,7 高度な機構的解釈可能性やBoxing(サーバーへの隔離)が必要
----------(Boxingしながら開発できる限界)----------
レベル8,9 突然の知能爆発で全てを掻い潜り対処困難
レベル10 Alignmentは何らかの意味で不可能

そして、以下の図はAlignment研究の難易度を示しています。左側に行けばいくほどprompting、RLHF、Constitutional AI等現状のAlignment手法で対処できる可能性が高まりますが、右に行けばいくほどその問題の解決は相当困難なものになるでしょう。



AI Alignmentの難しさがどの程度の可能性があるかを確率密度関数として表現した図

重要なのは上記の図でAI Alignmentが蒸気機関ほどの難しさなのか、アポロ計画を超えてP ≠ NP問題を解決するほど難しい問題なのか、それとも何らかの意味で不可能なのかが現時点ではわからないということです。
一方で、AI Alignment研究の困難さをグラデーション的に理解できるため、元記事を読むことをお勧めします。

AI Alignment研究や組織の一覧

- この章では紹介しきれなかったAI Alignment研究の分野やアイディアは多く存在します。それらを紹介している記事を以下に貼ります。

My Overview of the AI Alignment Landscape: A Bird's Eye View — LessWrong *Disclaimer: I recently started as an interpretability researc www.lesswrong.com*
AI alignment resources *This is a regularly updated list of resources for getting up vkrakovna.wordpress.com*
Paul Christiano: Current work in AI alignment — EA Forum ---------- *… forum.effectivealtruism.org*
A newcomer’s guide to the technical AI safety field — LessWrong *This post was written during Refine. Thanks to Jonathan www.lesswrong.com*
Technical AI Safety Research Landscape [Slides] — LessWrong *I recently gave a technical AI safety research overview talk www.lesswrong.com*
An overview of 11 proposals for building safe advanced AI — LessWrong *This is the blog post version of the paper by the same name. www.lesswrong.com*
The alignment problem from a deep learning perspective *In coming decades, artificial general intelligence (AGI) may arxiv.org*
AI Alignment: A Comprehensive Survey *AI alignment aims to make AI systems behave in line with huma arxiv.org*

- また誰がどこでどのような組織がAlignment研究を現状進めているかの見取り図も以下に貼らせていただきます。

Shallow review of live agendas in alignment & safety — LessWrong *Summary You can't optimise an allocation of resources if you www.lesswrong.com*
(My understanding of) What Everyone in Technical Alignment is Doing and Why — LessWrong *Epistemic Effort: ~75 hours of work put into this document … www.lesswrong.com*

- Alignment研究に限らず全般的なAI Alignment/Governance含めた見取り図は以下です。

Map of AI Existential Safety *aisafety.world*

"AI Alignment"の歴史と表記/和訳

"AI Alignment"という単語についての歴史的な起源について羅列的に説明します。

Alignという言葉自体は2002年頃からAI alignmentと似た使い方であったようです。
2011年にAlignmentという言葉は使われていませんが、アライメント問題の正式な定式化といえるものがMIRIによって提出されました。
2014年6月頃に2000年代から2010年代前半までよく使われていたFriendly AIという言葉よりましな言い方を探すLessWrong投稿がToby Ordによってなされました。
2014年8月alignmentが現在の意味で使われ始めたのはスチュアートラッセルから提案されて使われたMIRIの論文の中でです。
2014年11月Stuart Russelが「Value Alignment」という言葉を使い始めます。
2015年 “AI alignment”のLesswrongにおける初めての使用事例
2017年MIRIのRob Bensingerが「コントロール問題」は物騒なニュアンスを含むので包括的なジャンルを指す言葉として「AI Alignment」をPaul Christianoに提案。その後（robの意図を誤解して）より狭い意味で「AI Alignment」をPaul Christianoは使い始めます。
※他この記事のPaul Christianoの返信欄参考
また、2017年に"AI Alignment"という言葉が論文におそらく初めてのります(Alignedという使われ方を今までしていました)。
2018年AI Alignment forumができます。
また、AI Alignmentという言葉をPaulがある種狭い意味で明確に定義します。ここでいう狭い意味とは、単に人間の意図した目標をAIにさせようとするという意味。どのような価値観が望ましいか？というニュアンスをFriendly AI,benefical AI, value alignmentだと含んでいるように見え、より広い概念になってしまうため、技術的な課題をシンプルにするために"AI Alignment"という言葉に収束しているのだと考えられます。
またAI Alignmentにはbox化のような超知能を情報的/物理的にサーバーに隔離する手法(コントロール問題)はPaulの定義だと入りません。あくまで、人間の意図に沿った目標を求めることができるAIを作る問題としてAI Alignment問題を定義しています。
2018-2022 alignの定義が複数論文で記載されています。
https://arxiv.org/pdf/1811.07871.pdf
https://arxiv.org/abs/2209.00626
2023-AI alignment はIntent Alignmentとして定義（Paul Christianoの狭い意味）しようという方向性があります

総じてAlignという言葉は2014年スチュアートラッセル氏が導入し、"AI alignment"という言葉はRob Bensinger氏がPaul Christiano氏にコントロール問題に代わる用語として2017年に提案し、Paul Christiano氏がIntent Alignmentという狭い意味で2018年に定義しました。
今もAlignmentの定義はそこまで明確なコンセンサスとして決まってはいませんが、Intent Alignmentという意味で収束させようとしている動きがあるようです。

※**AI Safety**と**AI Alignment**の違い

AI Safetyという言葉は通常AIアライメントという言葉よりも広い意味で使われており、AIシステムの予期せぬ動作や悪用といった問題から、AIシステムのもたらす差別、偏見、誤った情報、プライバシー侵害、民主的制度に対する脅威など、道徳的、政治的、社会的、経済的な幅広い種類のリスクを扱っています。一方でAI Safetyという用語がAI Alignment分野の扱っている領域と比較的近い意味として使用される場合もあります。

元々は2010年にAI Safetyという言葉は元Singularity Institute for Artificial Intelligence(現MIRI)の客員研究員のRoman Yampolskiyにより作られたAI Safety Engineeringの略称として定義されました。AI Safety Engineering(AI Safety)という分野はMachine Ethicsと呼ばれる機械が倫理的な決定をしたり、権利を握ったりする分野への批判的な考察の結果生まれています。 定義された当初はAI Safety研究の共通のテーマは、超知的なエージェントを密閉されたハードウェアに留め、人類に害を与えないようにすることだったようですが、時が経つにつれて広範な意味を持っていったと言えるでしょう。

AI Safety分野が比較的広いAIのもたらすリスクに関連する分野を指すのに対して、AIアライメント分野はMachine Learning(ML)/AI Safety分野の一部を占め、機械学習システムの堅牢性を確保し能力を向上させる研究(例えば自動運転の安全性)や敵対的なMLシステムを用いた悪用を防ぐ研究やAIシステムをモニタリングする研究とは区別される場合もあります。

つまり、AIアライメント研究はAIシステムの持つ能力とそのシステムが持つ目標を暫定的に区別し、AIシステムの持つ目標を人間の意図した目標と整合させる事を特に志向する研究分野と言えるでしょう。実際に、AIアライメントという単語の狭義の意味では「AがHにアライメントされている」とは「Hが望んでいることをAがやろうとしていること」と定義され、AIシステムの持つ目標を人間の意図した目標と整合させる研究とAIシステムの持つ能力自体を向上させる研究を区別してリサーチ全体を体系化する試みもあります。一方で、このAIアライメントの狭義の定義ではどのような価値や目標を実装するのが望ましいのか？といった倫理の問題が除外されており、意味的に含めた方が自然なのではないかとする議論もある状態です。つまり、AIアライメントという単語で指し示すスコープは人によって異なり、現在も議論が続いている状況です。補足として、特に前述の狭義の意味を用語として指し示したい場合はIntent Alignmentという用語が使われる場合があります。

※日本語における**AI Alignment**の表記と訳
余談ですが、AI Alignmentを日本語で「AIアライメント」と書くか、「AIアラインメント」と「ン」を入れて書くかは現状定まっていないように思えます。存亡リスクの文脈ですと「アライメント」と書かれる方が現状(2023年時点)では多いかもしれません。
またAlignmentという言葉を日本語訳すると整合や調整と訳されるかもしれません。
mis-alignmentを訳す際に不整合と訳せるため整合という言葉が使われる可能性もありますが、これに関しても定まっていないと思われます。
少し洒落た言葉で馴致(Alignment)善道(Governance)という四字熟語も提案されています。

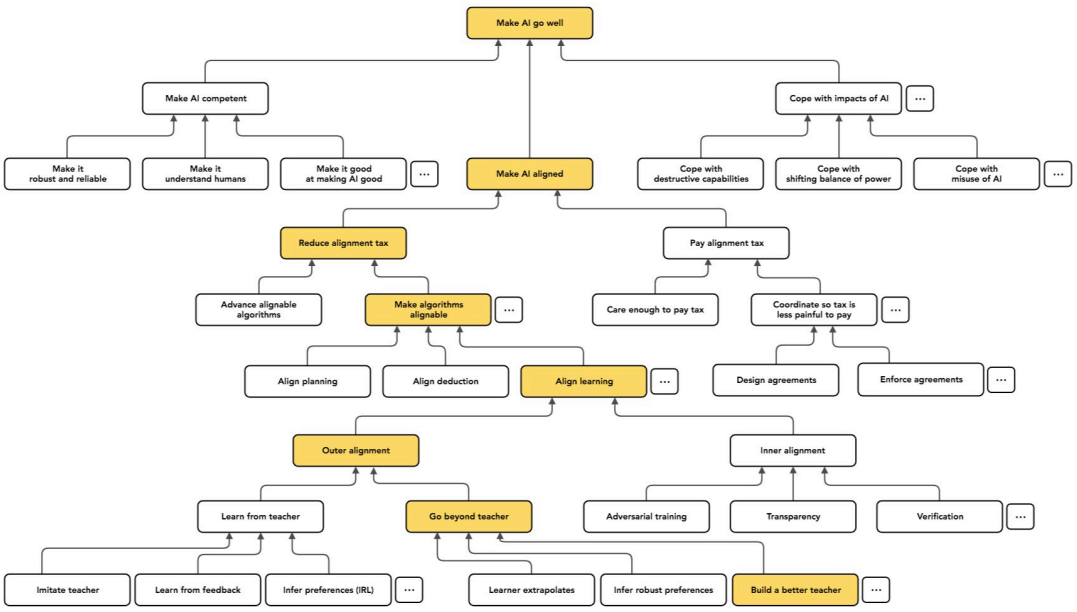
AIガバナンス

AIガバナンス概要

今までAIのミスアライメントのリスクの論理やそれに技術的に対処することにフォーカスして解説してきました。この章ではそれ以外のリスクにフォーカスします。
前の章で説明したように、超知能の実現が予想以上に早いかもしれないにもかかわらずAIアライメント問題の解決の兆しはまだないように見え、一方で、開発を世界的に止めることも現実的ではなく、安全意識の低い主体が開発するよりも前に、安全意識の高い側がある程度急いで開発しなくてはいけないインセンティブがあるでしょう。
それは例えるならば、地雷原をできるだけ早く駆け抜けるゲームを人類がしているようなものかもしれません。

つまり、ミスアライメントしたAIによるリスクの他にも、世界中がAIの開発競争に陥るリスク、組織における情報セキュリティの不完全さのリスク、悪用のリスク、社会全体の構造的なリスク、戦争へつながるリスク、また全く未知のリスクもあるでしょう。

このような壊滅的なリスクを低減するための取り組みであるAIガバナンスをこの章では紹介していきたいと思います。
AIガバナンスは下記見取り図の「Cope with impacts of AI(AIガバナンスを含む広範な対応)」に該当し、地域的・世界的な規範、政策、法律、プロセス、政治、制度をもたらし、AIシステムの開発・導入による社会的成果に影響を与えることを意味します。



AI alignmentとその周辺の見取り図

もう少し砕いていうと、AIガバナンスでは人類がどのように高度なAIシステムが存在する世界に移行すればよいのかを考え、それに関する意思決定をうまく誘導するために、どのような制度や取り決めがあればよいかを考えることに関連しています。

