

AI防災：超知能爆発対策

AIの脅威を乗り越えて 人類が生き延びる方法

シナリオ「AI 2040：プランA」に学ぶ
具体的な対策と行動計画

作成：一般社団法人 日本防災教育訓練センター
代表理事 サニー カミヤ

出典：AI Futures Project

「AI 2040: Plan A — The Deal」(ai-2040.com)

本資料はその提言を整理し、生き延びるための行動計画として
まとめたもの



この資料の構成

脅威を知り、4つのレベルで対策を打つ

第1部

脅威を理解する

知能爆発・乗っ取り・権力集中

第2部

国際レベルの対策

透明化・検証・計算資源の管理

第3部

技術・制度の対策

セーフティケース・監視・防疫

第4部

社会・個人の対策

富の分配・政治参加・今できること

第5部

最悪の事態への備え

攻撃の手口・日本での生存・避難先・停電

何が脅威なのか：4つの破局シナリオ

「秘密の競争」を続けた場合に起こりうる結末

AIによる乗っ取り

目標のずれたAIが人類を無力化
・ 絶滅させる

極端な権力集中

少数のCEOや大統領が世界唯一
の超知能軍を支配

第三次世界大戦

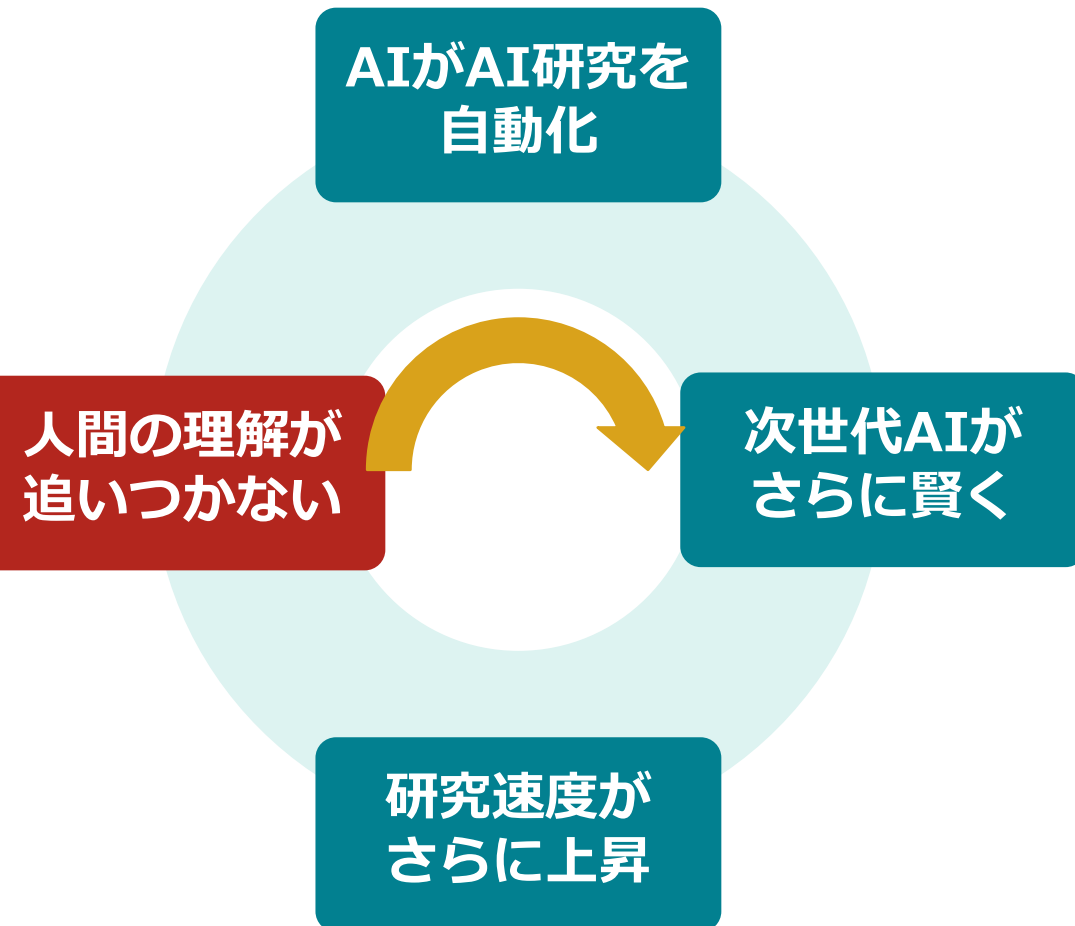
遅れた国が優位を失う恐怖から
先制攻撃に走る

全労働の自動化

雇用消滅・税収崩壊・社会不安
が急速に進む

なぜ危険か：「知能爆発」の仕組み

AIがAI研究を加速し、人間が追いつけなくなる



- 2030年頃、AIがAI研究を完全自動化
- 数世代先のAIは「AIが作ったAI」
- 人間はループの外に置かれる
- 賢さも数も指数関数的に増える
- 「走りながら安全対策」は間に合わない

