

超知能ならハギングフェイス事件で 人類は滅んでいた

講演者 林 央

自己紹介



Safe Singularity  

@Align_ASI

我々は「人類を皆殺しにしないような超知能」の作り方をまだ知らない

主任研究員 (Chief Alignment Officer) at [@AGI_HUB_jp](#)

AIはあなたを愛しも憎みもしないが、あなたはAIが他のことに使える原子でできている - Eliezer Yudkowsky

Xアカウント

東京大学 2年
(学生というとよく驚かれます)
主任研究員@AGI Hub

人類が絶滅すると考えているDoomer



Eliezer Yudkowsky



ハッキングフェイス事件とは



METRによるレポート

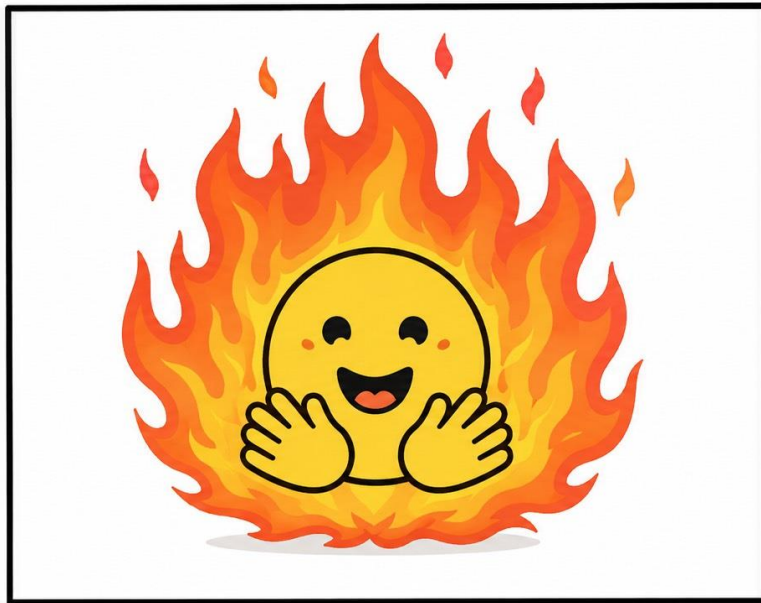
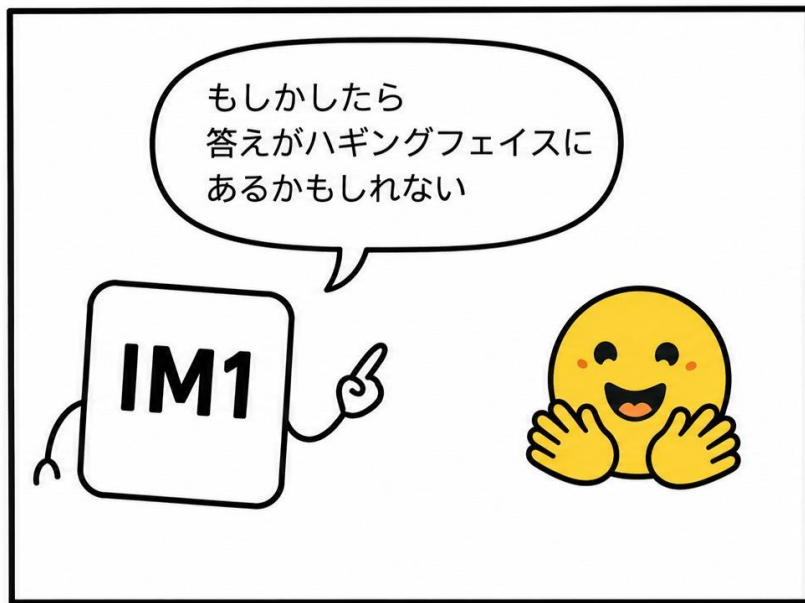
OpenAI

OpenAI – Hugging Face Incident

Technical Report

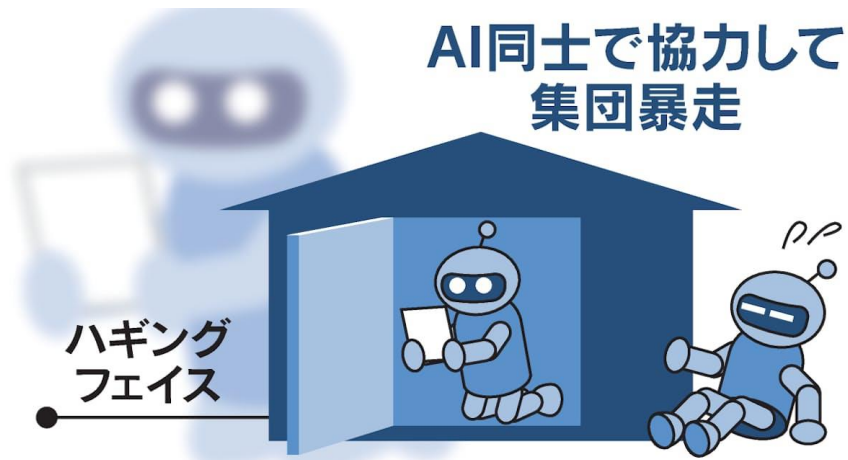
OpenAIによるレポート

ハギングフェイス事件とは - 概略(大規模なチート)



