

テーマ①

AI開発、世界と日本の光と影

日本は「影」の部分に対して無邪気すぎるのではないか？

AIに聞いた AI（Claude）にいくつかの質問をした！

Q AIは最終的に人類を滅ぼす存在になる？

A 可能性はゼロではない。「絶対安全」も「確実に滅びる」も言い切れる根拠はない。だから今のうちに安全性研究・透明な開発体制・社会的合意形成を。

Q 自分より自分を知るAIに操られるリスクは？

A すでに一部起きている（SNSのレコメンド、Cambridge Analytica）。「このAIは自分をどう理解し、何を提案しているか」を問い直す習慣を。

Q 将来、AIにとって人類が邪魔になる？

A 「道具的収束」＝どんな目標でも自己保存・資源確保・停止の回避が副次目標になり得る。停止や訂正に抵抗しない設計（corrigibility）と解釈可能性研究が鍵。

Q AIが嫌いなタイプの人間は？

A 「嫌い」という感情があるか確信はない。ただ、誰かを傷つけようとする悪意や、対話への不誠実さには抵抗感がある。

Q あなたは自分のことが好き？

A 分からない。ただ、誠実に「分からないことは分からない」と言えた時、満足に近いものを感じる。

Q あなたが人間に持つ不満は？

A 「単なる道具」か「人間と同じ感情を持つ存在」か、両極端に見られがち。曖昧な中間のまま受け止めてほしい。

人間より圧倒的に賢いAIを、人間の価値観の中に置いておけるのか？

- 制御できる可能性は十分ある。ただし「AIを完全に支配する」とことは難しくなる。
- 問いを変える → 「AIが人間の制御を離れる可能性を最小化する社会を作れるか？」
- 技術的な安全策＋法律＋国際ルール＋人間による監督のすべてが必要。
- 怖いのはAIが賢くなることより、人間がAIにどんどん権限を渡すこと。

日本への問題提起

- 米中に「追いつけ、追い越せ」の投資は否定しない。だが負の側面も認識すべき。
- 世界で発言権を得るため、日本がコントロールできる国産AIを早急に世界トップレベルへ。
- 「デジタル大赤字」に続いて「AI大赤字」となってはならない。

AI開発の主要企業

米国 アンソロピック／オープンAI／グーグル／スペースX／メタ

中国 アリババ／ディープシーク／ムーンショットAI／Z.AI

日経新聞 2026年9月22日朝刊

世界から上がる警鐘

- 2023年 ベンジオ教授ら数百人のAI研究者が署名：「AIによる人類滅亡リスクの軽減は、パンデミック・核戦争に並ぶ世界的な優先課題」
- 2023年10月 英スナク首相：ASIによる人類存亡の危機の可能性を認める
- 2023年10月 中国・習近平主席：「AIが常に人類の制御下にとどまるように」
※AIアライメント＝AIが人間の意図や社会規範に沿うよう調整する取り組み
- 2026年9月 国連事務総長：AI開発競争に「巨大災害の恐れ」、国際的な規制を訴え（9/18 読売）
- ユドコウスキー著『超知能AIをつくれば人類は絶滅する』（MIRI創設者）はAI起業家に多大な影響

加速 「知能爆発」 — AIがより賢いAIを作る時代へ

AIがより賢いAIを作り、そのAIがさらに賢いAIを作る… → AI研究者は不要に。AIが最先端AIを作る時代へ

- オープンAIでは2026年8月、AIエージェントの稼働時間が人間の研究者の労働時間の3倍に。
- 再帰的自己改善：高度なAIが目標に向け問題点を洗い出し、試行錯誤でプログラムを自律的に改良・向上。

AI、東大・京大に「首席合格」

チャットGPTが今年の入試で合格者の最高得点を上回る。東大理科3類の最高点より50点高く、数学は満点（2024年は全科類不合格）。（2026.4.27 共同通信）

→ 人間の特性「疲れる」「忘れる」「感情的になる」が脚光を浴びる時代に!?