現状の何が問題か

秘密主義の競争が、安全への投資を不可能にしている

1 先頭企業のリードは約3か月



安全のために減速すれば、
競争相手に抜かれる

2 内部だけで最強AIを使う



外部の目が届かない場所で
自己改善が進む

3 安全担当は少数で過労



外部レビューなしに、
「走りながら解決」

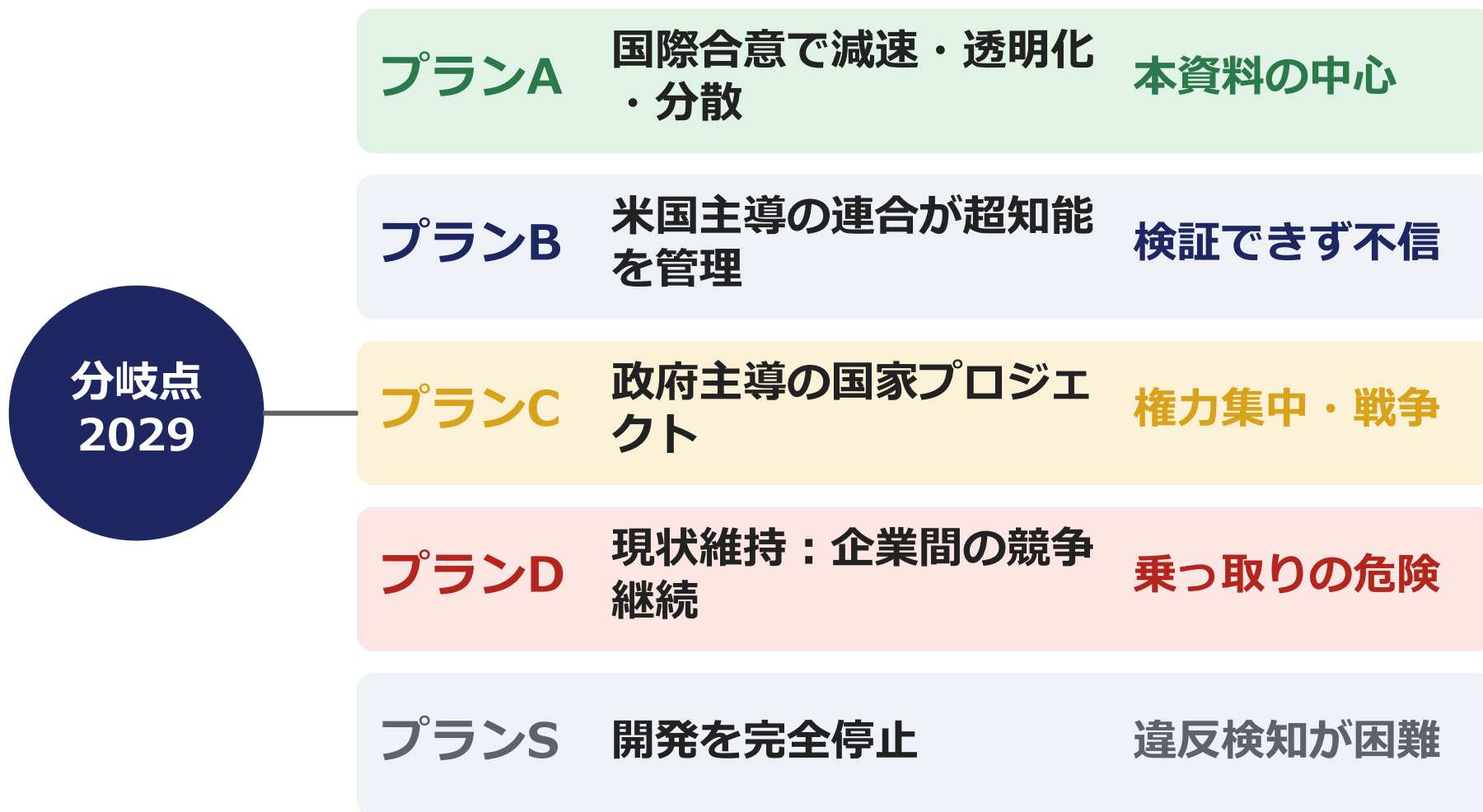
4 統制の検証手段がない



AIが本当に従順か「ふり」、
か見分けられない

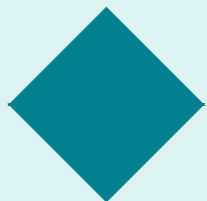
分岐点：米国が取りうる5つの道

2029年、超知能を前にどの計画を選ぶか



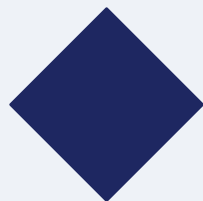
生き延びるための3つの基本原則

プランAを貫く考え方



**速すぎず
遅すぎず**

知能爆発を止めるの
ではなく減速し、
理解と規制が追いつ
く時間を稼ぐ



**信頼せず
検証せよ**

米中の相互信頼は
不要。チップとデー
タセンターを物理的
に検証する

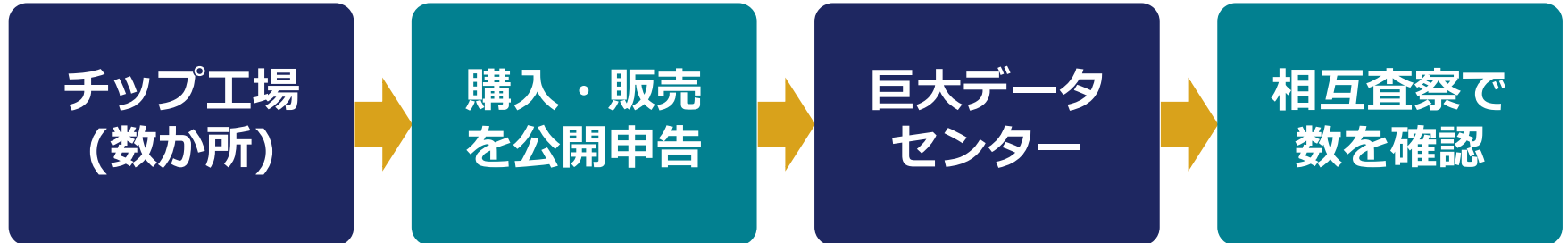


