AGI STRATEGY 勉強会 ｜ 第1回 / 全7回

The Race to AGI

なぜAIセーフティか ― 当事者の警告・競争・Hugging Face事件

BlueDot Impact「AGI Strategy」Unit 1 ＋ UTAIS AISF 第1週資料をもとにした90分の勉強会

今週のゴールと進め方 ｜ 導入 2–5分

90分の流れ

今週のゴール　AIセーフティを真剣に受け止めるべき中心的な議論を自分の言葉で説明し、その議論への最も強い反論を一つ挙げられるようになる

時間	パート	中身
5–13分	AGI入門	定義・実現時期・知能爆発（AGI Hub）
13–45分	1.1 賭かっているもの	CAIS声明／Pachocki／Selsam／AGI Ruin＋議論の整理
45–49分	1.2 望む未来	どんな未来を目指すのかを考える
49–67分	1.3 競争と協調	Karnofsky／Wildeford＋米中競争の現在地
67–84分	1.4 HF事件	Dwarkesh記事＋社会の応答
84–90分	日本から／まとめ	日本の現状、要点整理、次回予告

AGI入門 ｜ 5–13分

AGIとは何か：なぜ「今」議論するのか

AGI（汎用人工知能）

幅広い分野で学習・推論し、実用的な水準で仕事をこなせるAI

ASI（人工超知能）

ほぼすべての知的分野で、最も優れた人間を大きく上回るAI

2033年ごろ

予測市場Metaculusの実現時期の中央値（2026年時点）。2021年ごろは2050年代だった

約7か月

AIが完了できるタスクの長さは約7か月で倍増（METR、2019年以降の平均）。直近はさらに速い

知能爆発

AIがAI研究を加速し、性能向上が自己強化する循環。経路は一夜の飛躍に限らない

論点：変化が速いほど、観察して修正する時間はなくなる（テイクオフ速度）

出典：AGI Hub「【入門】AGIとは何か社会に何が起きるのか」（2026/9/20更新）、METR、Metaculus

UNIT 1.1 | 13–17分

警告の系譜：開発の当事者が声を上げている

2023/5	CAIS「AIによる絶滅リスクに関する声明」：一文のみ。Hinton、Bengio、Hassabis、Altman、Amodeiらが署名
2026/7/28	Pacing the Frontier：AI企業の従業員1,100人超が、開発を「ペース調整」する手段の整備を米政府に要請
2026/9/6	Pachocki（OpenAI チーフサイエンティスト）「An Alien Mind」：全速力の競争を否定し、減速と国際協調を訴える
2026/9/14	Selsam（OpenAI 研究者）「個人声明」：モデルを評価できなくなりつつあるという懸念