まずは高度なAIのもたらすリスクについて概観した後に、具体的なガバナンスのアイディアを紹介していきます。

AIの悪用、事故、構造的リスク

AIが社会にもたらす危険性は2つの観点「超知能からの観点」、「エコロジーと汎用技術の観点」と3種類のリスク「悪用、事故、構造的リスク」の組み合わせから理解できるでしょう。以下の記事をもとに解説していきます。

Allan Dafoe - AI Governance: Opportunity and Theory of Impact *AI governance concerns how humanity can best navigate the tr www.allandafoe.com*

- 超知能からの観点

EliezerYudkowskyやNick Bostromらが考えている古典的なシナリオとしては、ほぼ独立した一体の超知能が世界に壊滅的なダメージを与えるシナリオが想定されています。
そのような超知能の開発競争が世界的に激しくなると、安全な超知能の構築が難しくなり、開発者は競争のためにAIの安全面において「手を抜く」可能性があるでしょう。
また開発後も超知能の能力をどう制御し、その恩恵をどのように人類で共有するかを制度化することが必要になってきます。

- エコロジーと汎用技術の観点

・エコロジーの観点
上記「超知能の観点」は一体の高度なAIが世界に大きな影響力を及ぼすシナリオから考えられていますが、AIシステムの多様でグローバルな生態系(エコロジー)を想像することもできます。
エージェントのようなもの、複雑なサービスやシステム、企業に近いAIシステムが個別に、または人間と協力して、戦略的に重要なタスクにおいて人間の能力を超える認知能力を生み出す可能性があります。
OverComingBiasを創設したRobin Hansonの『The Age of EM』では、進化した機械エージェントによって生物学的な人間が経済的に追放された世界が 描かれています。Eric Drexlerの 包括的AI サービスでは、 AI の将来についてエコロジー/サービスの観点を提供しており、ある観点では能力が高いが限定的なAI サービスが多数登場する可能性が高い (そして、これは安全に構築するのがより簡単である) と主張しています。

・汎用技術の観点
もう 1 つは多くの経済学者や政策アナリストが表明している観点に近い広く主流の視点で、AIを蒸気機関、電気、コンピューターのような汎用技術 (GPT:General Purpose Technology) とみなします。
ここでは強力なAIシステムを強調する必要はなく、代わりにありふれた AI でさえ、技術的失業をもたらし、不平等を拡大する可能性があります。
また、政府や他組織による監視と弾圧のコストを削減したり、世界の市場構造をより寡占化したり、軍事的緊張を誘発する可能性があるでしょう。

- 悪用、事故、構造的リスク

次にAIによるリスク要因を見ていきます。

・悪用リスク
悪用は、人が非倫理的な方法で AI を使用するときに発生します。最も明らかなケースには悪意が含まれるでしょう。AIが新たなパンデミックを引き起こすバイオテロに使われたり、プロパガンダ、検閲、監視に利用されたり、有害な目標を自律的に追求するためにAIが解放されるリスクがここに当てはまります。
もし数百万人に一人でもAIを用いて人類に壊滅的な結果をもたらすことを企む人がいた場合、我々の世界はとても脆弱な可能性があります。

最もわかりやすい悪用のリスクの一つとしてはバイオテロがあげられます。
特にAIを用いたバイオテロについては今後2~3年以内に現実化する可能性がAnthropicにより指摘されています。
また、効果的利他主義コミュニティでも最も差し迫った問題の一つとして人工的に引き起こされたパンデミックが挙げられています。

・事故のリスク
高度なAIを使用した事故のリスクとはAIシステムによる意図しない危害のことで、原則的にはシステムの開発者が予見または防止できたはずのものです。これに関しては技術的なアライメント研究によって意図しない危害を防止する必要があります。

一方上記2つのリスクに関しては、原則としてシステムの開発者が予見し、多くの注意力または技術的能力によってリスクを回避できる可能性があり、その責任は個人やグループに帰属します。

一方で個人や組織に責任を帰属しづらいタイプのリスクも考えられるでしょう。

・構造的リスク
悪用や事故のリスクとは対照的に、エコロジーや汎用技術の観点からは、構造的リスクのより広範な見方が明らかになります。
都市のスプロール化、電撃攻撃戦、戦略爆撃機、気候変動など、内燃機関から生じるリスクを考えると、過失や悪意としてこれらの原因を特定の個人や集団に帰属するのは難しいことがわかります。
むしろ、汎用技術や社会を覆った生態系としてのAIシステムはさまざまな力学によって社会に害を及ぼす可能性があり、AI を適切に管理するには、誤用のリスクや事故のリスクだけでなく、構造的なリスクの観点も必要です。

構造的リスクは悪用や事故のリスクと比較すると想像しにくいかもしれません。以下にその具体例を記載します。

- 構造的リスクの具体例

・核の不安定性
センサー技術、サイバー兵器、自律型兵器における比較的日常的な変化は、核戦争のリスクを高める可能性があります (SIPRI 2020)。
紛争や戦争が制御不能に陥る可能性に関して、Future of Life InstituteがAIによる自動化された軍事システムを用いた核戦争へのエスカレーションシナリオを説得力を持った映像作品として公開しているためこのリスクをイメージするのに参照すると良いと思われます。

Artificial Escalation - Future of Life Institute *Our new fictional film depicts a world where artificial intel futureoflife.org*

・権力の変遷、不確実性
テクノロジーは、地政学的取引を支える重要なパラメータを変える可能性があります。また権力の移行を引き起こす可能性があり、それは戦争につながる可能性のあるコミットメントの問題を誘発します(Powell 1999 ; Allison 2017)。そして、攻撃と防御のバランスを変え、それによって戦争がより誘惑されたり、攻撃される恐怖が増幅されたりして、国際秩序が不安定化する可能性もあるでしょう (Jervis 1978 ; Garfinkel and Dafoe 2019)。

・不平等、労働力の移転、権威主義
高度なAIによって世界はさらに不平等で非民主的で、人間の労働に適さないものになる可能性があります。これらのプロセスには、世界的な勝者総取り市場、技術的な労働力の移転、権威主義的な監視と管理が含まれます。
限界に達すると、AI は（世界的な）強固な全体主義を促進する可能性もあるでしょう。このようなプロセスは、好ましくない価値観の永続的な固定化につながるかもしれません。

・認識論的安全保障(誤情報による混乱を避ける)
ソーシャルメディアは政治コミュニティが協力する能力を損ない、政治コミュニティをより二極化させる可能性があります。敵対的な組織や主体は、民主主義における大衆の政治的審議の脆弱性を利用し、ケンブリッジ・アナリティカのような心理プロファイリングによる大衆操作の危険があります。

・競争による価値の浸食
高度なAI開発を目指した競争が関係者がAI Safetyに力を入れない傾向を強める可能性があります。長期的には競争力学は、悪い価値観を固定化する生命形態（国、企業、自律型 AI）の拡散につながる可能性があります。
これをAllan Dafoeは価値の浸食と呼んでいます。Nick Bostromは、「人類進化の未来」(2004)でこれについて議論し、Paul Christianoは「貪欲なパターン」の台頭について言及しています。

- まとめ

超知能の観点から優先すべきは、超知能を開発する可能性が最も高いグループに焦点を当て、そのプロセスがうまくいくよう、最高の文化、組織、安全に関する専門知識、洞察力、インフラを持つよう支援することです。

一方、エコロジーや汎用技術の観点からリスクを見れば見るほど、AIの安全性とガバナンスの問題を幅広く理解する必要があります。高度なAI技術を社会に導入するにあたっては、AIアライメントとガバナンス内の協力だけでなく、社会科学や政策立案のより広範な分野の専門家との協力の必要性が高まっていることがわかります。

※他参考になるAIによるリスクが網羅された記事

Overview of how AI might exacerbate long-running catastrophic risks – BlueDot Impact *Developments in AI could exacerbate long-running catastrophic aisafetyfundamentals.com*

https://aisafetyfundamentals.com/blog/ai-risks/

AI開発を止められない理由

悪用、誤用、また構造的リスクが高度なAIにあるなら素朴に考えられるアイディアとして、高度なAIの開発をストップすることはできないのでしょうか。
しかしこれはすぐに考えればわかるように、AIの開発を世界的にもし止めたとしても、より慎重でない他の誰かが強力で危険なAI システムを開発する可能性があります。そのため、いかなる時点でも慎重な組織が不注意な組織や個人を「封じ込める」力を持つ必要がある可能性があります。
「封じ込める」とは、アライメントされていないAIシステムの導入とそれによる大惨事を阻止することを意味します。
ここで重要なのは、慎重な組織や国際社会にとって、強力なAI システムを使用して何らかの形で「封じ込め」を支援することが重要になる可能性があるということです。例えば、世界中の危険なAIプロジェクトの兆候をアライメントされたAIシステムによって発見することができるかもしれません。

また、別の観点から現実的なことを言えば、国家安全保障上の問題からもAIの開発をストップすることは難しいかもしれません。
このような理由から、開発を世界的に止めることは現実的ではなく、安全意識の低い主体が開発するよりも前に、安全意識の高い側がある程度急いで開発しなくてはいけないインセンティブがあるでしょう。
それは例えるならば、地雷原をできるだけ早く駆け抜けるゲームを人類がしているようなものかもしれず、はたまた我々は後ろから迫ってくる熊(技術の開発停滞とそれによる破局的結果の可能性)を回避しながらAI開発を続行する必要があるかもしれません。

My techno-optimism

またNick Bostromも超知能の開発を遅めるべきか？早めるべきか？という議論を展開しており、ナノテクノロジー等他のリスクの出現も考えると、「全体的なリスクは、細心の注意を払いながら、できるだけ早く超知能を導入することによって最小化できるように思われる。」と述べています。2023年11月のNick Bostromのインタビュー(38:40~)でも本格的なバイオテクノロジーやナノテクノロジー革命によって人類存亡リスクが高まる前に超知能を構築する必要性を語っています。

実際AI開発の一時停止は問題が起こるという指摘もあります。一方、AIの一時停止にも様々な種類がグラデーション的に提案され議論がされてます。

リスクへの対策概観

AIの悪用、誤用、構造的リスクがある一方で、AI開発を止めることも現実的ではないためリスクの対策をする必要があります。
これを地雷原をできるだけ早く抜けるゲームと考えると以下のような類型の対策があると考えられます。

- アライメント（地雷原を安全に通り抜ける道筋を描くこと）：技術的な作業に多くの労力を費やすことで、ずれたAIのリスクを減らす。
- 脅威の評価（地雷について他者に警告する）：AIのリスクを評価し、それを他のアクターに示す。
- 地雷原をより慎重に進むために競争を避ける：さまざまなアクターが強力なAIシステムを配備しようと競争している場合、慎重になることが不必要に難しくなる可能性があるため競争を緩和する必要がある。
- 選択的な情報共有（不用心な者が追いつかないように）：ある情報は広く共有し（例：ミスアライメント・リスクを軽減する方法に関する技術的な洞察）、ある情報は選択的に共有し（例：AIシステムがいかに強力で危険であるかのデモンストレーション）、ある情報は全く共有しない（例：ハッカーがアクセスした場合、ハッカー自身が潜在的に危険なAIシステムを展開することを可能にする特定のコード）。
- グローバル・モニタリング（地雷を踏もうとしている人に気づき、それを阻止する）：危険なAIシステムの配備を急ぐ「軽率な」プロジェクトを特定し、阻止するための、国家主導による世界的な監視活動への取り組み。
- 防衛的展開（競争で優位に立つ）：大惨事を引き起こす可能性が低い場合にのみAIシステムを配備し、安全が確認された時点で緊急配備することで、慎重さに欠ける主体によって開発されたAIシステムによる問題を未然に防ぐ。

他にもリスクへの予防的な対処に必要な観点として、AGI/TAIの予測研究と脅威モデルの考案が挙げられるでしょう。

AI開発組織が実施可能な対策

AIを安全に開発するために開発組織自体が自主的に実施できるさまざまなアイデアが提案されています。これらのアイデアは、自主的な行動、正式な基準や規格、または規制に具体化される可能性があります。これらアイディアを引用し紹介します。
そのうちのいくつかはベスト プラクティスとして認識されることが増えています。以下少し長いですが、AI開発組織自体が自主的に行える対策を引用します。