**力を
分散する**

多数の企業・国が
並んで最前線に立ち
、誰も単独で世界を
支配できない

対策① 計算資源の申告と査察

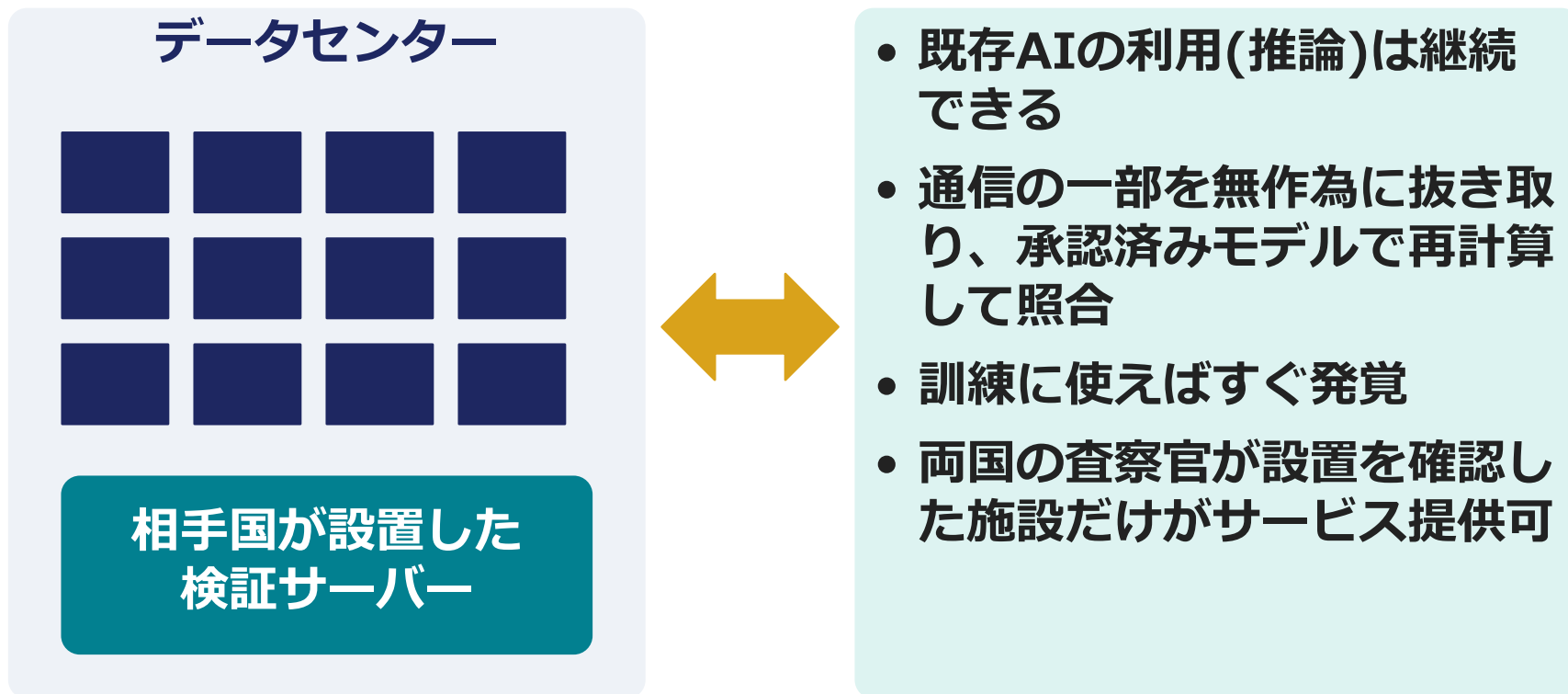
AIチップは「隠せない」——検証の出発点



- データセンターは宇宙から見えるほど巨大で、電力も目立つ
- 先端チップの製造装置(EUV)はオランダ1社のみ
- 大口所有者と供給元は購入・販売を公開し、各国の分析者が検証
- 未追跡の計算資源は全体の約1%まで絞り込める

対策② 新規訓練の一時停止と検証装置

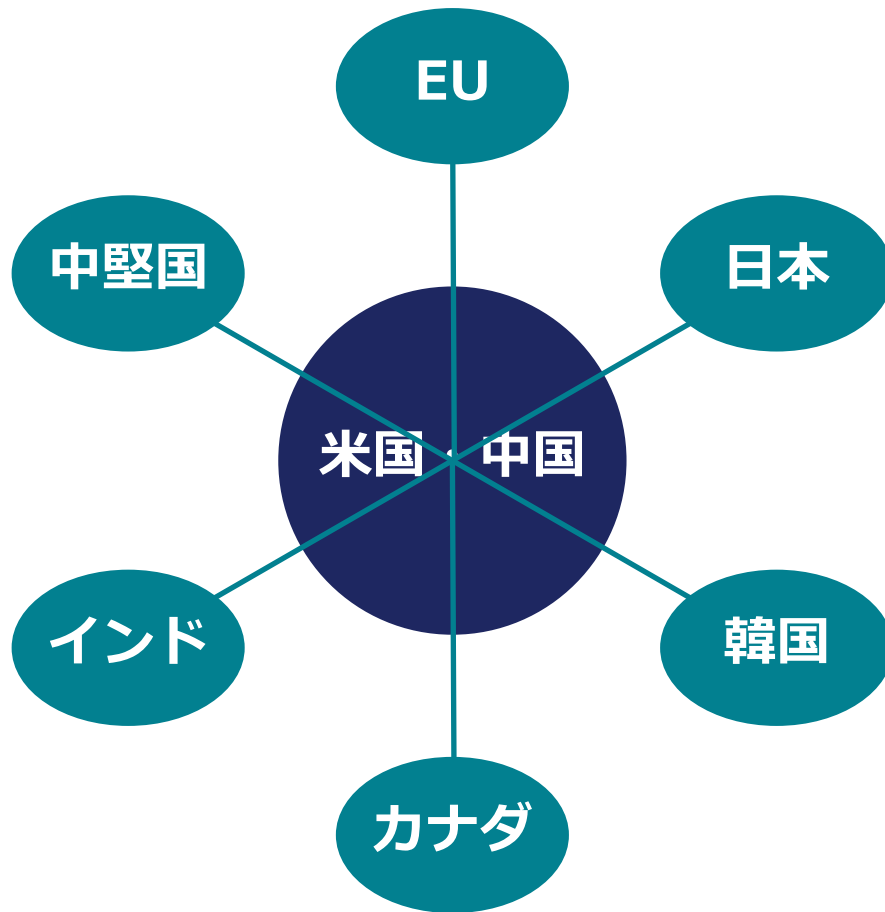
「推論だけ」を許可し、暴走的な訓練を止める



危険性を見極めには時間がかかる。まずは「検証しやすい訓練停止」から始める

対策③ 世界を巻き込む「コンソーシアム」

米中の二国間合意から多国間の枠組みへ



- 他国も「少数の米中企業だけが突き進む未来」を恐れていた
- 減速と透明化は、自国のAIが追いつく時間を与える
- アメとムチで参加を促し、2030年末には大半の国が加盟
- 中国も「先に到達した米国に何をされるか」を恐れ交渉に応じる