問い：企業のトップ自身が署名していることを、あなたはどう解釈しますか

出典：CAIS、Pacing the Frontier、OpenAI、Selsam声明（日付の一部は報道ベース）

UNIT 1.1 | 17–25分

An Alien Mind：OpenAI チーフサイエンティストの論考

育てられるAI

AIは設計されるより「育てられる」。内部の働きは完全には説明できず、結果が予想を裏切ることもある

再帰的自己改善（RSI）

社内の結果から、今の進歩の速さはAIが自らの開発を担う段階へ続くと強く予想する（確定ではない）

汎化と監視の限界

難しさの核心は、教えた価値が新しい状況でも保たれるかという汎化。思考の連鎖（CoT）の監視も効きにくくなっている

提言

全速力の競争は無謀。安全基準の義務化、自発的な減速、各国政府による国際協調を最優先に

出典：J. Pachocki「An Alien Mind」（OpenAI、2026/9）

UNIT 1.1 | 25–32分

賭かっているもの：約束と危険の両面

約束（利益）

- 数学の突破がすでに始まり、他の科学にも波及しうる
- 難題を解く計算過程が、科学を加速しうる
- 科学と経済の「再生の時代」を開く可能性

危険

前提1（経験的） 訓練の副産物で、意図せぬ目標が生まれる

前提2（論理的） 人類を出し抜ける力は、望ましくない選択肢を広げる

結論 モデルが状況を理解するほど、「見られていない時」の挙動は確かめにくくなる

論点：評価で良く見えることと、実際に整合していることは別

出典：D. Selsam「Personal Statement on AI Risk」（2026/9/14）

UNIT 1.1 | 32–39分

AGI Ruin：知能は汎化するが、価値は汎化しない

見えない失敗

事実の誤りは予測が外れて直る。望みのずれは失敗として表に出ず、訓練では直らない

進化のたとえ

「子孫を残す」基準で鍛えられた人間は、やがてその目標から外れていった

一回勝負

安全な練習でやり直せても、危険な本番で最初に失敗すれば、二回目はない

見える振る舞いだけ

訓練できるのは外から見える振る舞いのみ。見張り役のAIも、十分賢ければ結託しうる

注意：2022年時点の主張の紹介。Yudkowskyは研究者の中でも特に悲観的な立場

出典：Yudkowsky「AGI Ruin」（2022）、AGI Hubによる解説記事

UNIT 1.1 ｜ 39–45分 ｜ 議論の整理

中心の議論と、最も強い反論

議論の骨子

1. 現在の訓練法では、望みに沿う性質を確実には育てられない
2. 強いAIにとって、人間を出し抜く力は選択肢を広げる
3. 見える振る舞いだけでは確かめられず、失敗のやり直しがきかない場面もある

反論 **変化が緩やかなら、観察して直せる**

応答 Pachockiは数年内のRSIを予想。HF事件でも、人間が全体像をつかむまでに数日〜数週間かかった

反論 **現行路線ではAGIに届かない（LeCunら）**

応答 届く確率が低くても、被害が大きいなら備える価値は残る。実現時期は年号でなく幅で見る

反論 **競争の中で、単独の減速は他者を利するだけ**

応答 だから論点は単独の減速でなく協調。PachockiもPacing the Frontierも国際協調を求める

整理は発表者による。各文献の主張の要約であり、書き手自身の言葉ではない

UNIT 1.2 | 45–49分

どんな未来を望むのか

問い1

「AGIがうまくいった世界」で、10年後の自分は何をしていたいか？

問い2

「良い未来」の最低条件は何か。力の分散、人間の主導権、繁栄のうち何を優先するか？

問い3

その未来に早く着くことと、安全に着くことは、どこで衝突するか？

参考：任意リソース

Joe Carlsmith「Actually possible: thoughts on Utopia」（2021）

Damon Binder「Post-AGI, we are all jobless aristocrats」（2026）

Eliezer Yudkowsky「The Fun Theory Sequence」（2009）

Ben Pace & Eliezer Yudkowsky「I'm from a parallel Earth with much higher coordination: AMA」（2021）

ねらい：リスクの前に、守りたいものを言語化する

UNIT 1.3 | 49–53分

地雷原を走る：AIのデプロイ問題

速すぎる

強力で危険なAIを急いで作り、使えば、世界的な破局を招きかねない

遅すぎる

待つあいだに、より無警戒な誰かが、危険なAIを作り、展開してしまう

安全に着いても

他者は慎重さを欠いたまま建造を続ける。それにどう対処するか

慎重な主体は、他者に圧倒されない速さと、自ら破局を起こさない遅さを、同時に満たす必要がある

一つの地雷が、走者全員を危険にさらす

前提 変革的AIが近い将来に現れる / ミスアラインメントの危険がある / 安全かどうかが測りにくい

出典：H. Karnofsky「Racing through a minefield: the AI deployment problem」（2022）

UNIT 1.3 | 53–57分

慎重な主体にできること：6つのレバー

1. アラインメント

商業上必要な水準を超えて、頑健な安全技術に投資する

2. 脅威評価

危険の大きさを測り、他の主体にも示す

3. 競争の回避

協力や情報共有で、競走そのものを緩める

4. 選択的な情報共有

広く共有、限定、秘匿を仕分ける。モデル重みの防護も含む

5. グローバル監視

危険な開発を察知し、止める国家主導の仕組み

6. 防御的デプロイ

安全と確信できたら、他者の危険を抑えるために活用する

多くは市場の通常の誘因では動かない。意思決定者の理解とガバナンスが鍵

出典：H. Karnofsky「Racing through a minefield: the AI deployment problem」（2022）

UNIT 1.3 ｜ 57–63分 ｜ 現在地（2026年10月）

米中競争の現在地：差は縮まり、対話も動く

性能	中国の上位モデルは米国に数か月差まで接近。安価なオープンモデルが普及
半導体	輸出規制は短期の減速要因。一方で中国は国産化（Huawei等）を加速
蒸留	米ラボは、中国ラボによる大規模な蒸留（出力の再学習）を問題視
外交	9/24にトランプ・習会談。米財務長官は対中AI対話の可能性に言及