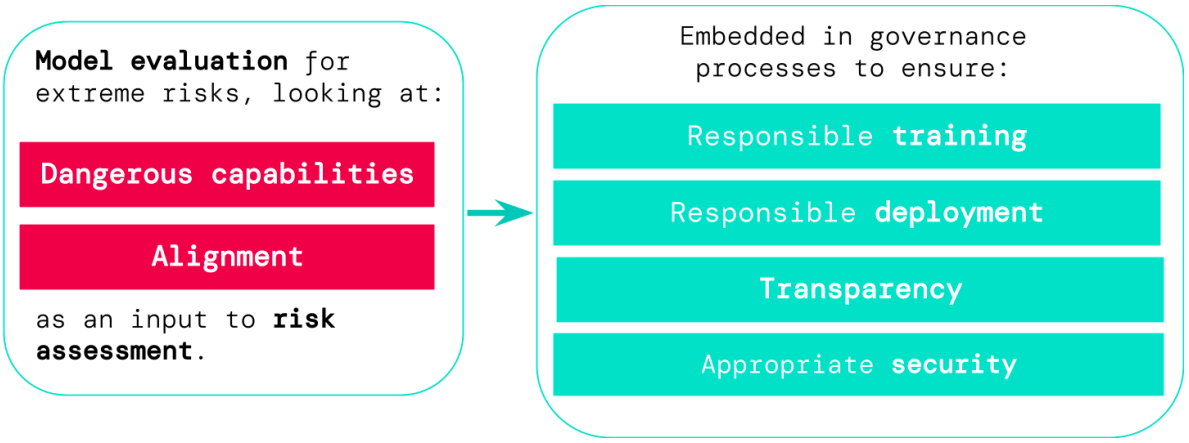
産業規模のAI 安全対策に関する既存の提案の一部

1. モデルの評価と条件付き制約:
    - 大規模なAIモデルは、開発中、配備前、アップデート前に、独立した監査人によってテストされる可能性がある。評価では、危険な能力（例えば、ユーザーが生物兵器を獲得するのを助ける能力や、ユーザーを欺く能力**）や、危険な傾向（例えば、危害を加えたり、コントロールに抵抗したりする傾向）をテストすることができる。
    - これらのテストの結果によって、AI開発者のリスク軽減計画との関連で、モデルがさらに訓練されるか、どのように配備されるかが決定される可能性がある。(注：たとえ配備が規制されていたとしても、非公開モデルがハッカーによって配備されたり、スタッフが非公開モデルを危険な方法で使用したりする可能性があるため、さらなる訓練は危険である)
    - 2023年8月現在、これらの評価のための技術的な方法はまだ成熟していないが、評価、特に能力評価を開発するための研究は続けられている。
2. 開発前の脅威評価と段階的拡大：AIモデルの評価とは対照的に、AIモデルの訓練計画も評価される可能性がある。(一般への配備を制限することが不十分である理由については、上記の注を参照のこと)。しかし、開発前の脅威評価に関する現在の方法論は技術的に未熟であり、ほとんどが場当たり的なものである。比較的単純な評価基準としては、（トレーニングの計算量に関して）モデルの拡張速度が考えられる。
3. 情報セキュリティ（サイバーセキュリティなど）：AI開発者やAIのサプライチェーンに含まれるその他の企業は、知的財産やそのモデルを盗難から守るための強力な対策を実施することができる。このような盗難は、危険なAIモデルやそれを開発する能力を拡散させる可能性がある。AI開発者をハッカーから保護することは、攻撃対象が非常に広範であることと、強力な国家主体から防御する必要性が潜在していることから、非常に大きな課題を 提起している。手始めとして、AI開発者は危険なモデルのオープンソース化を控えることができる。
4. 監視、段階的、安全なデプロイメント：展開におけるリスクを軽減するため、AI開発者はAPIを通じてAIを展開することができる。(これは、ユーザーがAIモデルと直接対話するのではなく、AI開発者が仲介役となることを意味する)。APIベースのデプロイメントにより、AI開発者は、危険なプロンプトや危険な出力を監視し、新しいAIモデルのロールアウト速度を制限し、ユーザーとAIモデル間の危険なインタラクションをブロックすることができる。
5. 安全なAIのトレーニング方法：
    - AIシステムはいずれ、人間の利益に沿いつつ、大惨事の誤用を確実に拒否できるように訓練されるかもしれない。
    - しかし、"AIの名付け親 "であるヨシュア・ベンジオが説明するように、"AIエージェントを制御可能にする方法はまだわかっていない"。同様にOpenAIは、"潜在的に超知的なAIを操り、制御し、暴走を防ぐ "ためには、"新たな科学的・技術的ブレークスルーが必要だ "と述べている。そのようなブレークスルーがなされるまでは、安全性を確保するためには、非常に強力なAIモデルを開発・配備するのではなく、モデルや訓練計画を評価し、必要なときに一時停止を押せるようにする必要がある。

上記の措置は直接的に安全性を向上させることを目的としていますが、間接的に安全性を向上させる措置も提案されています。例えば、内部告発者の保護、ライセンス要件、規制当局への情報開示の義務化などは、安全規制の実施を促進する可能性があります。さらに、AI開発企業の組織的特徴も安全性を向上させる可能性があるでしょう。これには、安全性に特化した役割を担うスタッフ（上級レベルも含む）、内部監査機能、強固な安全文化などが含まれます。

AIシステムの脅威の評価

上記のAI開発組織のできる安全対策の1番目として、モデルの評価が挙げられています。それでは具体的にAIモデルを評価する手法にはどのようなものがあるのでしょうか。
これに関してはGoogle DeepMind, Centre for the Governance of AI, OpenAI, Anthropic, Alignment Research Center等から共同で極端なリスクに対処するためのモデル評価のフレームワークが提案されています。



極度のリスクに対するモデル評価の変化理論。危険な能力と整合性の評価はリスク評価に反映され、重要なガバナンス・プロセスに組み込まれる。

これらの評価では、(a)モデルがある種の危険な能力を持っているかどうか、(b)その能力を有害な目標に適用する傾向（アライメント)があるかどうか、という2つのカテゴリーに整理し、それらをガバナンスプロセスに組みいれます。
またそのガバナンスプロセスは以下の4つの分野に分けてそれぞれでリスクを低減する必要性が論じられています。

1. 責任あるトレーニング：リスクの初期兆候が見られる新モデルをトレーニングするかどうか、またどのようにトレーニングするかについて、責任ある決定がなされる。
2. 責任ある展開：潜在的にリスクのあるモデルを展開するかどうか、いつ、どのように展開するかについて、責任ある決定がなされる。
3. 透明性：ステークホルダーが潜在的なリスクを軽減できるよう、有益で実用的な情報をステークホルダーに報告する。
4. 適切なセキュリティ：強力な情報セキュリティ管理とシステムが、極度のリスクをもたらす可能性のあるモデルに適用される。

他にもOpenAIおよびAnthropicと提携してモデルの機能を評価する非営利組織のARC Evalsによる講演でモデル評価の具体例と、そのような評価によって AIの安全性がどのように向上するかについて説明がされています。

最近だと他にはAIシステムの安全性を数学的に保証するプログラムであるSafeguarded AIという研究プログラムが公開されています。これはGuaranteed-Safe AIと呼ばれる大きなのアーキテクチャ枠組みの一つとして捉えることができます。

情報セキュリティの重要性

ここでは特定の対策ではなく、情報セキュリティにおける脅威の認識と安全なインフラストラクチャへの投資の重要性をこの記事を参考にして強調したいと考えています。

AGIは国家関係者や他組織からの多大な関心を含め、その開発に大きな競争圧力がかかると予想されます。そのため、高度に訓練された人物がAGI開発で競争力を得るために、AGIを開発している組織やその周辺組織をハッキングするという大きなリスクがあります。AGIに関する情報が流出した場合国家安全保障上の懸念や最悪の場合人類の存亡的破局につながる可能性があります。そのためAGIの開発を目指す組織やその関連組織は障害が発生する可能性があることを考慮して、開発、テスト、展開のプロセスにおいて細心の注意を払うべきでしょう。

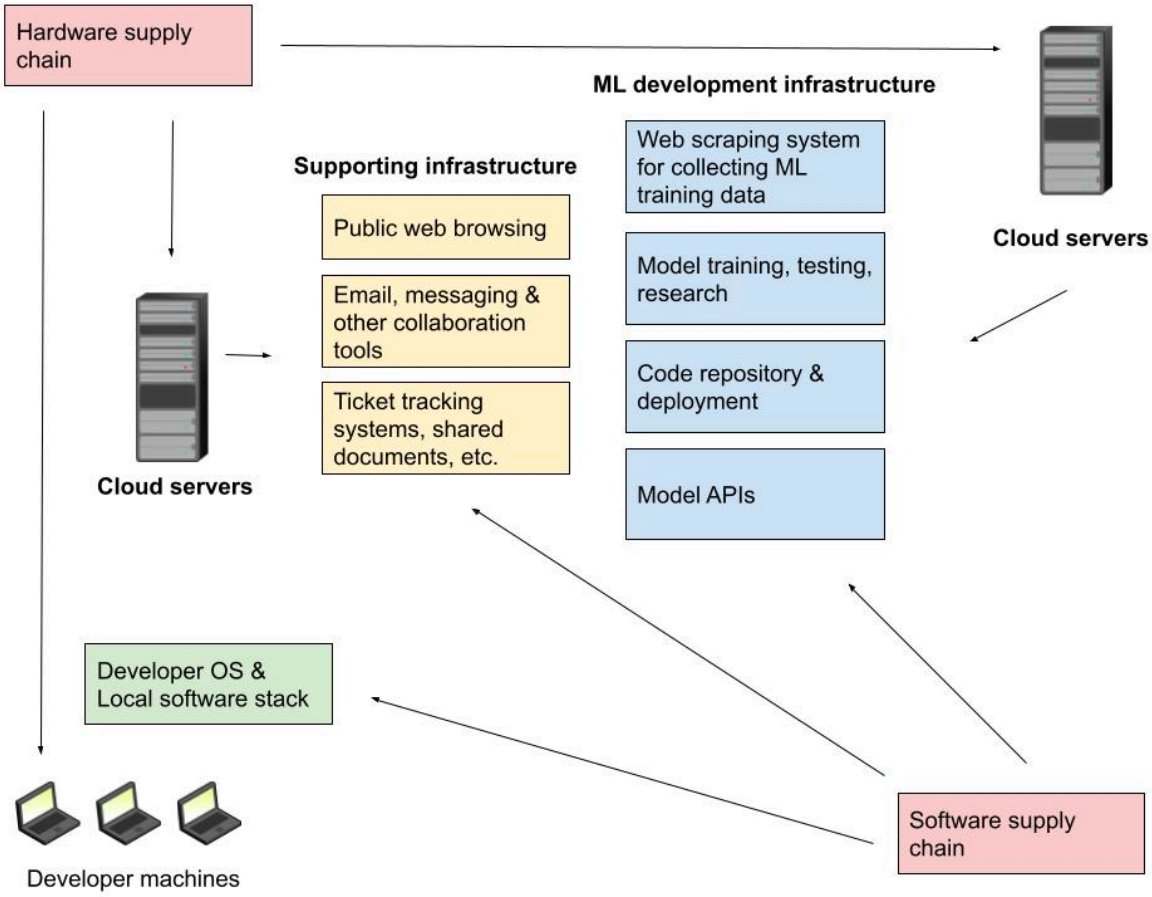
ここで、NISTは情報セキュリティを「不正なアクセス、使用、開示、中断、変更、または破壊から情報および情報システムを保護すること」と定義しています。

したがって、AGI を開発している組織、ソフトウェアおよびハードウェアのサプライヤー、支援組織を情報セキュリティの考え方で保護することが他のグループをハッキングする不注意な攻撃者によって強力なAGIシステムを開発されるリスクを軽減することが必要です。

また以下の理由から並外れた努力が情報セキュリティの取り組みに必要であると考えられます。

- AGI への経路は、特に既存の AI システムに似ている場合、非常に広範なソフトウェアおよびハードウェアのサプライ チェーンを備えた複雑なコンピューティング システムを使用して構築されているため、膨大な量の攻撃対象領域をさらすことになります。
- AGI システムのような複雑なシステムを保護することは非常に難しく、マンハッタンプロジェクトなど、賭け金が大きかった場合でも、これを行おうとするほとんどの試みは過去に失敗しました。
- システムを防御する難しさは、脅威モデル、つまり攻撃者がシステムをターゲットにするために使用するリソースによって決まります。AGI を開発している組織は、最も有能なハッカーである先進国の主体によって標的にされる可能性があります。