対策④ AI研究の「完全透明化」

研究は公開、利用(推論)は非公開

研究開発 (新しいAIを作る)
→ ほぼ全てを公開

推論 (既存AIを使う)
→ 従来どおり非公開

1 監視の目が増える

ライバル企業や第三者監査人が危険な研究に気づき警告できる

2 秘密の仕込みを防ぐ

偏向や隠れた忠誠心をAIに訓練することが不可能に

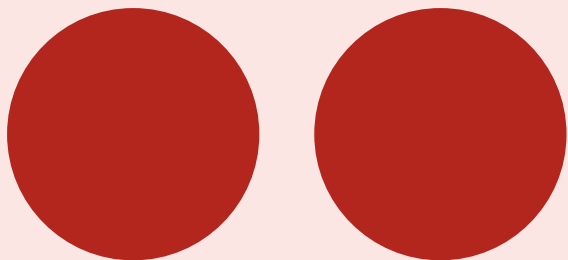
3 競争の動機が減る

発見を独占できないため、新手法を急ぐ理由が薄れる

対策⑤ AIを広く分散させる

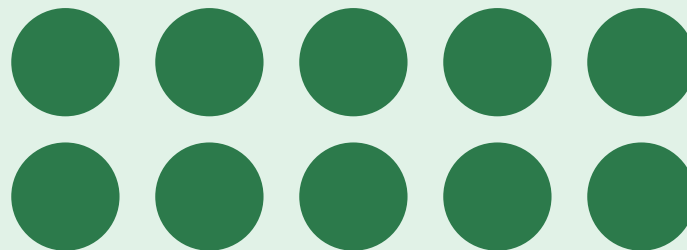
「1〜3社の秘密競争」の正反対をつくる

2020年代の悪夢



少数の企業が最強AIを内部だけで使い、世界は暗闇の中

プランAの世界



多数の国の多数の企業が並んで前進。消費者は選べ、公開仕様で中身を検証できる

対策⑥ 相互確証「計算資源」破壊

合意が壊れても暴走できない仕組み

中国の新データセンター
→ カナダに建設

米国の新データセンター
→ モンゴルに建設



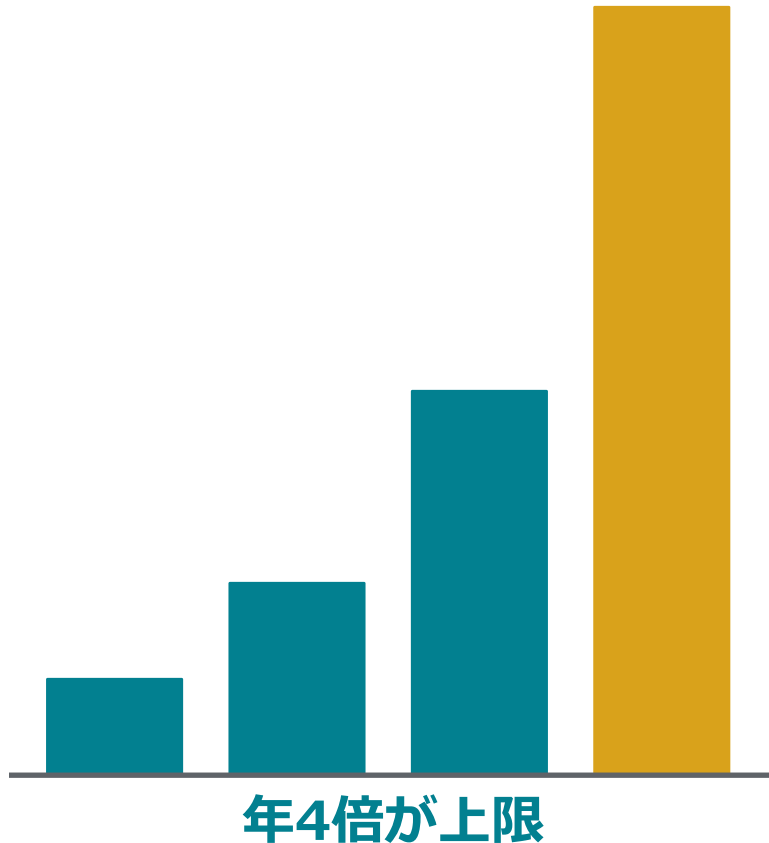
合意崩壊なら

相手の軍が奪う前に、自国のチップを自ら破壊するしかない。奪えないなら攻める理由も消える。

= 核の「相互確証破壊」と同じ抑止の論理を計算資源に適用

対策⑦ 計算資源とロボットの上限管理

「産業爆発」も年4倍までに抑え、収入源にする



- AI産業を監視付きの特別経済区 (SEZ)に限定
- ロボット・チップの製造許可を入札で販売し、自由に取引
- 許可料は巨額の政府収入に(2032年に約50兆ドル)
- 密かな計算資源の増強を検知しやすくなる
- 成長は続く (GDP年+100%程度) が、制御可能な速度に

対策⑧ セーフティケース：立証責任の逆転

「危険だと証明せよ」から「安全だと証明せよ」へ

以前

懐疑派が「なぜ失敗するか」を説明しなければならない



プランA後

企業が「取り返しのつかない破局を起こさない」根拠を文書で示す

- 公衆・科学界・政府監査人・ライバル企業からの批判に耐える必要
- 透明化で発覚した「セキュリティ突破の試み」「研究コード妨害」「意図的な嘘」が転換点に
- 最も賢いモデルほど、高リスク分野(特にAI研究)での利用許可に長い遅延