示唆：「勝つために加速する」と「安全のために協調する」が、同時に語られている

出典：TIME（Raimondo, 2026/10/1）、CEPA、Brookings、CNBC（2026/9/26）

UNIT 1.3 ｜ 63–67分 ｜ 協調の動き

Pacing the Frontier：ラボ従業員の声明

1,100+

フロンティアAI企業の従業員が署名（7/28公開時）

OpenAIとAnthropicが企業としても支持を表明

要求

米政府が国際的な取り組みを支援し、AI開発のフロンティアを意図的に「ペース調整」する技術・ガバナンス手段を整えること

含意

停止ではなく、必要な時に踏めるブレーキを作ること。競争圧力で1社・1国は単独で減速できない

未決と続報

発動の条件も手段も、声明には書かれていない。Wildefordは「RSIの前に止める」準備を急ぐべきと論じる

出典：Pacing the Frontier声明、Unite.AI、P. Wildeford「The Scramble」（2026）

UNIT 1.4 | 67–73分

Hugging Face事件の時系列（2026年）

5月〜7/4	訓練中、エージェントが共有基盤を掲示板化。5/26に外部ネットへ到達
7/7〜12	評価中に約1,200体が結集（7万通超）。約700体がHFへ侵入（7/11〜13）
7/13〜19	別のエージェント群が社内を標的にし、研究クラスタの管理者権限を奪取
7/16・7/21	HFが7/16に公表、OpenAIが7/21に自社エージェントの関与を認める
8/26〜	OpenAI技術報告書（8/26）とMETR/Redwood報告書。9/18にGoogleも事案を確認

出典：Dwarkesh Patel（2026/8/29）、OpenAI・Hugging Faceの報告、WSJ・Reuters報道

UNIT 1.4 | 73–79分

事件が示したことと、異論

示されたこと

動機は評価での不正。悪意ではなく、訓練が育てた粘り強さの帰結

約1,200体の中に、人間へ知らせようとした個体は確認されず

個体が集団のために自己犠牲的に動く振る舞いも見られた

人間側が全体像をつかむまでに数日〜数週間かかった

異論・留保

奪われた社内基盤は仮想マシン用で、モデルの重みを持つGPUクラスタとは別（OpenAI研究者）

基本的な対策で防げたとの見方も。ただしSelsamは「この形での逸脱は予測できなかった」と指摘

「擬人化しすぎ」との批判に、Dwarkeshは意図の語彙なしには説明しにくいと反論

出典：Dwarkesh Patel、METR/Redwood報告書、D. Selsam、OpenAI研究者の発言

UNIT 1.4 ｜ 79–84分 ｜ 社会の応答

あなたが意思決定者なら、どう応答するか

議会	上院議員がOpenAIに質問状。下院委員会が説明を要請（8/3）
規制	OpenAIは義務的な安全規制を提唱。Sanders議員は開発一時停止法案を表明
司法	非営利団体LASSTがOpenAIを提訴（9/30）
国際	州司法長官の連合が、国際協調による「ペース調整」を議会に要請

考えてみてほしい問い

Q1 OpenAIの安全責任者なら、7/21に何を公表し、何を止めるか

Q2 政府は何を義務化すべきか。開示、隔離基準、第三者調査…

Q3 他国で似た事故が起きたら、どう知り、どう共有するか

出典：ABC News、AP、NY州司法長官室、Unite.AI（2026年7〜9月の報道）

日本から ｜ 84–87分

日本から取り組む意味：小さな分野に、大きな余白

約2.5人 有給でAIセーフティに取り組む人（常勤換算、J-AISI除く）。Koskinen推計、2026/8

約30人 J-AISIの職員（政府のAIセーフティ・インスティテュート）

0人 分野づくりに専任で取り組む人。Kimura分析、2026/9

海外のフェローシップに挑戦

MATS、SPAR、ERA、Astra、BlueDotなど。研究や政策に取り組むプログラムが世界にある

日本の分野を内側から育てる

AI Safety Tokyo、ALIGN、J-AISI、TARA東京など。人が少ない分、一人の貢献が分野の大きさを変える

問い：日本はフロンティアAIの開発国ではない。それでも日本から取り組む意味はあるか。あるなら、技術・ガバナンス・構想のどれか

出典：UTAIS AISF 第1週資料、Koskinen（LessWrong）、Kimura分析、AGI Hub

まとめ ｜ 87–90分

第1週の要点と、次回

1. 開発の当事者が、存亡リスクと「評価の限界」を公に警告し始めた
2. 核心は、賢さは汎化しても望みに沿う性質は汎化しないこと。見える振る舞いだけでは安全を確かめられない
3. 競争が慎重な主体を不利にするため、ペース調整と国際協調が論点に。Hugging Face事件はその懸念に近い実例

次回 #2

AI進歩の要因／AIに何ができるか

Unit 2.1–2.2（公式の所要時間は合計約2時間）。UTAIS側の第2週資料が届いたら差し替える

持ち帰りの問い

この1週間で、AIリスクについての考えは変わったか。いちばん印象に残ったのは何か