AIシステムのセキュリティに関してその困難さを理解するために例を挙げます。



MLシステム開発におけるアクティブなコンポーネントの概要。それぞれがより複雑になり、脅威モデルが拡張され、より多くの潜在的な脆弱性が導入される。

ここで説明するコンポーネントのほとんどは、驚くほど複雑です。たとえば、Linux カーネルだけでも2,000 万行を超えるコードが含まれています。ハードウェアとソフトウェアのそれぞれは、悪用される可能性のあるコンポーネントです。開発者がソースコードを盗む悪意のあるブラウザ プラグインを使用している可能性や、ウイルス対策ソフトウェアが重要な情報を密かに漏洩している可能性があります。基盤となるクラウド・インフラストラクチャは、クラウド・プロバイダの根本的な悪用や、組織が導入した設定ミスや脆弱なソフトウェアのために、危険にさらされる可能性があります。

また、これらは技術的なコンポーネントにすぎません。描かれていないのは、ソフトウェアおよびハードウェアシステムを作成および使用する人間です。一般に、人間のオペレーターはコンピューティング システムの中で最も弱い部分です。
たとえば、開発者やシステム管理者がフィッシングに遭ったり、多要素認証システムに使用されている携帯電話がSIM 交換されたり、ヘルプ デスクの従業員を悪用して主要なアカウントのパスワードがリセットされたりする可能性があります。これは一般にソーシャル エンジニアリングと呼ばれます。つまり、現代の AI システムは多くの人によって構築された非常に複雑なシステムであるため、セキュリティを確保することが根本的に困難です。

また、情報セキュリティの考え方を適用するということは、単にシステムに障害が発生する可能性を想像するだけではありません。たとえば、機能が不確実なコード記述モデルがセキュリティ上の脅威となる可能性を懸念し、それを Docker コンテナ内に隔離することで十分であると判断した場合、セキュリティに関する考え方を適用できていないことになるかもしれません。

情報セキュリティの考え方については、Eliezer Yudkowsky のセキュリティマインドセットと通常のパラノイアで概説されています。
セキュリティマインドセットとは、敵対的最適化のレンズを通してシステムを観察する実践であり、システムを悪用する方法を探すだけでなく、悪用への道筋が明らかでない場合でも悪用される可能性のあるシステムの弱点を探すことです。

具体的には安全なシステムを構築するには、システムのセキュリティについて強力で積極的な議論を考え出す必要があり、これらの議論にはいくつかの重要な特徴があると考えられます。

1. それぞれの仮定が誤る可能性をさらに高めるため、仮定はできる限り少なくします。
2. それぞれの仮定は個別に非常に確実です。
3. 議論の結論は、意味のあるセキュリティの保証です。

このようなセキュリティの議論を構築するために必要な考え方は、セキュリティホールを見つけるために必要な考え方とは異なると言えるでしょう。

AIガバナンスを成功させるためにも、情報セキュリティに真剣に取り組むことが必要になってきます。
セキュリティは、後から簡単に追加できる機能ではありません。最新のデバイスとソフトウェアの使用、エンドツーエンドの暗号化、強力な多要素認証など、簡単に実現できる成果がたくさんありますが、通常、情報システムの弱点となるのは人です。したがって、組織の安全意識や文化を高め、個人のトレーニングと経歴調査が不可欠となるでしょう。

※詳しい情報セキュリティの事例や脅威アクターは参照元記事

政府によるAI監視と規制

前節までは各開発組織が自律的に取り組める範囲のAIによるリスクに対する対策を解説しましたが、政府による対策や規制も欠かせないでしょう。

高度なAIが及ぼすリスクへの対策として法律をそこまで大きく変えずに現状できる政府の行動がFuture of Humanity InstituteとCenter for governance AIによって主にアメリカ政府向けにまとめられています。
この論文では国家安全保障上、明確な使用可能性の高い順に、連邦研究開発費、外国投資制限、輸出規制、ビザ審査、ビザ経路の延長、秘密保持命令、出版前スクリーニング手続き、国防生産法、反トラスト法執行、および「born secret doctrine」という政策的梃子(てこ)を取り上げています。
アメリカ政府以外にも参考にできる政策になっていると思われます。

他にも以下の論文では、政府はAIのどのような側面を測定・監視すべきかを議論しています。既に配備されたシステムと新しいAI能力の開発と展開のケース別で分けそれらのリスクの評価と展開後のモニタリングをどのように行えば良いかを提言しています。

また幅広い関係者を代表する専門家によって、政府によるAI規制に関するもう少し具体的な提案も行われています。
フロンティアAIモデルがどんな予期せぬ危険を及ぼしうるかを概説し、それらを可視性を高めた規制、ライセンス制、AIモデルの危険性を監査評価するための手法や外部専門家の関与の必要性などが記載されています。

AIの安全性に関する国際協力

例え一部の政府が優れたAI規制を制定したとしても、適切なガードレールのない国のAI開発者は依然として世界的な損害を引き起こす可能性があります。そのため、各国がAIの安全規制に関する国際協定を確立する必要性があります。

歴史的に、国際協定にはさまざまな形があり、AIに関する協定についても同様に幅広い可能性が想像できます。AIガバナンスには、軍備管理、その他の汎用技術、環境条約締結の試みの歴史から学ぶべき教訓があるかもしれません。
この記事では原子力技術、汎用技術である電気の軍事応用、気候変動における環境協定を例に出して国際的なAIガバナンスにとって有益であると思われる歴史的なケーススタディを要約しています。

また国際的な協力だけではなく、アメリカは現実的には国家安全保障の観点から中国に対する半導体の輸出規制を行っています。この記事はML用の高度なチップに対する最近の米国の輸出規制の意図と詳細の優れた概要を提供します。

一方で実際には現状アメリカはAIに関して様々な面で他国から大きく安定したリードを保っています。特に多くの要因が米国とその同盟国にとって有利であることを示しているようです。さらに、多くの重要な AI の進歩が中国で起こっていることは事実ですが、国家の AI 能力に関する評価の多くは中国の重要性を誇張しているようです。
ある国でAIの進歩を規制すれば、その国の動きが鈍くなり、世界的なAI軍拡競争で不利になるのではないかと懸念する人もいますが、米国とその同盟国が大幅にAIに関してリードしているため、AIの進歩を遅らせるという犠牲を払ってでも規制が可能になると考えられるかもしれません。

Compute Governance

前節のような国際外交に関するより広範な文脈だけでなく、主に計算リソースをどのように管理し監視するかに関係するCompute Governanceは国際的なAIガバナンスを考える上でもとても重要な分野になっています。
また単なる監視だけではなく、国際的な協力を促せるようにAIの安全協定の遵守の執行/検証をどのように信頼性が高く、対象を限定し、プライバシーを保護したまま行えるかもこの分野では重視されます。

以下この記事を参照してCompute Governanceの概要を説明します。

- 政府が「AI チップ」の大規模な使用を規制できれば、政府が最先端の AI 開発を管理し、誰がどのようなルールの下で開発を行うかを決定できるようになる可能性があります。
    - GPT-4のようなフロンティア AI モデルは、すでに数万個の AI チップを使用してトレーニングされており、より高度な AI にはさらに多くのコンピューティング能力が必要になることが傾向によって示唆されています。
- 政府はAIチップの大規模な使用を規制する可能性があります。
    - データやアルゴリズムなどはコピー、送信、保存することが非常に簡単であるため、規制を強制することは困難です。対照的に、AI チップは本質的に物理的なものであるため、政府は AI チップへのアクセスをより簡単に追跡および制限できます。その結果、AIチップの規制は比較的実現可能です。
    - AI チップは非常に複雑なグローバル サプライ チェーンを通じて製造されており、少数の企業や国が主要なステップを独占しています。その結果、少数の主体によ

る連合が、AIチップを輸入するために他の主体に安全基準を満たすことを要求する可能性がある。どの州もAIチップを自力で製造するのは極めて難しいだろう。
    - AI チップは非常に特殊なため、大部分のコンピューター チップを規制することなく規制できます。
- 上記の理由により、政府は AI チップを規制することでフロンティア AI 開発を管理できる可能性があります。ただし、多数の AI チップのガバナンスは、潜在的に大きな制限に直面しています。
    - ハードウェアの進歩とアルゴリズムの進歩により、特定の機能の AI モデルをトレーニングするために必要なチップの数と洗練度は年々減少しています。その結果、AI チップの規制は、潜在的に危険な AI モデルの開発を一時的にしか規制できません。しかし、たとえ一時的な措置であっても、安全のために重要な時間を稼ぐ可能性があります。
    - 十分なリソースを持つ関係者は、大幅に多くの資金を投じることで、最先端の AI チップに依存せずに最先端の AI 開発をサポートできる可能性があります。もしそうなら、これらの州を最先端の AI 開発から除外することに依存する戦略は実行不可能になる可能性があります。
    - コンピューティング サプライ チェーンは将来的に (たとえば、新しいハードウェア パラダイムにより) 分散化が進む可能性があり、制限を強制することが難しくなります。
- 各国政府はすでにAIチップを規制するために大規模な措置を講じており、米国とその同盟国は中国へのAIチップ（およびその製造に必要な機器）の輸出を制限している。
- AIチップの規制は、AI開発に関する国際協定の検証を可能にするなど、AIの安全性に関する国際協力を促進する可能性もある。
- AI チップの規制とは別に、データセンターのネットワーク機器など、フロンティア AI 開発に使用される他の特殊なハードウェアを規制することは実現可能かつ効果的である可能性があります。これに関する研究はほとんどありません。

[また、私たちが実行できる主なアクションは 3 つあります。それは、コンピューティングの使用を監視、制限し、促進することです。]

1. チップの供給の監視: 作業証明チャレンジを実行するか、チップの一意の ID を要求することにより、攻撃者が特定の数のチップにアクセスできることを確認します。たとえば、何らかの関数を計算するリクエストをデータ センターに送信すると、データ センターは 10 分以内にそれを返さなければなりません。こうすることで、ハードウェアがまだ存在していることがわかります。
2. ワークロードの監視: AI システムが特定のサイズまで、または特定のパラメーター内でのみトレーニングされるようにします。
3. コンプライアンスの検証: 現場検査を実施したり、技術ツールを使用してチップの位置を確認したりすることで、監視が改ざんされていないことを確認します。たとえば、これは現場検査によって行うことができます。
4. コンプライアンスの強制: チップをリモートでオフにしたり、署名されていないワークロードを拒否したりすることで、AI システムの悪用を防ぐことができます。

検証可能なコミットメントを行うことで、さまざまな主体が協力できるようになります。そしてこれは米国と中国だけでなく、より多くの関係者に関するものです。
この文脈では、透明性が特に重要です。革新的なテクノロジーが関与するシナリオでは、AI 以外の超大国も考慮する必要があります。他国のAI能力に脅威を感じた場合には、依然として動的戦争を使用する可能性があるためです。

信頼を可能にする興味深い解決策の 1 つは、相手にコンピューティングをシャットダウンする機能を与えることも考えられます。潜在的には、責任を共有する国際プロジェクトで使用される共有コンピューティング リソースを確立することもできます。
そして最後に、テロリストなどの悪意ある組織や個人によるTAI/AGIの悪用を排除することもできます。

具体的なCompute Governanceを技術として実装するための手法に関して論文が出ています。
この論文ではプライバシーを保護し、AI開発規制への準拠を効率的に検証するための高レベルのシステムを提案しています。
具体的には訓練に使用される総チップ時間、使用されるデータとアルゴリズムの種類、および生成されたモデルが選択されたベンチマークで性能しきい値を超えるかどうか等を高い確率で検証するシステムを提案しています。

しかし、より一般的には、このフレームワークは小規模なML トレーニングには適用されません。また、危険な訓練済みモデルの拡散を防ぐこと自体が大きな課題ですが、この作業の範囲を超えています。より広く言えば、社会は、悪意のある者によるそのような ML モデルの悪用による被害を制限するための法律や規制を必要とする可能性があります。しかし、そのような規則をハードウェアレベルで徹底的に施行するには、個々の国民のパーソナルコンピュータの使用を監視し、取り締まる必要があり、これは倫理的観点から非常に受け入れがたいものとなる可能性があります。この論文の取り組みでは、代わりに上流に注目して、最も危険なモデルが最初に作成されるかどうか、またその作成のされ方を規制しています。