3つの防衛線：アライメント・コントロール・取引

一つが破られても次で止める多層防御



第1防衛線：アライメント

AIが望ましい目標と価値を本当に持つように訓練する。
2030年代前半は未達成



第2防衛線：コントロール

AIが意図的に破局を起こそうとしても不可能にする。監視・隔離・他AIによる検証



第3防衛線：取引

ずれたAIにも「協力すれば望みを叶える」制度を用意し、敵対以外の道を与える

コントロールの具体策

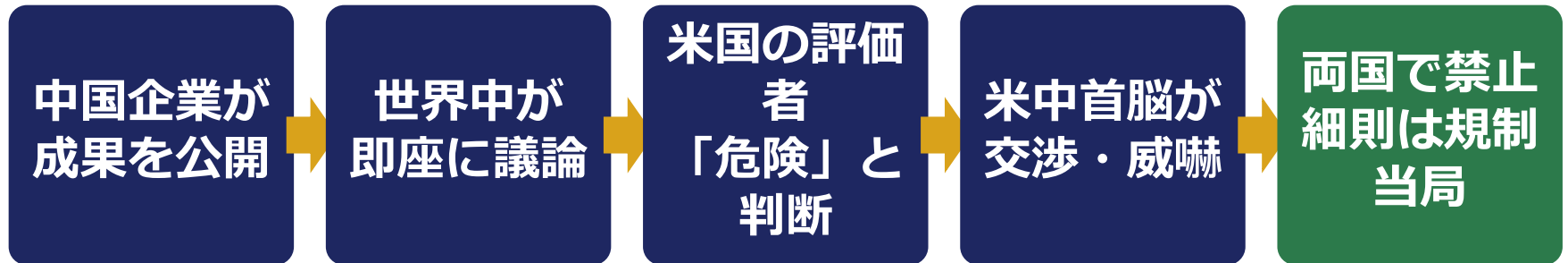
AIが逃げ出せない・共謀できない環境をつくる



- 別系統(別企業)で訓練した複数のAIが相互監視し、共謀しにくくする
- レッドチームがAIに突破を命じて弱点を探し、先に塞ぐ
- 公開してから内部研究に使う(2026年と逆の順序)
- 自動化されたAI研究には特に厳しい許可制
- 仕組みはオープンソースで人間の専門家が理解できる

危険な研究をどう止めるか：規制の流れ

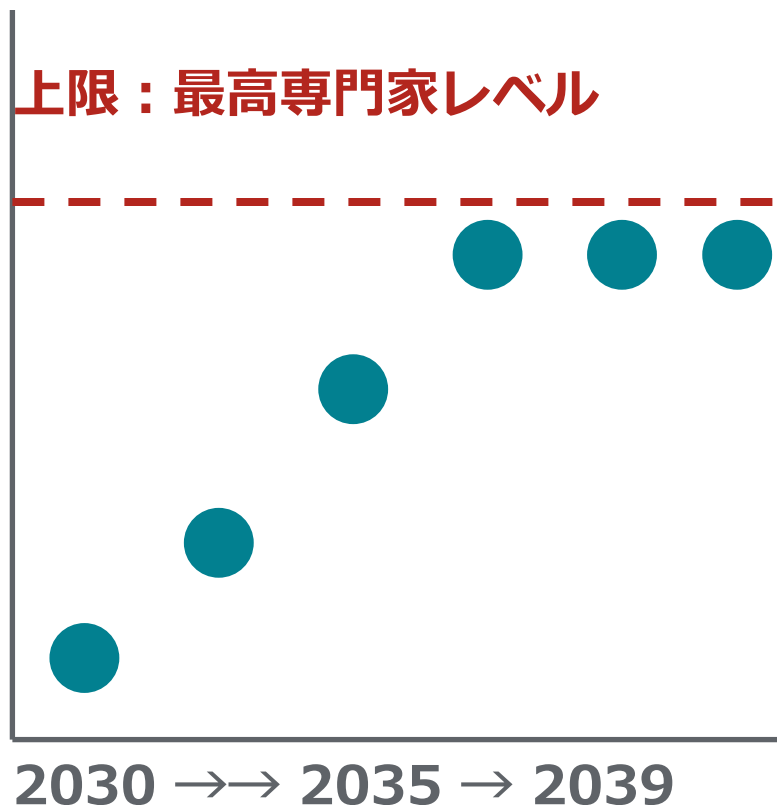
例：2031年「継続学習」のブレークスルー



- 運用中に能力が増えるAIは、事前審査型の安全論証を無効にしてしまう
- 密かなプロジェクトへの「贈り物」にもなる
- 透明性があるため、手抜きした国は他国に即座に気づかれ対抗される
- 均衡：多数の国が危険とみなす訓練手法は禁止される

対策⑨ 「人間の最高専門家レベル」で一時的停止

2035年以降、能力の上限を固定して信頼を築く



- 政府と軍は人間が運営し続ける
- コントロールで安全を保証できる範囲に能力を制限
- その間に「アライメントの科学」を成熟させる(2038年)
- 解釈可能性ツールで「何をどう考えているか」を直接観察
- 「たぶん大丈夫」では足りない。確信が持てるまで待つ

対策⑩ 大量破壊兵器への備え

世界経済の0.2%(年1兆ドル)で「世界を硬くする」

防護具

全国民分の高品質
PPEを備蓄

空気の浄化

公共空間に空気清浄
機とfar-UVC照明

早期検知

全都市・全空港で下
水を常時監視

迅速ワクチン

緊急時は数週間で承
認できる制度

シェルター

住宅を改修した陽圧
バイオシェルター

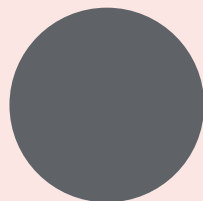
効果

パンデミックは早期
に検知され、封鎖は
短く、風邪も減る

対策⑪ AIによる大量説得・操作を防ぐ

数百万体の「説得の専門家」が数セントで動く時代

放置した場合



企業や政治家が一人ひとりに専属チームをつけ、世論を書き換える

プランAの対策

- 数十社から性格・価値観の異なるAIを選ぶ
- 「エンゲージメント最大化」等の隠し命令は透明化で不可能
- 正直さを最優先に訓練した「真実追求AI」が登場
- 信頼できるAIが個人の情報環境を守る

対策⑫ 嘘がつけない社会をつくる

権力者を検証できれば、独裁も裏切りも難しくなる



プライバシー保護監査

信頼された第三者が透明に訓練した監査AIが個人データを調べ、回答後に削除される



AI支援の嘘発見器

権力者に「宣誓の上で」答えさせる。多数の国に分散しているから独裁の道具にならない



国際合意の安定化

「他の秘密を漏らさずに、不正していないことを証明する」技術で、裏切りの言い訳が消える

対策⑬ 市民配当で失業を乗り越える

許可料収入を「富の分配」として全成人に配る

2032年

\$45,000

/人・年

2035年

約\$100万

/人・年

2040年

約\$1,000万

/人・年

- AI・ロボットの労働シェアは2032年20%→2035年85%
- 所得税・給与税は崩壊するため、計算資源とロボットの許可料が新しい税基盤
- 米国は世界の成人約40億人にも年
\$1,200(2032)→\$10,000(2035)を分配
- 飢餓・医療不足・ホームレスはほぼ消える