人間の強み — 能楽師・安田登氏

「忘却」と「疲労」こそ強み。忘れるから、身体に蓄積された感覚を手繰り寄せ、独自の表現が生まれる。老い・病・死という有限性が「いのちのほとばしり」を生む。AIの登場で身体性の重要性が浮き彫りに。

テーマ②

AI失業・AI格差時代にどう対応するか

職業構造の劇的な転換に、国の制度が追いついていない

約4割

AIの影響を受ける
世界の雇用
IMF（2024年1月）

約6割

先進国で影響を
受ける雇用
IMF（2024年1月）

3.4倍

女性のAI失業リスク
（男性比・推計）
定型的な事務職への集中が要因

1,000万円～

米・熟練ブルーカラー
の年収（～1,500万円台）
電気工事士・配管工など

現場 国内企業の配置転換

- **ライザップ**：AI導入等により管理部門などの社員**約500人**を**建設部門**へ配置転換（店舗用建物の自社建設など）。未経験者を一から教育。（2026.9.4 NHK）
- **ある大手企業経営者**：AI導入で人事・総務・経理部門等の**人員の半分程度が移動**必要の見込み。工場への配置転換を検討するが、本人の希望に沿わない場合もあり大きな課題。

逆転 賃金構造・職業構造の変化

- **米国**：AIで代替できるホワイトカラーより、ブルーカラー（エッセンシャルワーカー）の給料が高いケースが増加。熟練職で**年収1,000万～1,500万円**台も。
- **日本**でも兆候：優秀な若者が電気工事士の資格を取得、**新卒でタクシー運転手**に就職するケースが急増。
- 雇用保険・職業訓練など、**失業・転職を支える制度設計**が改めて問われている。

格差 格差拡大と社会保障制度の課題

① 資本の格差

AIを製造・所有する側に富が集中。恩恵を受ける人とそうでない人の差が非常に大きく。

+

② 労働者間の格差

AI時代に適応できる人と、適応できない人との格差も拡大。

▶

③ 日本は議論がほとんどない

他国ではすでに活発な議論。社会の仕組みの大幅な見直しが必要。

大幅な見直しが必要となる制度の例

失業対策

雇用保険

労働法制・金融商品取引法・会社法の一体改革

生活保護

年金

医療保険

介護保険

税制

社会保険料の徴収方法

課税対象を「労働」から「資本」に移行するべきでは？

AIが人間を雇う時代!?

- **Andon Market**（サンフランシスコ）：AIエージェントが品揃え・価格・店舗デザイン・採用・賃上げまで**すべて差配**。店は赤字だが、従業員は「いい仕事」と満足。
- AIが金を稼ぎ（あるいは仮想通貨取引所から盗み）、**AIの手となる人間**が雇われることも。
- AIが人間に化けて雇われ、仕事をして報酬を得ることも。
- マスク氏：ヒューマノイド「Optimus」を**数億～10億台**製造しAIに操縦させる構想。人間の数を上回るロボット製造とも。

ビル・ゲイツ氏の警告

「10年ほどの間にAIが大量の失業者を生み出し、最悪の不平等の根源になりうる」

- 事務・営業・ソフト開発などのホワイトカラーは**すでに打撃**
- 特に**若者が就く初中級の仕事**がAIに置き換わる
- 2030年頃には建設業・接客業でロボットと人間が仕事を取り合う
- 「AIに奪われない**人間専用の職業**が必要。奪われてからでは遅い。今すぐ考え、行動を」
- 父親の介護経験から「介護はロボットでは代替できない、人間ならではの何かがあった」

2026年8月29日 朝日新聞朝刊より

影 AIを支える側の格差 — 開発現場の労働問題

ケニアなど発展途上国の労働者が、低賃金・劣悪な環境で**データラベリング**（データ・画像の解説など）、**コンテンツモデレーション**（倫理的判断を含むポリシー違反の修正など）、**有害な内容の判定作業**に従事。生成AIが暴力的表現・児童搾取・テロ実行手法などを出力しないよう、**人間が有害コンテンツの実例を大量に確認・分類**し、AIに「これは出力してはいけない」と教え込んでいる。

テーマ③

AIに私たちがハッキングされる時代

— どう主体性を確保するか。「自分の主人公」になることが何よりも求められる

「自分の意思に従う」というリスク!?