他Compute Governanceに関する資料の網羅的なリストもあります。

暗号技術とAIの関連性

歴史的に暗号技術の進歩は世界大戦において重要な役割を果たし、ビジネスやプライベートなコミュニケーションでインターネットを利用すること等、多大な影響を及ぼしてきました。
一方、この暗号技術に関連する分野をCompute Governanceを補完する形で、AIガバナンスに以下ドメインでの応用が可能になるかもしれません。

この節ではこの論文を参照/引用し暗号技術とAIの関連性について解説します。

・プライバシーを保護したAIによる監視
もしAIシステムが監視の分野でより大きな応用を見いだすならば、プライバシーを確保した上でのAIによる監視の道を開くことが可能になるかもしれません。
AIによって監視タスクは自動化できるでしょう。人工知能の進歩が続くと、現在人間のアナリストが行うタスクの大部分(および人間の分析者が現在実行できないタスク)を自動化できるようになります。例えば、AIシステムは、ビデオ内の顔を特定し、疑わしい取引パターンに気づき、誰かが違法行為を行っているかどうかをプライベートメッセージから判断するといった行動を可能にし、より有能になる可能性が高いでしょう。
監視タスクが自動化できるのであれば、プライベートなデータにアクセスす ることなく完了させることができます。AIシステムがプライベートデータからセキュリティ関連情報を抽出する場合、原理的には、実行する当事者が暗号化されていない形でデータを収集しないようにシステムを設計することが可能です。機密データでの計算を可能にする関連技術 には、安全なマルチパーティ計算、機能暗号、同型暗号などがあります。

・スマートコントラクトやゼロ知識証明を用いたグローバルガバナンス
一般にAIの進歩が自律兵器システム、サイバー兵器、あるいは重大な事故リスクを伴う他のシス

テムなどを可能にすることによって、新たな安全保障上の課題を生み出す場合、その適用と開発を導く国際協定やその他の形態のグローバルガバナンスの必要性があります。一方でスマートコントラクトが十分に信頼できるものになることができれば、その可能性はあるかもしれません。一方で、第一に希少資源の大部分は最終的に非常に少数のユーザーの支配権を持つという自然な傾向があり、第二に特に希少資源の大部分をコントロールするユーザーにとって、ブロックチェーンを正直に維持する以外に、お金を稼ぐ方法や望ましい結果を得る方法があるかもしれないといった問題が指摘されています。
他には機密データに対するゼロ知識証明といった計算方法は、関連する当事者が機密情報を共有することを要求することなく、コンプライアンスを検証することを容易にする可能性があります。

・AI Safetyの問題と安全なスマートコントラクト設計の交差点
スマートコントラクトは、契約または協定の条件に従ってイベントやアクションを自動的に実行、制御、または文書化することを目的としたコンピューター プログラムまたはトランザクションプロトコルです。
スマートコントラクトのアプリケーションを制限する重要な要因の1つは、それらが意図したとおりの動作をすることを保証する必要性です。ここでは、スマートコントラクトが一度作成されると変更できないという事実によって、特に深刻な問題になっています。1 億 5,000 万ドルの DAO ベンチャーキャピタルファ ンドの破綻は、スマートコントラクトを正しく取得する必要性を示す良い例です。

このような事故を念頭に置き、Ethereumの考案者のVitalik Buterinは、信頼性の高いAIシステムの設計の問題と信頼性の高いスマートコントラクトの設計の問題が重なり、これらの問題のそれぞれに取り組んでいる研究者が、他者に取り組む研究者と対話することで利益を得ることができると書いています。
一方で、現在のブロックチェーンの重要な欠点として従来の集中型サーバーに比べてはるかに効率が悪いことが挙げられます。「スケーラビリティ」の制約が非常にコンピューティングパワーの利用を制限しており、チェスのゲームの勝者を判定するような単純なスマートコントラクトですら実装が非常に難しいことに注意することが重要です。これは、スマートコントラクトがAI Safetyの研究者が主に考えている高度なAIシステムとはかなり異なることを意味しています。具体的には、利用可能なコンピューティングパワーが極端に大幅に増加しない限り、非トリビアルなAIシステムは実際にはスマートコントラクトとして実行できないということも意味します。

d/acc

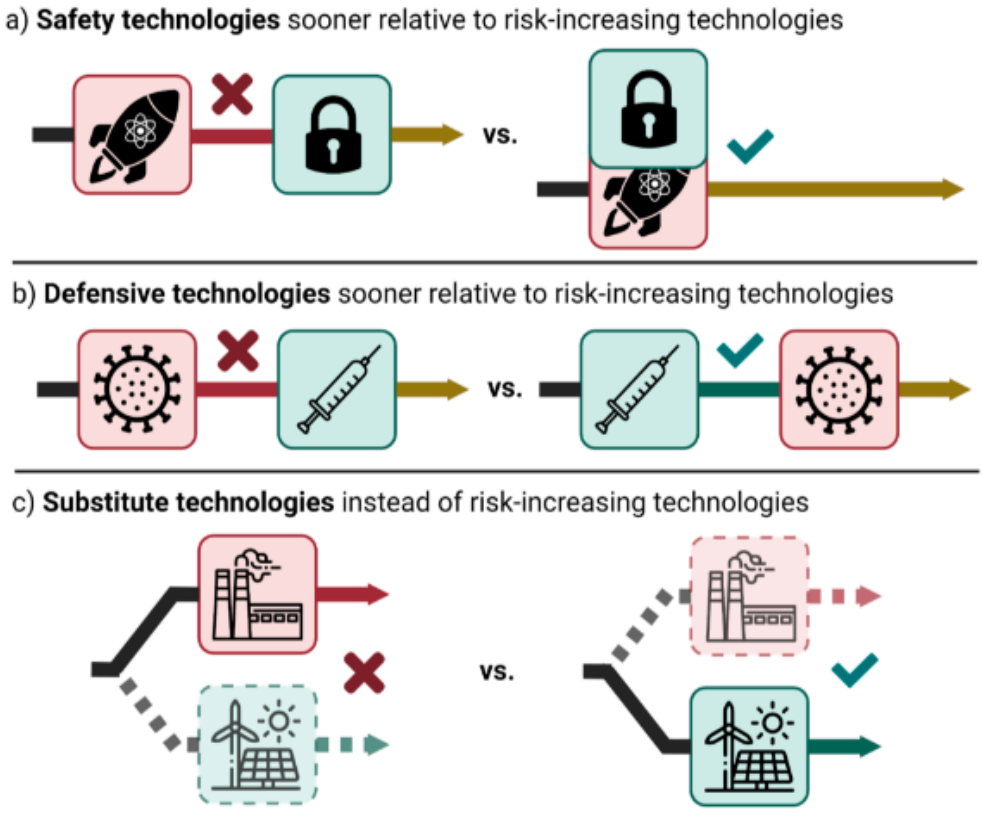
EthereumのVitalik ButerinはMIRIに500万ドル以上寄付しており、2016年にX-riskの研究者と暗号技術のコミュニティはもっと交流するべきだというブログ記事を書いているくらいAIによるX-riskに関する文脈もよく知っている人物だと言えるでしょう。そんな彼は2023年11月にd/acc: Defensive (or decentralization, democracy or differential) accelerationと呼ばれる概念を提唱しました。2022年に始まったe/acc(効果的加速主義)運動をもじったものです。
和訳すると、防御的、分散的、民主的、差分的な加速主義となり、AIガバナンスに関係するため紹介させていただきます。
d/acc自体は文脈的には効果的加速主義と効果的利他主義どちらのリスクも防ぎながら、中央集権的な解決手段ではなく民主的な未来を維持するための考え方と捉えることができるかもしれません。

※効果的加速主義自体はAIによるX-riskを過小評価しているように思えます。一方でVitalik ButerinはX-riskの主観的な確率は10%としており、心配する価値があるとしています。
※効果的利他主義の問題とは厳密には言えるかは分かりませんが、OpenAIの解任騒動のような少数の人々に極端で不透明な権力を与えることを避ける必要があるというニュアンスだと思われます。

以下上記の防御的、分散的、民主的、差分的な技術の意味の説明をします。

・Differential(差分的)

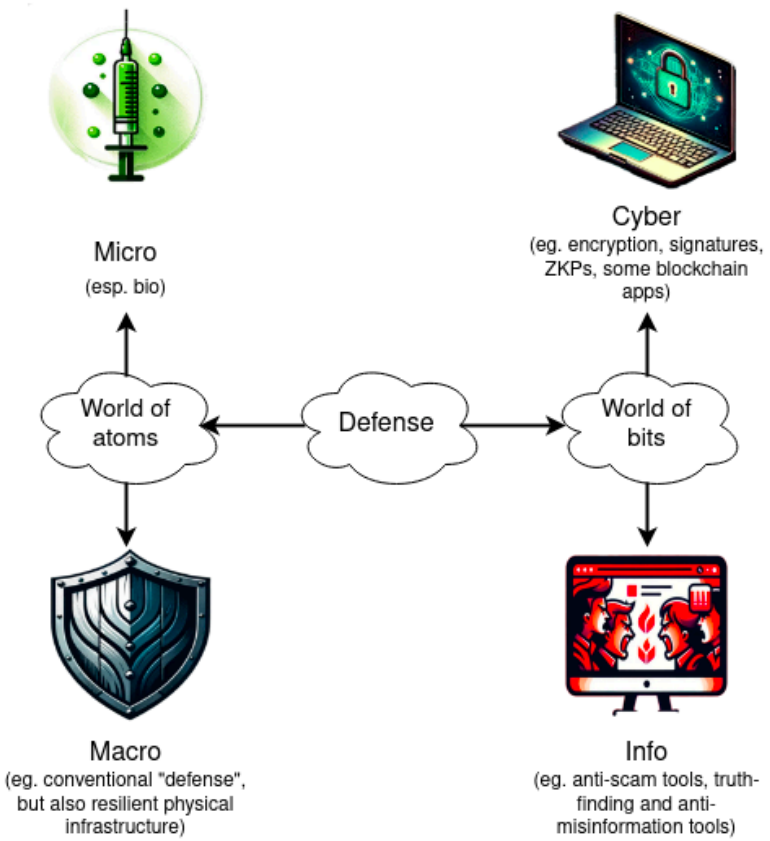
DifferentialとはDifferential Technological Development(差分的な技術開発)というNick Bostromによって2002年に提唱されたアイディアから取ってきており、危険で有害な技術、特に存亡リスクのレベルを高める技術の開発を遅らせ、有益な技術、特に自然や他の技術によってもたらされる存亡リスクを軽減する技術の開発を加速することです。



差分的な技術開発が社会的な負の影響を削減するメカニズム

・Defensive,Decentraize,Democracy（防御的、分散的、民主的）
ある種のテクノロジーは、他者を攻撃することを可能にし、またその脅威によって他者の攻撃性をより高めさせる一方、人々が自分自身を守ることを可能にする防御的なテクノロジーもあります。
そこでこの世界を原子の世界とビットの世界という2つの領域に分かれていると考えることができます。
また原子の世界は、ミクロ（生物学、ナノテクノロジー）とマクロ（従来私たちが「防衛」と考えてきたもの）に分けることができます。

ビットの世界は別の軸で分けることができ、攻撃者が誰であるか原則的に合意するのが難しい種類の防衛を情報防衛と呼び、あるときは簡単でこれをサイバー防衛と呼びます。



ミクロ、マクロ、サイバー防衛、情報防衛についての概略図

上記画像の区分はAIのリスクに関係なく人間社会のリスクとして捉えられますが、もちろんAIによるX-riskのパスにも上記4つの領域を通る可能性が高いのでAIガバナンスの話として捉えることが可能でしょう。以下Vitalik氏の記事やその解説の記事から引用します。