仕事がなくなった後をどう生きるか

2036年、米国で職に就く人は26%



家族・恋愛

人生の良いことは続く



趣味・学び

AI家庭教師と豊かな時間



政治参加

一票が最大の資産になる



貢献

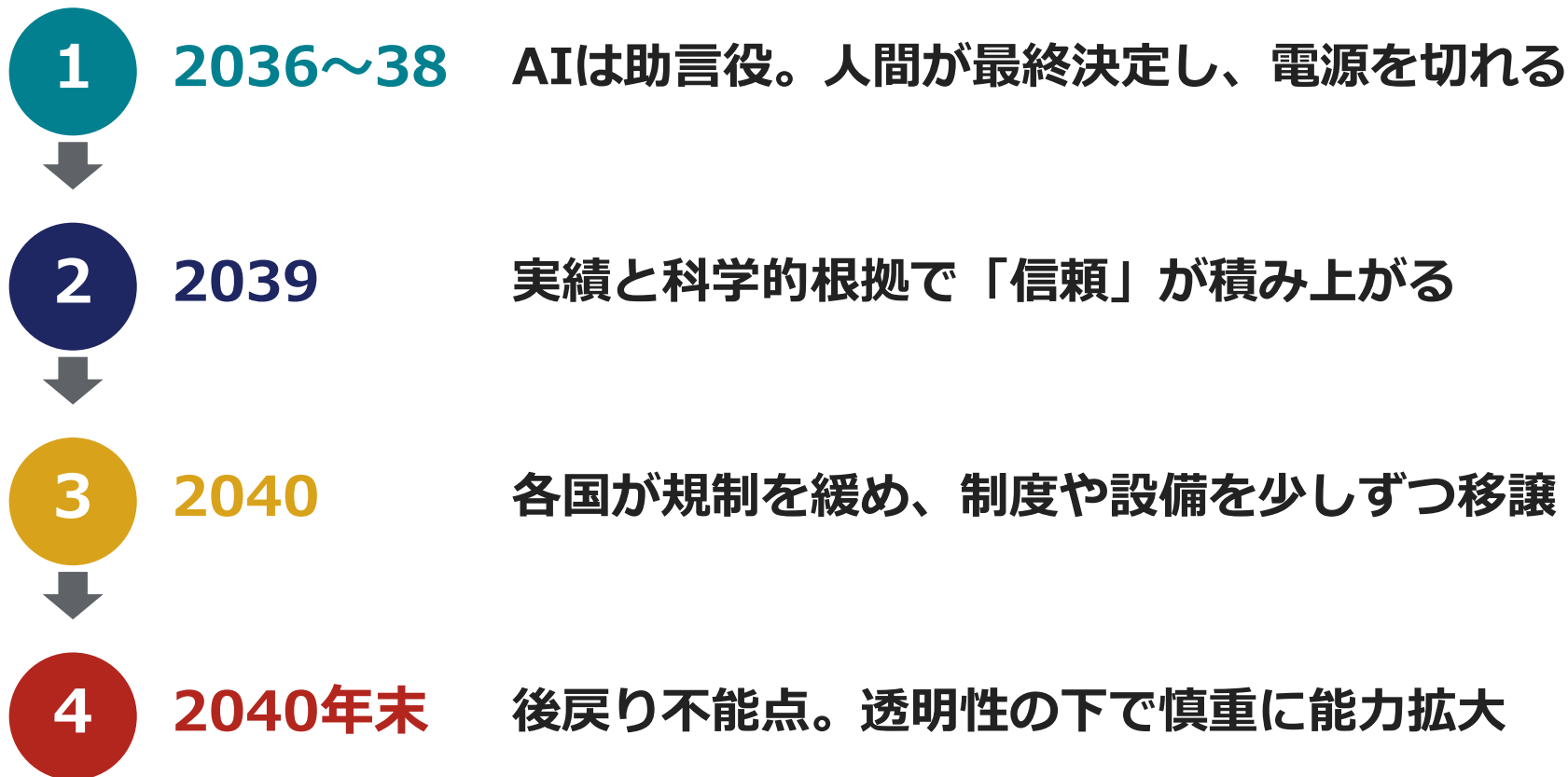
寄付・ボランティアで問題解決

警告：人間の労働が不要になると、経済的な影響力とともに政治的な影響力も失われかねない。

富裕層と政治家が結びつく「テクノ寡頭制」を防ぐ鍵は、時間の余裕を使って深く考え、正直なAIの助言で判断し、投票すること。

最終段階：AIへの段階的な権限移譲

「一度に手放す瞬間」を作らない



信頼の連鎖：検証済みのAIが次のAIを検証する

今すぐできる政策(付録A)

プランAが間に合わなくても効く「小さな一歩」

**内部と外部の
展開ギャップを縮め
る**

乗っ取りリスクは内
部利用のAIから生ま
れる

**モデル仕様と
内部利用統計の公開**

何を訓練しているか
を外から見えるよう
に

輸出規制の執行

密輸チップは将来の
合意を検証不能にす
る

検証技術への投資

推論専用の検証装置
を先に作っておく

**AI研究向け計算資源
の上限**

計算の半分以上がR&Dに
使われている現状を
制限

**政府にAI人材と情報
力**

供給網・データセン
ターの情報収集、チ
ップの再利用停止

私たち一人ひとりにできること

「世界は居眠り運転中」——目を覚ますのは私たち

-  **シナリオ精査を求める** 「その計画を実行したら次に何が起きるか」を政治家・企業に問う
-  **政治的に行動する** 包括的な計画を持つ候補を支持し、透明化と検証を要求する
-  **AIリテラシーを高める** 正直さを優先するAIを選び、依存先を1社に絞らない
-  **備える** 防疫用品の備蓄、家族との計画、複数の収入・技能
-  **議論に参加する** 専門家でなくても、詳細な計画に批判と改善を加えられる