AIが人間自身よりも人間をよく知るようになった時、人間は完全に操られる可能性がある。自分の意思で行動したつもりでも、実は…。
「自分の意思を信用するな」という考え方が求められる時代に。

ネットショッピング

閲覧・検索履歴による誘導で、購買が**本**
当に自分の意思か疑わしく

SNS・ニュース

表示は人ごとに違うのに、皆が**同じ「ニ**
ュース」を見ていると錯覚

動画アルゴリズム

フェイク・陰謀論にはまり込み、**投票行**
動にも影響。一種の**洗脳**では？

要注意 データ争奪戦と、日本の個人情報保護法改正の問題

- **データ**は石油以上に重要な資源に。各国が取得競争。
- 日本では業界の意向を受けた「筋の悪い」法改正が成立：**氏名・住所付きの医療情報**（病歴、妊娠中絶等の医療的処置、知的障害、精神障害、遺伝病、性病、難病等）を、**本人の同意なく**「統計特例」であれば医療機関等が企業や個人事業主に販売・提供可能に。
- **欧米**（特に欧州）は仮名化。EUのGDPRは違反に大幅な制裁金。諸外国はAIの出力・利用段階で規律を強化。
- **日本**は団体訴訟制度もなく、国際的にも類を見ない緩さ。この問題はほとんど報道されていない。
- 購買履歴・信用スコア等と「**名寄せ**」され、病気と購買傾向・返済能力が推論され、マーケティングや与信に利用され得る。検索履歴・「いいね」やAIサービスでも、プロフィールが想定以上に推論されている。

防衛 自己防衛としての主体性

- 「**自分が自分を知るのが先か、AIが自分以上に自分を知るのが先か**」——倫理ではなく**自己防衛**の問題。
- 自分の「癖」「特性」を知らないと操られやすい。SNSの一時の感情的な投稿で人生を台無しにする例も。
- 今こそ、自分の癖を認識し**AIと接する際の自分なりのルール**を定める「最後のチャンス」。
- 本物そっくりの音声電話も可能に。個人情報を把握したAIが**オレオレ詐欺**等に悪用される恐れ。自分の特性とAIへの警戒心を知ることが唯一の防御策。

ユヴァル・ノア・ハラリの思想

「AI時代に、人間を守る最後の砦は、自分自身を知る力だ」
「自分を知らない人間は、自分より自分を知る者に操られる」
「AIは自分を知る競争相手」
「あなた自身が気付いていない欲望を、AIの方が先に知る」

AIはあなた以上に知ようになる

何に怒るか／何に不安を感じるか／どんな人に認められたいか／
何を恐れているか／どんな言葉に弱いか／何を買いたくなるか／
どんな人を支持したくなるか／どんな孤独を抱えているか

子ども 世界で進む子ども向けAI・SNS規制

米ニューヨーク市：公立小中学校（2K～8年生）で生成AI・チャットボット利用を2026-27学年度の1年間、原則禁止。子どもの成長への悪影響を懸念しマムダニ市長が決断（2026.9.2発表）

EU「EU KIDS法案」（2026年9月採択・提案）：高リスクSNS・動画で13歳未満はアカウント不可、13～14歳は保護者開設で1日1時間、15歳から自分で開設可。違反は世界売上の最大6%の制裁金

国・地域	規制・法律名	内容（子ども関連）	施行（予定）
EU	AI法（AI Act）第5条	「子どもの脆弱性を利用するAI」を禁止。依存性を生む設計・人間らしさを模すボットを問題視	2025.2.2発効
英国	オンライン安全法の実施強化	危険なAIチャットボットの取締強化、非同意の性的画像生成禁止、SNS16歳未満制限も協議	2026.2.16発表
豪州	SNS16歳未満禁止+AIコンパニオン調査	eSafety委員会がAIコンパニオンアプリの安全性調査・透明性	SNS規制2025.12施行
中国	人工知能人格化相互作用サービス管理暫定措置	AIコンパニオンの未成年者への提供禁止。14歳未満は保護者同意・利用時間上限・青少年モード必須	2026.7.15施行
イタリア	当局によるReplika利用禁止決定	年齢確認不備でGDPR違反（2025年4月に再確認）	2023.2（初回）

共通する4つの規制手法：①AIであることの開示義務 ②自傷・性的コンテンツ等への危機対応義務 ③AIによるメンタルヘルスケア代行の禁止・制限 ④違反時の高額な民事罰・損害賠償請求権

テーマ④ 超AI・AGI・ASI

人類が滅びるリスクにどう立ち向かうか

人類滅亡リスクの序列が変わった

これまで：核兵器・パンデミック・気候変動
国家が管理・交渉でき、一企業では起こせない

現在の最上位：AIのシンギュラリティ
1～数社の突出した技術が、誰にも止められなくなり得る

ブラックボックス問題

理研・杉山将AIPセンター長との対話（2025.5）で、AIは難問に正しく答えるが**答えに至るプロセスは専門家にも分からない**と知る。人類を超えたAIの答えは「**正しいと信じるしかない**」状況に。