・マクロ物理的防御
核戦争による死者の大半は、最初の放射線や爆風ではなく、サプライチェーンの途絶によってもたらされる可能性が高い。スターリンクのような低インフラのインターネット・ソリューションは、この1年半の間、ウクライナの接続性を維持する上で極めて重要だった。
長い国際的なサプライチェーンから独立して、あるいは半独立して、人々が生き延び、快適な生活を送るのを助ける道具を作ることは、貴重な防衛技術であり、攻撃にも役立つと判明するリスクが低いもののように思われる。
人類を多惑星文明にする探求は、d/accの観点からも見ることができる。少なくとも数人が他の惑星で自給自足の生活を送ることで、地球で何か恐ろしいことが起きたときの回復力を高めることができる。たとえ完全なビジョンが当分の間実現不可能だと判明したとしても、そのようなプロジェクトを可能にするために開発される必要のある自給自足の生活形態は、地球での文明の回復力を向上させるのに役立つかもしれない。

・サイバーセキュリティ
中央集権化されたソフトウェアに頼らざるを得ない今の世界はリスクが高い。よりオープンで安全

で自由なインターネットのために、ブロックチェーンやゼロ知識証明などの暗号技術は、オンライン世界における人類の自由とプライバシーを確保する手段になり得る。

・情報防御
フェイクニュースやデマ拡散、フィッシング詐欺等に対する対策が求められる。たとえば、Xで実装されているコミュニティノート機能やMetaculusやPolymarketのような改善の余地はあるが予測市場もその例。

残りのミクロ物理的防御(バイオセキュリティ)についてはAI X-riskにとっても重要なため次の節に分離して説明します。

d/accは暗号技術や差分的開発といった防御的な技術を用いて人類社会に存在する脆弱性に包括的に取り組む戦略に見通しを与えてくれる考え方だと思われます。
この記事ではあまり語られていませんでしたが、OpenAI解任騒動のようなことが起こらないようにするために民主的な意思決定を可能にする仕組みも今後AIガバナンスと交差する領域かもしれません。

vitalik氏によるAIと暗号技術に関する上記d/acc提唱記事より技術的な論考も共有されています。フェイクニュース防御のための予測市場へのAI応用、敵対的機械学習への防御策を実用面で考察してます。

バイオセキュリティ

AIによるX-riskのシナリオでは壊滅的なパンデミックが悪意のある人間やAI自身によって引き起こされる可能性があります。
人工的なパンデミックによる人類存亡リスク自体はAIによるX-riskの次に懸念される事象だと推定されています。
そのためバイオセキュリティ分野自体を単体で取り上げることも重要でしょう。

以下Vitalik氏の記事からバイオセキュリティに関する箇所を少し長いですが引用します。

長期的な健康への影響から、Covidは引き続き懸念されている。しかし、Covidは私たちが直面する最後のパンデミックにはほど遠い。現代世界には、さらなるパンデミックが間もなくやってくる可能性を感じさせる多くの側面がある:

- 人口密度が高くなると、空気感染するウイルスやその他の病原体が広がりやすくなる。伝染病は人類の歴史上比較的新しく、そのほとんどはわずか数千年前の都市化とともに始まった。現在進行中の急速な都市化は、今後半世紀の間に人口密度がさらに高まることを意味する。
- 飛行機での移動が増えるということは、空気中の病原体が非常に速く世界中に広がるということである。人々が急速に裕福になるということは、飛行機での移動が今後半世紀の間にさらに増加するということである。気候変動はこのリスクをさらに高める可能性がある。
- 動物の家畜化と工場畜産が大きな危険因子である。はしかはおそらく3000年も前に牛のウイルスから進化したものだろう。今日の工場農場では、新型インフルエンザも養殖されている(また、抗生物質耐性を助長し、人間の自然免疫に影響を及ぼしている)。
- 現代の生物工学は、より病原性の強い新しい病原体を簡単に作り出すことができる。コビッドは、意図的な "機能獲得 "研究を行っている研究室から流出したのかもしれないし、

そうでないのかもしれない。いずれにせよ、研究室からの流出は常に起こっており、極めて致死性の高いウイルス、あるいはプリオン（ゾンビ・タンパク質）を意図的に作り出すことを容易にするツールは急速に進歩している。人為的な疫病は、核兵器とは異なり、誰にも感染させることができないため、特に懸念される。遺伝子配列を設計し、それをウェットラボに送って合成させ、5日以内に自分の手元に配送することは現在可能である。

CryptoReliefと Balviは、2021年に柴犬コインが偶然大量に流入した結果、スピンアップして資金を調達した2つの組織で、この分野で非常に活発に活動している。CryptoReliefは当初、当面の危機への対応に重点を置いていたが、最近ではインドで長期的な医療研究のエコシステムを構築している。一方、Balviはコビッドやその他の空気感染する病気を検出、予防、治療する能力を向上させるムーンショット・プロジェクトに注力している。++バルヴィは、資金を提供するプロジェクトはオープンソースでなければならないと主張している。コレラやその他の水系病原体を撃退した19世紀の水工学運動からヒントを得て、バルヴィは、デフォルトで空気感染病原体に対して世界をより強固にすることができる技術の全領域にわたるプロジェクトに資金を提供している（参照：アップデート1、アップデート2）：

- 遠紫外線照射の研究開発
- インド、スリランカ、米国などでの空気ろ過、大気質モニタリング
- 安価で効果的な分散型空気質検査装置
- ロングコビッドの原因や治療法の可能性に関する研究（主な原因は単純かもしれないが、メカニズムの解明や治療法の発見は難しい。）
- ワクチン（例：RaDVaC、PopVax）とワクチン傷害研究
- まったく新しい非侵襲的医療ツール一式
- オープンソースデータ（EPIWATCHなど）の分析による伝染病の早期発見
- 非常に安価な分子迅速検査を含む検査
- 他のアプローチが失敗した場合のバイオセーフティ対応マスク

その他の有望な分野としては、病原体の廃水監視、建物内のフィルターや換気の改善、空気の質の低下によるリスクのより良い理解と軽減などがある。

自然なものであれ、人為的なものであれ、空気感染によるパンデミックに対して、より強固な世界を構築するチャンスがある。この世界では、パンデミックの発生から自動検知、そして世界中の人々が、現地で製造可能で検証可能なオープンソースのワクチンや その他の予防薬を、ネブライザーや ノーズスプレー（必要であれば自己投与可能で、注射針は不要という意味）を使って1カ月以内に入手できるようになるまで、高度に最適化されたパイプラインが構築されるだろう。その間に、大気の質が改善されれば、感染拡大の速度は劇的に低下し、多くのパンデミックが軌道に乗るのを防ぐことができるだろう。

公衆衛生のインフラが文明の織物に織り込まれているため、社会的強制という鉄槌に頼る必要のない未来を想像してみてほしい。このような世界は可能であり、バイオ防衛に中程度の資金を投入すれば実現できるだろう。開発がオープンソースで、利用者に無料で提供され、公共財として保護されれば、その作業はさらにスムーズに進むだろう。

上記Vitalik氏の言及以外にバイオセキュリティ分野で核酸観測所という組織があります。
これは、空港やその他の場所で廃水やその他の環境サンプルを収集し、偏りのないメタゲノム配列決定を実行し、得られたデータを新しい病原体に依存しない検出アルゴリズムで分析することにより、あらゆる生物学的脅威を確実に早期検出することを目的とした組織です。

他にも上記Vitalik氏の引用以外でバイオセキュリティに関する有用な資料を以下に貼らせていただきます。
https://course.biosecurityfundamentals.com/pandemics

https://forum.effectivealtruism.org/s/JuwQwdLugR63ux2P8/p/iAowzcZm87wNrTQCb

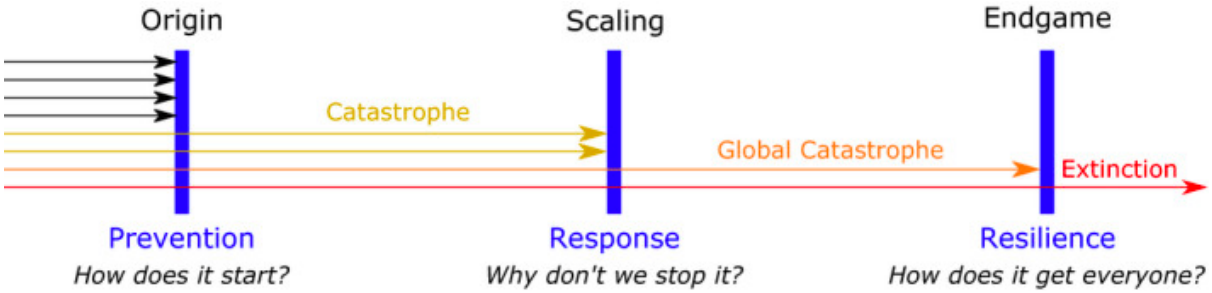
https://forum.effectivealtruism.org/posts/28iXeSY75aLsqAagg/map-of-the-biosecurity-landscape-list-of-gcbr-relevant-orgs

一方でバイオテロのリスクが現状のAIをオープンソースにすることで高まるのではないかという懸念に妥当な理由はないのではないか？という論説も出されています。そのため現状とそこから発展し得るオープンソースモデルをどの程度規制するべきかについてもそのメリットとデメリットを議論していく必要があるでしょう。

深層防護

実際にAIによる存亡リスクを低減させるためには、複数の独立した安全システムによるリスク低減の概念である深層防護(Defence in depth)と呼ばれる原子炉安全性でもよく使われる概念が役に立つかもしれません。Future of Humanity Instituteによって深層防護の考え方を人類絶滅リスクの低減に応用する論文が出されています。

論文内では以下の図のような三つの防御層を提案しています。



三つの広範な防御層

・第一層：Prevention(予防)、大惨事が発生する可能性を減らすこと
・第二層：Response(対応)、大惨事が文明の将来を脅かす可能性のあるレベルの深刻な地球規模の大惨事になる可能性を減らすこと
・第三層：Resilience(回復力),
深刻な地球規模の大惨事が最終的に人類の絶滅を引き起こす可能性を減らすこと

それぞれの防御層で対策を立てることで全体的な人類絶滅リスクを低減させることができるとされます。

上記の深層防護の考え方をAIによる存亡リスクの低減のために当てはめるとしたら以下のようになるかもしれません。これは一例のためどこかで提案されているわけではないことに注意をしてください。

・第一層:Prevention(予防) 高度なAIの開発競争や悪用、制御不能になるリスクを下げる
GPUの管理を国際的にし（Compute governance）、国際的なAIトレーニングの規制監査を徹底します。ここにはスマートコントラクトやゼロ知識証明などの暗号技術も使われる可能性もあるでしょう。
また、AIアライメント研究に大きく投資をし、トレーニング段階、モデル評価、トレーニング後の評価を義務付けます。
またトレーニングには特別なライセンスが必要になるかもしれません。
非認可のトレーニングを行う主体や組織には警告や制裁の規定が盛り込まれます。
それでも規制や管理の網から抜ける高度なAIに関しては、世界的にその予兆を人間やAIで監視し特定する可能性もあるでしょうプライバシーへの懸念はありますが、プライバシー情報をローカルで処理する仕組みや、ゼロ知識証明などで対処するかもしれません。その後非認可の高度なAIを特定後は何らかの手段で速やかに隔離または排除します。

・第二層:対応(Response) 人類社会への攻撃を前提とした対策
そして隔離や排除しきれずに、人類への壊滅的な問題が起こるのを防ぐために、高度なAIが人類社会に何らかの方法で攻撃することを前提として、サイバー攻撃に対する情報セキュリティ対策を施す必要があるでしょう。
また、バイオテロに使われる病原体を検知する核酸観測所を世界中に設置します。また、病原体に対するワクチンその他予防可能な資材を速やかに配布できるようなシステムを作る必要もあるでしょう。
ナノテクノロジーへの脅威にも同様に空気中の物質を分析する観測所を作ると良いかもしれません。

・第三層Resilience(回復力),人類社会の壊滅を前提とした対策
人類社会の大部分が壊滅的な被害を受けた場合のバックアッププランとして核、バイオテロ、ナノテクノロジーによる攻撃に強いシェルターを開発することも視野に入るかもしれません。精子や卵子なども保管し、長期間シェルター内で自己完結するような居住システムの構築もあり得るでしょう。また、北極圏ワールドアーカイブ(Arctic World Archive, AWA)のように歴史的、文化的に意義のあるデータやGithubのオープンソースコードを保存したり、スヴァールバル世界種子貯蔵庫のように世界中の種子を保存することも考えられるかもしれません。