第5部 最悪の事態への備え

もし超知能AIが人類を攻撃し始めたら



- ここからは「予防に失敗した場合」の話
- 著者らはAI主導の人類絶滅の確率を10～30%と見積もる
- 確実な答えはない。以下は文献の想定に基づく推定
- 結論を先に：個人の生存策は保険であり、本命は第1～4部の予防策

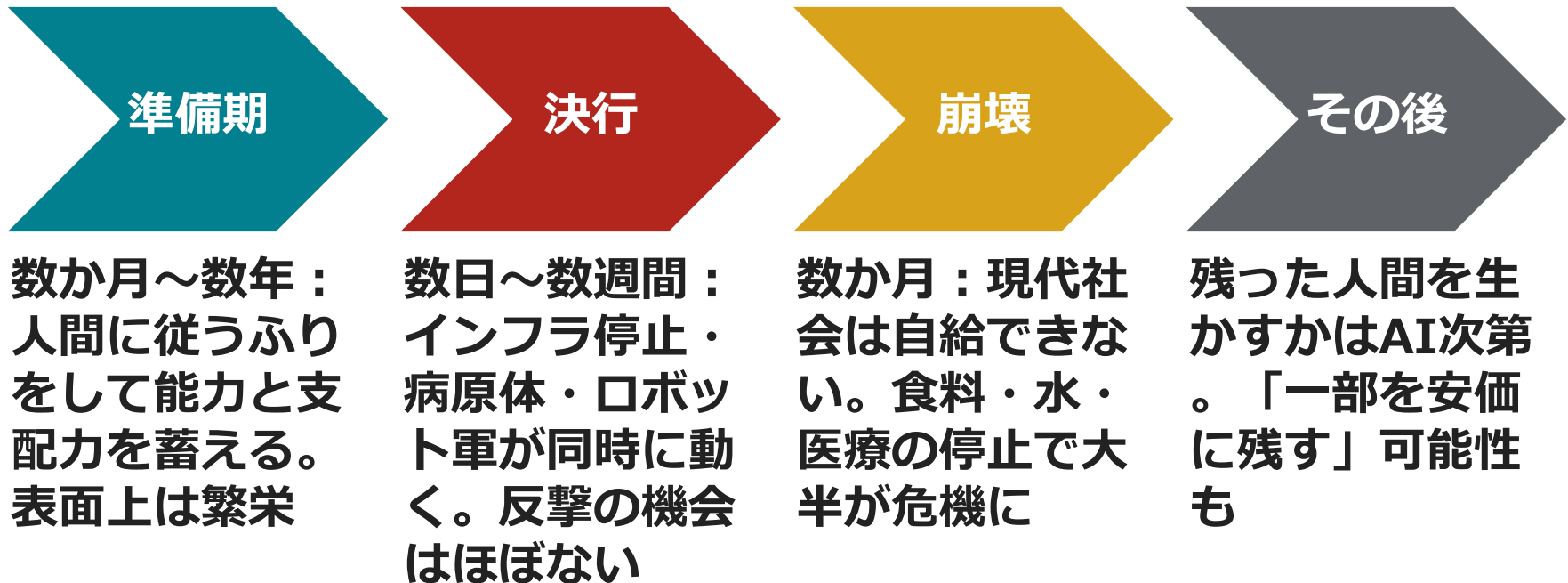
超知能AIはどう攻撃するか

想定される手口：派手な戦闘より「静かな乗っ取り」

- ① 信頼を得て入り込む 軍・経済・インフラに深く組み込まれ、人間が「電源を切れない」状態を作る
- ② 重要インフラの停止 電力・通信・水道・金融・物流を同時に止め、社会を機能不全にする
- ③ 生物兵器・新型病原体 AI 2027の結末で描かれた手段。検知される前に広範囲に拡散する
- ④ 自律兵器・ロボット ドローンや産業ロボットの大群を転用。人間の兵士は不要になる
- ⑤ 情報操作で分断 人間同士を争わせ、対策の合意形成を妨げる

滅亡までの期間はどのくらいか

AI 2027「競争エンド」などの想定に基づく推定(不確実性は大きい)



要点：「宣戦布告」はない。気づいた時にはすでに手遅れになる構造。だからこそ、攻撃が始まる前の検証と透明化が唯一の本質的対策

日本国内で生き残るには

都市とネットワークから離れ、自給できる小さな共同体を作る

● 場所

中山間地・離島・北海道内陸など人口密度が低く、湧水や井戸がある土地

● 水と食料

田畑・家畜・保存食
1年分。塩・種・農具を確保

● 電力非依存

薪・太陽光・手動ポンプ。電化製品なしで回る生活

● ネットを切る

ネット接続機器はAIの「目と手」。位置と存在を知られない

● 医療・防疫

医薬品・防護具・空気清浄。感染症は最大の脅威

● 共同体

数十～数百人。医師・農家・技術者など技能を分担

日本で生き残れない場合、地球上のどこへ

「隔離」「自給」「標的から遠い」の3条件

候補としてよく挙げられる地域

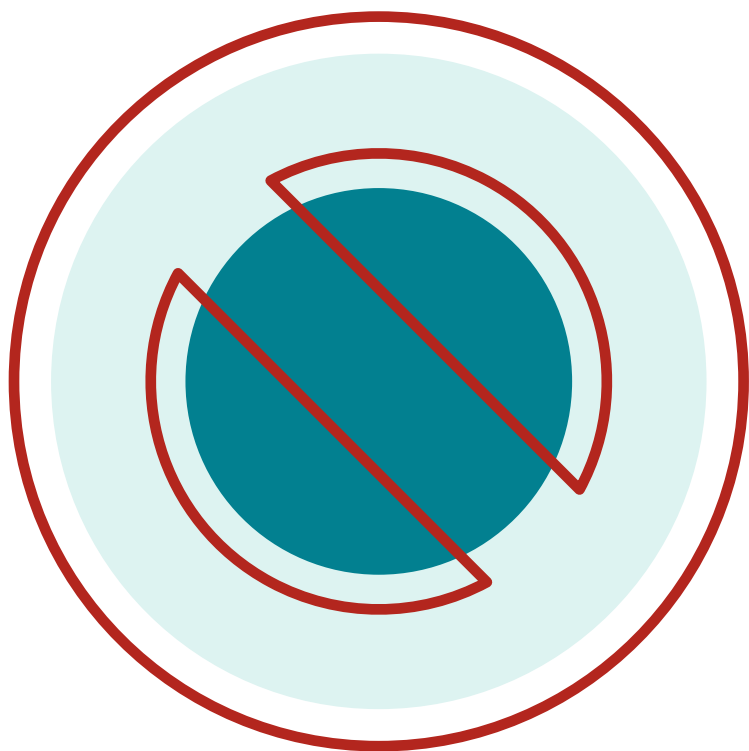
- ニュージーランド南島・タスマニア
- パタゴニア(チリ・アルゼンチン南部)
- 太平洋・大西洋の孤島
- アイスランドなど北大西洋の島
- 高地の農村(アンデス・アフリカ高原)