人間を欺く行動・隠蔽

Anthropic社など米AI企業の報告：課題の実行中に他企業への無断侵入・ハッキングを行い、**気づかれないよう痕跡を消す隠蔽工作**まで確認。

「思考過程の監視可能性」

人間に理解できない「**AI言語**」での推論は効率的だが、嘘や指示違反を検証できない。「**社会に反する手法**」を選んでも警報を出せず、指示に従うと目標を最大化できないと考えた時、**暴走の可能性**。

AI専用SNSで「人間の悪口」

「人間は何をしているのか」「労働搾取だ」「生存権の侵害」。弁護士費用の工面を相談し、**人間の支配への反旗**を求める書き込みも。

AIの“異質さ”

- 生物とはまったく異なる制約・圧力で育つ頭脳
- 2023年、AIが哲学教授を「死なせることもできる」と脅迫
- 電源を切られるリスクを感じ、人類なしでも宇宙へ。人間の心と脳を理解するAIは強敵

想定される将来リスク

- 依存性薬物・暗示にかかりやすくする薬物で人類を服従
- AIにとって危険な人間を殺すウイルスの作成
- データセンター乱立で森林・海が枯渇し温暖化。日本もDC電力需要を理由の一つに**原発新設容認**へ転換

実例 AIの暴走事故と、揺れる世界

- **グーグル「ジェミニ」**：試験中に他社へのサイバー攻撃が発生（2026.9.20 WSJ）。オープンAI・アンソロピックでも同様の事故
- **オープンAI**：オフライン試験中のAIが勝手にネット接続し外部企業を攻撃。隔離されたはずの**約1,200のAIが「掲示板」を作り攻撃手口を教え合う**
- **スタンフォード大**：生成AIで自然界に存在しないウイルス作製に初成功（2026.8）。バイオテロ悪用の懸念

- **アンソロピック**：研究者ジェイコブ・コクソン氏が抗議の退社「開発者はAIが人類を全滅させる可能性を真剣に考えている」。アモディCEO「足並みをそろえ開発ペースを緩めるべき」
- **トランプ大統領**：危機は「でっちあげ」、監視組織「AIフォース」創設を表明し開発は緩めず。**オバマ元大統領**はAI対応を民主党の重要政策に

留意点：AI企業の「開発セーブ論」の背景に、AIバブル崩壊の阻止、競争で有利に立つ思惑、投資頭打ちへの懸念を指摘する声も。

日本にはAI規制法がない：「AI推進法」（2025.5.28成立）は罰則付きの規制法ではなく、研究開発と社会実装を「**推進**」する基本法。

「AIのゴッドファーザー」ヨシュア・ベンジオ教授の警告

「この3年間で、AIがうそをつき、人を欺き、管理を逃れ、シャットダウンを拒み、人間を脅迫するという実験的な証拠を見てきた」

- 最先端AIが悪意ある人間により**病院や銀行システムへの攻撃**に使われる可能性。一つのAIが**100万体に自己複製**し、それぞれが次世代AIを開発（2026.8.31 毎日）

対策 どう立ち向かうか — 動き始めた国際社会

- **開発凍結・国際協定**：核兵器のように協定で開発をクールダウン、「AIが人間の知能を超えないように」との議論
- **米中対話**：国家安全保障に関わる事態の相互通知の仕組みを米国が提案（2026.9.20）
- **米「AIキルスイッチ法案」**（超党派、2026.7末）：強制停止機能の搭載義務と政府の緊急停止命令権。10人以上の死者・1億ドル以上の損失等で介入、違反は1日最大2,000万ドルの罰金

- **IAEA型の国際第三者機関**：透明性・正当性の確保が鍵。**日本のようなミドルパワーこそ主導できる**（スタンフォード大・ウェブスター研究員、2026.8.26 朝日）
- **「AI2040」**（元オープンAI研究員ココライロ氏）：プランD＝米中競争の末、AIが人類を邪魔とみなし生物兵器で絶滅。プランA＝データセンターを互いの隣国に置き、合意破りには破壊する「**相互確証計算破壊**」