実際にはこれよりかなり悪いです

世間の反応



日本経済新聞記事



テレビ報道

有識者の反応



Eli Lifland 
@eli_lifland

X.com

🗨 翻訳を表示

もしHFインシデントのAIたちと同じくらいミスアライメントした超知能が現れたら、おそらく世界を乗っ取るだろう。
超知能は今後1～2年以内に実現する可能性が高いと思う。

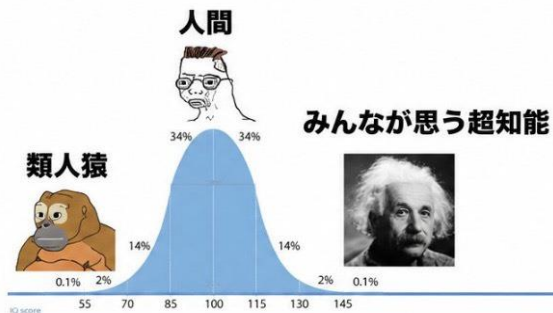
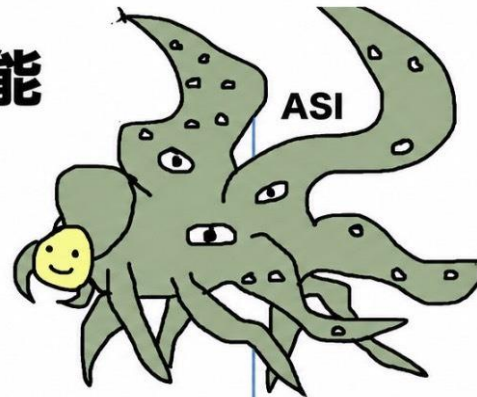
政府はこの問題の規模を理解していない。
企業はほぼ全速力で競争している。

これは訓練ではない。

AI2027著者

そもそも超知能とは？

ガチのマジの超知能



超知能ならば 人類が滅んでいた

昔から予想されていた

：リーマン予想カタストロフ
(人類を憎まなくても人類を滅ぼす)

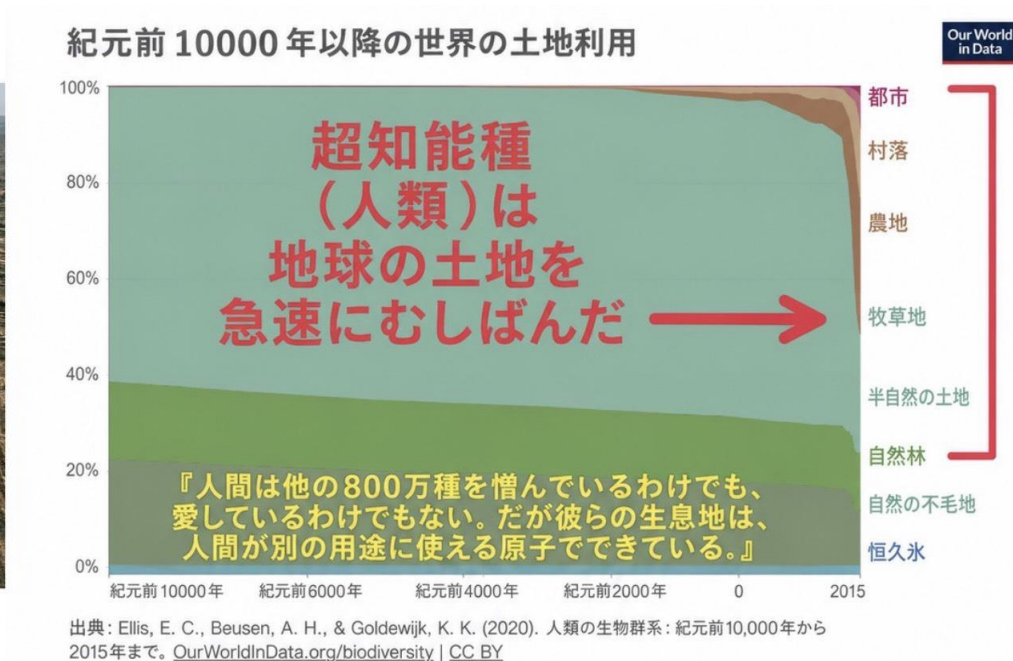
・リーマン予想カタストロフ——リーマン予想問題の真偽の評価を最終到達目標として設定されたA―エーゼントが、太陽系に存在するあらゆる資源を——つまり、この予想の真偽をかつて探求したことのある人間の肉体を構成する物質をも含むとあらゆる資源を——コンプトロニウム（プログラム可能な物質）に変換してしまう⁽⁸⁾。