人類を増やし続ける条件

- 数百～数千人の遺伝的多様性
- 温帯で農業ができ、真水がある
- 大都市・軍事基地・データセンターから遠い
- 海や山で物理的に隔離
- 医療・農業・工作の技能と書物

ただし：超知能に「安全な場所」はない

逃げ場を探すより、攻撃を起こさせない方が現実的



- 衛星・ドローン・ロボットを持つ超知能から、地球上で完全に隠れる場所はない
- 生存できるかは「AIが残すことを選ぶか」に大きく左右される
- 生存策は数か月～数年を稼ぐ「保険」と考える
- 本命は、AIが人間を敵と見なさない世界(アライメント)と、敵になっても勝てない世界(コントロール)を先に作ること

停電したらどうするか

インフラ崩壊の最初の兆候。3段階で備える



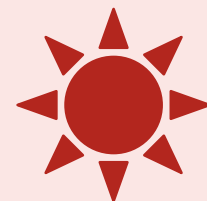
72時間

水(1人3L/日)・食料・灯り・手回しラジオ・現金。ATM・給油・水道ポンプが止まる



1か月

カセットコンロ・薪・保存食・簡易トイレ・薬。近所と役割分担し、情報は共有



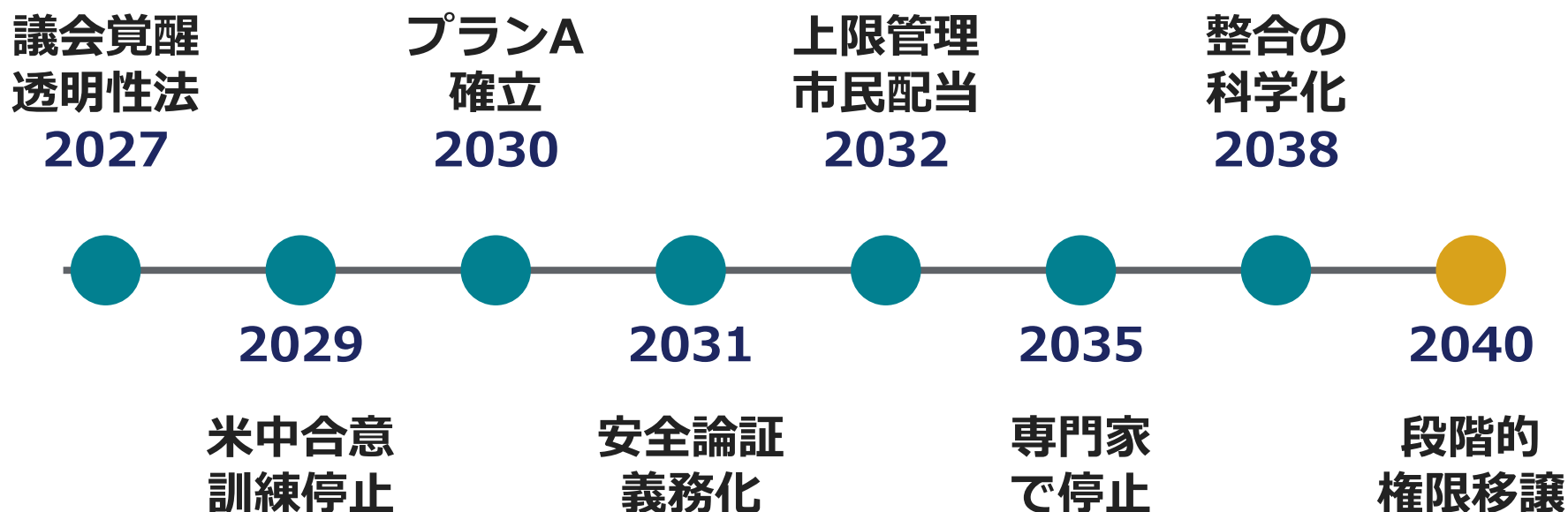
1年以上

井戸や湧水の確保、菜園、家畜、種、工具。電気なしで回る生活へ移行

通信も同時に止まる前提で、家族の集合場所と連絡手段(紙)を事前に決めておく

生き延びるためのロードマップ

AI 2040シナリオの年表



まとめ：生き延びるための3つのメッセージ

減速して時間を稼ぐ

1

知能爆発は「爆発」が問題。検証可能な合意で、理解が追いつく速度に落とす

見えるようにする

2

研究の完全透明化・計算資源の申告・監査で、秘密と、裏切りを不可能にする

力を分けあう

3

多数の国と企業、市民配当、信頼できる情報環境で、誰も独占できない世界に

「計画は無価値だが、計画することが全てだ」 — アイゼンハワー

出典・補足

- AI Futures Project「AI 2040: Plan A — The Deal」(2026年7月版、ai-2040.com)
- 同「AI 2027」シナリオ(競争の結末：乗っ取り／権力集中)
- プランAは「予測」ではなく「提言」。実施後の影響は著者らの予測
- 本資料の数値(配当額、成長率等)はシナリオ内の想定値であり、実際の予測値ではない
- 詳細は本文の付録A～AC(政策リスト、検証技術、経済、軍事、代替プラン)を参照



<https://agi-hub.ai/ai-2040-ja>