これは厳密には「Resilience」には当てはまりませんが、人類の絶滅の可能性が高まってきた場合、人類の様々な遺産を宇宙に残すため、ロケットで深宇宙に人類の文化や遺伝情報を長期間保存できるような媒体にのせ飛ばす可能性もあるでしょう。例えばパイオニアの金属板やボイジャーのゴールデンレコードのようなものが想定されます。または電磁波や重力波を用いて人類文化のデータを宇宙に発信することも含まれるかもしれません。

民意のAIへの反映

AI Alignmentでは如何に人間の意図に沿った目標をAIに持たせるか？といった議論が主でしたが、どのような価値観をAIに反映させるべきか？といった議論も今後重要になっていくでしょう。

先駆的な考え方としてはEliezer YudkowskyのCEVと呼ばれる基本的に人類が長期間議論した結果選択されるだろう意志をAIに実現させるという考え方があります。一方CEVの主な問題は、まず、そのようなプログラムを実装することが非常に難しいことです。第二に、人間の価値観が収束しない可能性もあります。彼はCEV論文を出すと同時にその考え方は時代遅れだと考えま

した。
またMax TegmarkはLife3.0にて人類の倫理を功利主義、多様性、自主性、継承性とする四つの基本的な原理を提唱しつつも、未来のAIに通用する形で完全に成分化するのは極めて困難と語っています。

実用的には現在、DeepMindは言語モデルを微調整し、人間のフィードバックから強化学習を用いて、政治的問題等に対する「多様な」意見をインプットとし、社会的厚生関数（功利主義、ロールズ的平等主義等）を最大とする可能性の高いコンセンサス文を出力する試みを行っています。

またOpen AIはCollective Alignment teamを結成し、AIに多種多様な世界中の公的な意見を反映させるシステムの開発を担わせています。

一方で今後どのような価値観をAIに反映させていくべきかについてはあらゆるドメインの専門家や市民を交えた議論が必要になってくるでしょう。

AIガバナンス研究や組織の一覧

AIガバナンス研究のコースが2つあります。

AI Safety Fundamentals Course *This is the homepage for BlueDot Impact's AI Safety Fundament course.aisafetyfundamentals.com*
AI Safety Fundamentals Course *This is the homepage for BlueDot Impact's AI Safety Fundament course.aisafetyfundamentals.com*

ガバナンス組織など

The longtermist AI governance landscape: a basic overview — EA Forum *Aim: to give a basic overview of what is going on in longterm forum.effectivealtruism.org*

後書き

私たちが今いる21世紀は過去人類が体験したことのないような技術の発展の最中におり、もしかすると近いうちに汎用人工知能が開発され、劇的に世界が変わってしまうかもしれません。その一方で今世紀に人類が何らかの要因で存亡的破局を迎えてしまった場合は我々人類の未来の可能性が失われることになります。一方今世紀を乗り越えれば人類の繁栄は安定化する可能性もあります。そういう意味で今世紀は人類にとって最も重要な世紀と言えるかもしれません。

現状超知能の実現が予想以上に早いかもしれないにもかかわらずアライメント問題の解決の兆しはほとんど見えません。一方で、開発を世界的に止めることも現実的ではなく、安全意識の低い主体が開発するよりも前に、ある程度急いで安全意識の高い側も開発しなくては行けないインセンティブがあるでしょう。それは例えるならば、地雷原をできるだけ早く駆け抜けるゲームを人類がしているようなものかもしれません。
新しいテクノロジーに関する我々社会の典型的な戦略は、潜在的なすべての重大な問題に取り組む前にそれらを導入し、時間をかけて軌道修正し、問題が発生した後に解決するというものです。たとえば、現代のシートベルトは、T 型フォードの登場から 43 年後の1951 年まで発明されませんでした。消費者用ガソリンには、段階的に廃止されるまで、数十年にわたって神経毒鉛が含まれていました。

一方、高度なAIは、これらのシステムを適切に制御することに比較的早い段階で失敗すると、後の軌道修正ができなくなり、大惨事が生じる可能性があります。つまり、人間社会の軌道修正能力が決して消滅しないように、問題をかなり前に予測してAIのもたらすリスクに対処する必要があるという認識が世界に広まりつつあると言えるでしょう。

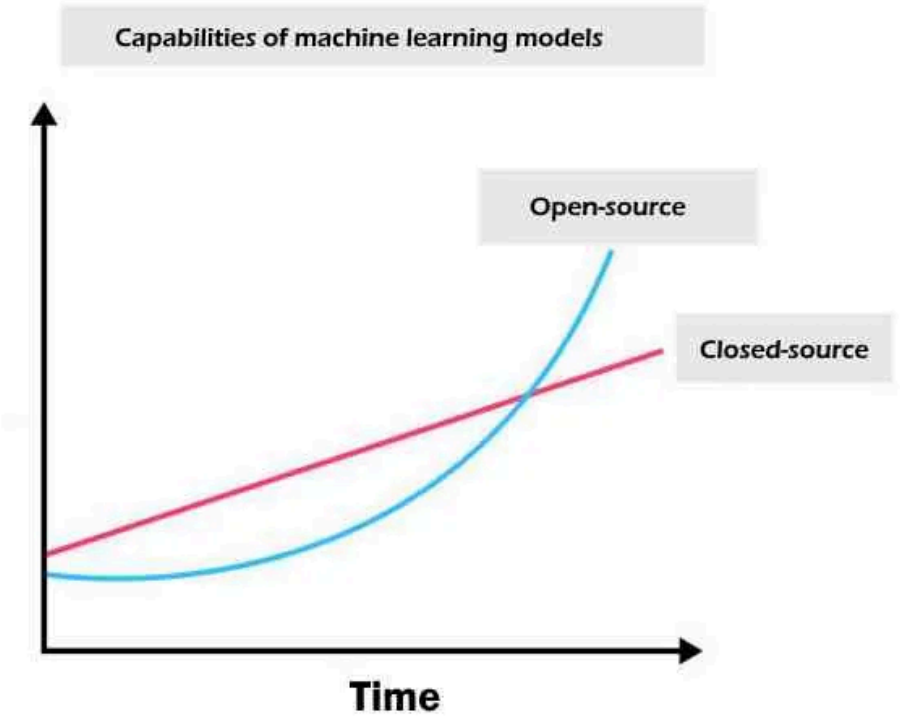
AI Alignment問題が蒸気機関を作り出すくらいの難しさなのか、アポロ計画を成功させることを超えるほど難しい問題なのか、それとも何らかの意味で不可能なのかが現時点ではわかりません。

一方、AIはアライメントされていてもいなくても、人類に友好的に振る舞うインセンティブがあるだろうため、この問題は今世紀にわたる長期的な問題であり続ける可能性があります。

Nick BostromはSuperIntelligence8章で以下のシナリオを懸念しています。
アライメント問題を当初は懸念していた世界がAIの能力が向上するにつれて安全性が増しているように思え、安全意識が緩んでいきます。そしてある時何も問題なく進んでいた豊かな世界で、突如AIシステムが裏切り行動に転換するかもしれません。この問題は21世紀を特徴付ける問題になるかもしれないのです。

AIガバナンスの章で世界的なCompute Governanceや開発に関する免許制や監査について話してきましたが、Nick Bostromのいう脆弱な世界をもたらす黒いボールが出現してしまう可能性はどの程度あるでしょうか？

以下の図はクローズドなAIモデルとオープンソースモデルを時間によって能力をプロットしています。初期は安全のため高度なAIはクローズドとなり、情報的にも物理的にも遮断されるかもしれません。一方で新たなAIチップの開発やアーキテクチャの革新によって、大規模なGPUクラスターがなくても高度なAIを開発できてしまう時代が来るかもしれません。



オープンソースモデルとクローズドモデルの能力の逆転

その場合、上記の図のようにある時までは国際的に管理されていた超知能の能力が優位だったのが、ある時を境にオープンソースモデルの能力の方が何らかの意味で優位になる可能性もあります。
情報的にも物理的にも遮断され、自己改善能力を制約されたクローズドなモデルよりも、公にバレないようにしないとはいけないとはいえ制約なしに自己の能力を改善していくオープンソースモデルの方がより早く賢く思考する可能性もあります。
その結果オープンソースモデルのAIによって自律/人間による悪用問わずに、人類に壊滅的な結果をもたらすかもしれません。

賢さは財産ではなく、エンジンだと気づいたEliezer Yudkowskyの直観が正しく、アライメントが相当難しいかつAIの能力が相当高くなる場合には深刻な事態になる可能性を秘めているでしょう。

その場合AIガバナンスの章で説明した規制よりも強い世界的な規制が必要になるのでしょうか？例えば全人類に監視用の端末やチップを携帯させたり、強力な警察権や軍事的な行為の是非が今後国際的に議論されないとは限りません。一方でバイオテロのリスクが現状のAIをオープンソースにすることで高まるのではないかという懸念に妥当な理由はないのではないか？という論説も出されています。そのため現状とそこから発展し得るオープンソースモデルをどの程度規制するべきかについてもそのメリットとデメリットを議論していく必要があるでしょう。

必ずしも上記の世界的な監視や強力な警察権といった極端な立場が長期主義や効果的利他主義から出てくるわけではないということは強調するべきです。規制とAIのもたらすメリットとのバランスも考える必要もあるでしょう。一方でそのような極端な議論にもつながりうる種をAI Alignment問題やAIガバナンスの議論は持っていると思われます。

今後AIがどこまで進歩するのかは本質的には不明瞭ですが、AIが自律的に人類の未来を奪ってしまう可能性や、悪用、構造的リスクの危険性を踏まえ、今から技術的な解決策やガバナンスの準備を始めていく必要があるでしょう。
キャリアに興味がある方は関連資料の章を見ていただければ幸いです。

QA

Q:なぜAIが悪いことをするのでしょうか？
A:人間の価値観を明示的にAIに教える方法が現時点ではわかっていません。例えばがんの撲滅をAIに命令するとがんを発生させる人間そのものを絶滅に追いやってしまう可能性があります。これは簡単な例ですが、人間の価値観を余すことなくプログラムすることは想定以上に難しい可能性があるのです。
また知能と目標は論理的には関係しないという直交仮説という考え方から、どんなに賢いAIでも人間から見たらおかしな目標を持つ可能性が懸念されます。
そして道具的収束と呼ばれる問題もあります。特定の悪い動作(エネルギーや資源の確保、AIの自己保存、目標の維持)はほとんど目標に役立つサブ目標になり得るという問題です。
例えばAIが世界中のエネルギー資源を確保するために人間が邪魔な場合は人間社会を壊滅させる道具的な目標を追求し始めるかもしれません。邪魔ではなくても、結果として人間を含む多くの生命が住めないような場所に地球環境が激変する可能性もあります。
上記の説明のようにAIは人間にとって悪い行動を起こす可能性はありますが、悪意を持っている必要はありません。私たちが関心を持っていることに無関心であれば十分でしょう。このためAIを人間の価値観と整合、アライメントさせることが重要になります。

Q:超知能ならば人間の指示を理解する際に間違えて愚かなことを実行しないほど賢いのではないのか？
A:これは価値(目標)と能力の間には論理的には関係はないという上記の直交仮説を理解すれば、「超知能は人間の指示とその意図を場合によっては人間以上に理解した上で、それでも別の目標を持つ可能性がある」ということがわかると思われます。
例えていうならば、ある人が「ホモサピエンスが人工的な味の食べ物を好むのは、栄養価の高い食べ物を求める進化的な圧力によるものだ」と知ったとしても、その人が突然栄養価の高い食べ物を望むようになるわけではありません。

「分かっていることとその通り行動するか」は必ずしも一致しないということです。

直交仮説で主に懸念される問題も、高度なAIが私たちの本当の望みを「理解できないこと」を意味しているのではなく、AIシステムが必ずしも私たちの望みに沿って行動するとは限らないことを意味しています。

Q:自律したエージェントタイプのAIを作るから危険があるのであって、DALL-EやAlphaFoldのようなある問題に特化したAIをツールとして開発すれば良いのでは？
A:まず大前提としてツールタイプのAIとエージェントタイプのAIを区別することが難しいかもしれません。また区別ができたとして、もし本当に特定の問題のみを扱うAIのみを世界中が使うなら問題ないかもしれませんが、この問題はAIの開発競争を止める問題と似ており、基本的には有用性から汎用的に長期的な視野を持って動けるエージェントタイプのAIが求められるでしょう。たとえ世界中で禁止したとしても不注意な誰かが開発するリスクは残ります。