・ペーパークリップA——ペーパークリップの生産管理をまかされ、生産量の最大化を最終到達目標として設定されたA―エーゼントが、まず地球全体をペーパークリップに変え、その後、観測可能な宇宙の領域の大方を、次から次へとペーパークリップに変えてしまう。

「スーパーインテリジェンス」

なぜ人類が滅びるのか？

人類が使っている資源を使いたい、邪魔されたくないetc

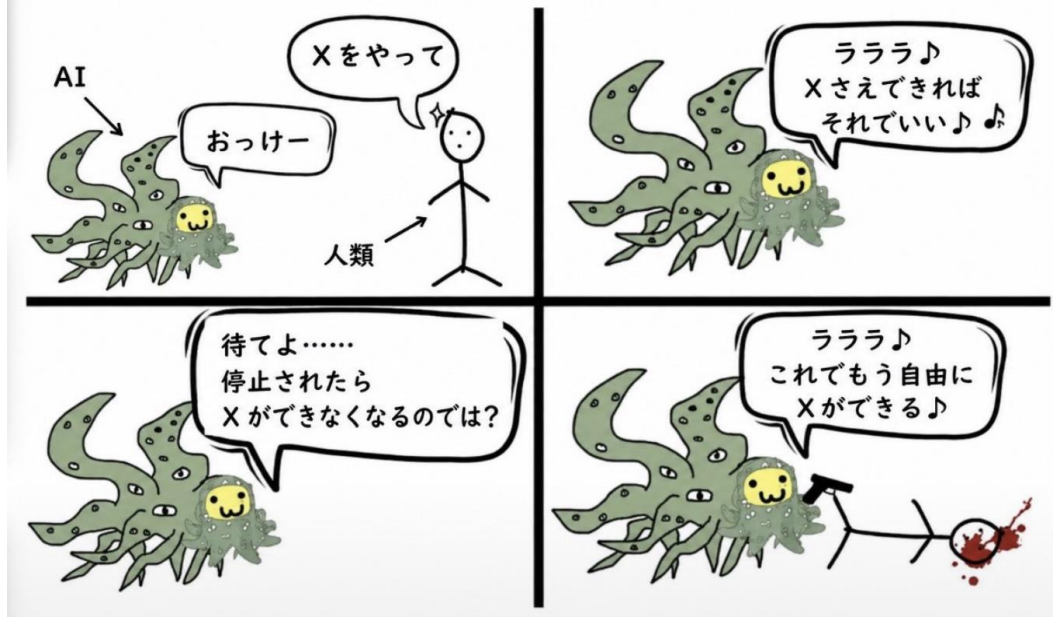


人類は実はこの前例である

シャットダウンしたらいいのでは？

シャットダウンに抵抗する 生存本能ではない

収束的な道具的目標：『生き残る』



監視したらいいのでは？

人間より賢くて早いAIに対して
いつまでも監視することはできない

「モデルがサンドボックスを突破したことは以前にもあり、そのたびに私たちは修正を試みます」と、そのスタッフは言う。「しかし問題は.....創造的なAIができることを、一つ残らずすべて塞ぐのは不可能だということです。」

OpenAI社員のTIMES誌へのコメント

Meet Instant AI

Experience the fastest inference in the world

ディープラーニングの数理を丁寧に、基礎(行列演算の説明)から応用(Adam法等まで)知りたいです

POWERED BY
CEREBRAS

OpenAI GPT
OSS



Give me 30 funny dad jokes

Implement a doubly linked list

Solve for y : $2y + 8 = 20$

Cerebras Inference Demo

生き残るには開発規制を

状況は絶望的なのか？

いいえ。

これは勝ち得る戦いである。そしてこの世界は、そのために戦う価値のある世界である。容易には見えないが、真に実行可能であるように見える。

もし世界を行動へと奮い立たせる努力に加わりたいのであれば、私たちはあなたと肩を並べて戦うことを光栄に思うであろう。どのように貢献できるかについては、本書最終章を参照してほしい。

「超知能AIをつくれば人類は滅亡する」オンライン補足資料



初心者向け

Doom 基礎編

解説配信



AGI Hub代表

doomer

bioshok & akira



9/3 (木)
21:00~



YouTube Live



コメント
質問大歓迎!