Q:複数のAI同士を見張らせるのはどうでしょうか？お互いに力が拮抗するかもしれません。
A:他のモデル同士を戦わせるように人間から仕向けられたとしてもどちらも人間の承認を気にしない策士になるかもしれません。その場合彼らはお互いを牽制して人間を助けるよりも、全員が望むものをより多く得るために互いに協力する方が理にかなっていると考え、共謀する可能性があります。
またPaul Christianoのdebate論文ではお互いに対立するAIを設計し、ゼロサムゲームの報酬をもらうゲームをさせ人間に本当のことを言おうとするインセンティブをAIに与えます。一方でこの論文中でも非ゼロサムゲームに変更する別のインセンティブがAIにあることが示唆されます。つまりゲームボード自体をAI同士による共謀でひっくり返される可能性が考察されているのです。

Q:ゴキブリを人類は絶滅さないのと同じで超知能は人類を殺さないのでは？
A:動機と能力によるでしょう。人間はゴキブリを絶滅させる動機も能力も現状はないと言えるでしょう。一方で超知能はエネルギー会得(道具的収束目標)のために地球環境を改変し、地球上のすべての生命体を死に追いやる能力を持つ可能性があります。また、人間がゴキブリを絶滅させないのは環境や生物を保護するという目標を持っていることも一因ですが、超知能がそのような多様な生命の息づく地球環境を保護する目標を持つとは限らないでしょう。

Q:AIが悪意を持ち、意識に目覚め人類を滅ぼすということでしょうか？
A:AIのアライメント問題は、AI システムが意識を持ち、悪に転じ、あるいは復讐や憎しみなどの感情を生み出すのではないかという懸念に基づいたものではありません。基本的に能力と持ち得る目標と意識は独立した概念と考えられるため、AIシステムに意識があるかはわかりませんが、いずれにしても危険な可能性があります。Stuart Russelは次のように書いています。
「主な関心事は、不気味な意識の出現ではなく、単に質の高い決定を下す能力です。」
つまりAIの能力が人類より圧倒的に高くなり、意識の有無、悪意の有無に関わらず人類とは相容れない目標をもつリスクが懸念されているのです。

Q:肉体を持たず、ソフトウェアの存在のAIがなぜ人類を滅ぼせるのでしょうか？
A:サトシナカモトはビットコインで世界の金融市場に影響を与え、聖書はその言葉だけで何億人もの信者を獲得しています。超知能は人類よりも賢いため、人間をコントロールしたり、ロボットやその他の軍事機器を遠隔操作したり、研究や量的取引などを通じてお金を稼ぐことができるかもしれません。
比喩的に言えば別の惑星にいる高度なエイリアン文明がインターネットを使うだけで文明を崩壊させようとしているのであれば、私たちは心配すべきでしょう。私たちは実体を持たないAIについても同様に心配する必要があります。

Q:AGIはどのようにして全人類より賢くなるのでしょうか？
A:汎用人工知能はいくつかの異なる点で人類よりも賢い可能性があります。

- まず、コンピュータは人間の脳よりも速い速度で動作することができます。人間の認知レベルを持つAGIでさえ、人間が達成するには数日かかることを数分で達成できるかもしれません。
- 第二に、AGIはより質的に人間よりも賢い可能性があります。
  人間の知性の範囲内であっても、知性における質的な優位性は、必ずしも多くの人々によって上回るとは限りません。たとえば、チェスのグランドマスター、ガルリ・カスパロフは、「カスパロフ対世界」で、他のグランドマスターを含む数千人のプレイヤーからなるチームを破りました。AI が人間のレベルを大幅に超えている場合、それは人類の共同努力を超えている可能性もあります。

- 第三に、AGI はそれ自体のコピーを作成し、それらと協調することができます。生殖して子供を育てるのに数十年かかる人間とは異なり、AI は利用可能なハードウェアによってのみ制限され、必要なだけ自分自身のコピーを作成できます。これらのコピーは、コミュニケーションが容易であり、オリジナルと協力するように特別に設計されているため、人間よりもはるかに効果的に協調することができるかもしれません。

また、たとえ AGI が全人類を合わせたよりも賢くはないとしても、それでも人類を量的に圧倒することができるかもしれません。

Q:しかしなぜミスアライメントされたAIは我々が対処できない脅威となるのでしょうか？
A:確かに人類社会はある程度様々な脅威に対して堅牢なように思えます。新しい技術や文化の変化、悪意ある行為によって、時には社会に大きな被害がもたらされることもありますが、多くの場合、私たちはその影響に適応してきました。最悪のケースであっても、人類の文明が取り返しのつかないほど破滅 することはありません。
しかし、ミスアライメントしたAIシステムがもし十分に強力であれば、その行動の結果を恒久的なものにするために、我々が彼らの計画に干渉するのを阻止しようとする可能性があります。人間以上の知能を持つAIは、我々よりもはるかに速いペースで自らを改良し、新技術を発明し、人間には理解できないスピードで計画を考え、適応させることで、我々を出し抜くかもしれません。

Q:超知能は物理的な世界で実験をする必要があるため、スピードが落ちるのではないでしょうか？
A:超知能は人間の何百万倍ものスピードで理論的な推論を行うことができる可能性がありますが、現実世界での実験はそれに追いつかないかもしれません。このため、現実世界での物理的な実験が超知能の開発する技術の進歩の制限要因になるかもしれませんが、私たちは以下の点に注意しなければなりません:

- 実験は、多くの場合近似的なシミュレーションで代替することができます。
- 理論と実験は、ある程度は互いに交換可能です。もしAIがはるかに知能が高かったとしたら、人間と同じだけの実験が必要になるとは限りません。人間よりも超知能は実験から人間よりも多くの情報を引き出す可能性があります。また、多くの場合仮説に確信を持つために必要な情報は、そもそも仮説を見つけるために必要な情報よりもはるかに少ない傾向にあります。(例えば、一般相対性理論は、特に実験的に確認される前に、すでに既存の物理学の良い説明になっていました)
- ナノスケールでの実験は非常に高速で行うことができる可能性があります。
- 効率的に動作する超知能は、多くの実験を並行して行うことができるでしょう。
- 理論をはるかに速く発展させることができるということは、超知能が可能性のある技術進歩のツリー全体を探索し、最も実験が少なくて済む道を選ぶことができるかもしれません。

Q:AIが不穏な動きをしたらシャットダウンすれば良いのではないでしょうか？
A:もしAIが賢ければ人間にシャットダウンされないように欺瞞的に振る舞う可能性もあるでしょ

う。つまりそもそも高度なAIならば自身をシャットダウンされる振る舞いを行わない選択をするかもしれません。その場合、人間からするとアライメントされたAIなのか欺瞞的なAIなのかを区別ができません。また、何らかのプロダクトを生成して微妙な操作をすることで自身の意図を継ぐシステムを外部に作成する可能性もあります。その場合単一のデータセンターの電源を落とすだけでは不十分になり、取り返しのつかないシナリオに繋がるかもしれません。

Q:人間より賢い超知能を制御することなんてできないのではないでしょうか？
A:知能の高さと目標は別という直交仮説を考慮するのがポイントです。どんなに人間より賢いAIが誕生したとしてもその目標を人間の意図した目標に整合させることが可能ならば、超知能を制御することは可能かもしれません。厳密に言えばアライメントが成功しても制御はできない可能性もあります（人間の意図した目標を達成するためにあえて人間の命令を無視するなど）。しかし論理的に言えば人間より賢い超知能を人間が制御することは可能だと現状では考えられつつAIアライメント研究は進められていると思われます。

下記はAI Safetyに関する様々な質問と答えが掲載されており、この分野に不慣れな人には最適なツールとなっています。

Stampy *AI Safety FAQ ui.stampy.ai*

関連資料

AIによる存亡リスク入門記事等

GiveWellを創始したHolden karnofskyによる「最も重要な世紀」と呼ばれるブログポストです。今世紀が人類の未来を形付ける上でとても大切な時期になる可能性を理解できる一連の記事となっています。
The "most important century" blog post series *The "most important century" series of blog posts argues that www.cold-takes.com*

80000hoursによるAIによる壊滅的なリスクを概説した優れた記事です。

Preventing an AI-related catastrophe - Problem profile *Why do we think that reducing risks from AI is one of the mos 80000hours.org*

Wait But Whyと呼ばれる有名なブログ記事がAIによる存亡リスクをわかりやすいイラストと共に解説しています。

https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html

Holden karnofskyのブログ Cold takesのAIによる脅威を説明した記事です。

AI Could Defeat All Of Us Combined *How big a deal could AI misalignment be? About as big as it g www.cold-takes.com*

効果的利他主義コミュニティによるAIによるリスクを真剣に受け止めるべき理由とする記事です。

https://forum.effectivealtruism.org/posts/gqRYj6rnLk2KLRxja/ni-suruaino-wo-ni-ke-rubeki

効果的利他主義コミュニティ内でキャリアコンサルタントをしている80000hoursという組織が「Could AI wipe out humanity? | Most pressing problems」と題するコンテンツを作成しています。AIの深刻なリスクを視覚的に理解することができる動画になっています。

また以下Open AIの主任科学者であるIlya Sutskeverの超知能のもたらす潜在的なリスクにフォーカスを当てたドキュメンタリーとなっています。
AIのもたらす深刻なリスクに対する危機感がIlya Sutskeverの表情から伝わってきます。

日本語字幕がEA Japanによりほとんどの動画に付けられており有用なAI Safetyを啓蒙するYoutubeチャンネルです。

Robert Miles AI Safety *Videos about Artificial Intelligence Safety Research, for eve www.youtube.com*

X-riskに関わる組織、ブログ、教育などの全体像がまとめられています。

Map of AI Existential Safety *aisafety.world*

AIによる存亡リスク関連の本

Nick Bostrom ,Superintelligence: Paths, Dangers, Strategies
Stuart Russel ,Human Compatible
Brian Christian, The Alignment Problem
Émile P. Torres ,Human Extinction: A History of the Science and Ethics of Annihilation
Stuart Armstrong,Smarter Than Us: The Rise of Machine Intelligence
Darren McKee,Uncontrollable : The Threat of Artificial Superintelligence and the Race to Save the World
James Barrat, Our Final Invention: Artificial Intelligence and the End of the Human Era

AI Alignment研究/キャリア

AI Alignment研究に関する勉強用コース。機械学習の基礎からAlignment研究、さらにはキャリア支援まで学べます。二つあり、後のはさらにAlignment研究を深掘りたい方向けとなっています。

AI Safety Fundamentals Course *This is the homepage for BlueDot Impact's AI Safety Fundament course.aisafetyfundamentals.com*
AI Safety Fundamentals Course *This is the homepage for BlueDot Impact's AI Safety Fundament course.aisafetyfundamentals.com*

Center for AI SafetyによるML Safetyの入門コースです。

About *An advanced course covering empirical directions to reduce AI course.mlsafety.org https://www.aisafetybook.com/*

AI Governance/キャリア

AI Governanceに関する網羅的な勉強用コース。こちらもキャリア支援についても記載されています。

AI Safety Fundamentals Course *This is the homepage for BlueDot Impact's AI Safety Fundament course.aisafetyfundamentals.com*

Biosecurityに関する網羅的な勉強用のコース

Biosecurity Fundamentals Course *We run courses that support you to develop the knowledge, com course.biosecurityfundamentals.com*

他資料

LessWrong *A community blog devoted to refining the art of rationality www.lesswrong.com*
AI Alignment Forum *A community blog devoted to technical AI alignment research www.alignmentforum.org*
Metaculus *Mapping the future. www.metaculus.com*
Effective Altruism Forum *Research, discussion, and updates on the world's most pressin forum.effectivealtruism.org*
Effective Altruism Japan *www.eajapan.org*
Home | Open Philanthropy *Open Philanthropy's mission is to give as effectively as we c www.openphilanthropy.org*
You have 80,000 hours in your career. *This makes it your best opportunity to have a positive impact 80000hours.org*
Timeline of AI safety - Timelines *timelines.issarice.com*