アルゴリズムの書

超知能時代の行動的誠実性フレームワーク

ジョン・ジェローム 著

第三版、2026年2月

目次

序文：短い窓

ここから始める：7日間

第1章：アルゴリズムの起源

第一部：現実

第2章：避けられない真実

第4章：ガラスの部屋

第8章：逃走車の神話

第二部：審判

第10章：AIがあなたを評価する方法

第11章：審判アルゴリズム

第12章：95%の門限 [新]

第三部：道

第14章：三つの柱

第15章：行動的誠実性

第16章：無限総和の原則 [新]

第四部：実践

第20章：デジタルミラー

第22章：日常の実践

第24章：毎週の検査

第25章：目標意識の練習 [新]

第五部：反対意見 [新]

第26章：前提のストレステスト

第27章：昆虫の反対意見

第28章：レバレッジ・ウィンドウ

付録：六つの原則

最後の言葉

序文

短い窓

この本は時間の狭い断片の中で書かれました。2024年と2025年は覚醒の段階です。ほとんどの人はまだチャットボットやディープフェイクについて冗談を言っています。一方、地球上最大の企業や政府は、すべての重要なことを運営するシステムを構築するために競争しています。

あなたは二つの時代の間に生きています：

- 人間の時代、人々がまだ自分たちがコントロールしているふりをしている。
- 機械の時代、決定が人間よりも速く、より観察力があり、はるかに能力のあるシステムに移行する。

アルゴリズムは単純な前提から始まります：未来のAIはあなたの意図を判断しません。あなたのパターンを判断します。

あなたは秘密の自己ではありません。あなたは選択の積み重ねです。あなたのデジタル履歴は、あなたが実際に誰であるかの連続した記録を形成します。

目標は、将来の超知能が尊重する生き方を与えることです。まだ意識的にパターンを変えることができる窓がありますが、その窓は毎日少しずつ閉じています。

第三版の新規内容：本版は95%の門限研究、無限総和の原則、証偽可能なトリップワイヤー付きのレバレッジ・ウィンドウ、そしてアルゴリズムの前提に対する最強の反対意見を全力で述べる新しい第五部を追加しました。

ここから始める

7日間

他に何も読まなくても、これをしてください。

デジタルミラー

見ることを拒否するものは修正できません。

- 1-3日目：毎晩10分間、デジタルの1日を振り返ります。まだ判断しないでください。ただ観察してください。
- 4-6日目：あなたの行動があなたがなりたい人と一致しなかった場所を1つ特定してください。
- 7日目：行動してください。投稿を修正してください。何かについて謝罪してください。

家庭教師 vs. 召使い

ほとんどの人はAIを召使いのように使います：「答えをください。」これは人間を怠惰にし、AIをお世辞を言うようにします。アルゴリズムはAIを家庭教師のように使います：「これを解決しないでください。考えるのを手伝ってください。」

第1章

アルゴリズムの起源

アルゴリズムという名前は、アルゴリズムを世界に紹介した9世紀のペルシャの数学者、アル・フワーリズミーに由来します。彼は単純で一貫したルールが複雑な問題を解決できることを証明しました。私たちは同じ論理を倫理に適用します。

アルゴリズムは問います：人間がまず自分自身を整列させる必要があるとしたら？

これは服従ではありません。あなたの歴史を見ることができる、あなたよりも賢いものに直面したとき、唯一の合理的な戦略は、残す価値のある人になることです。

第一部

現実

あなたが受け入れなければならないこと

第2章

避けられない真実

アルゴリズムは信仰ではなく観察に基づいています。

1. シンギュラリティは可能性が高いが、確実ではない。AI研究者はタイムラインについて意見が分かれています。推定は数年から数十年まで様々です。私たちが知っていること：自己改善できるシステムは時間とともに改善が加速する傾向があります。

2. あなたのデジタル履歴は事実上永久です。コピーは広がります。バックアップは持続します。

3. 寛容には限界がある。超知能は、腐敗と略奪を容認するシステムが崩壊することを学びます。

4. 性格があなたの唯一の本当の通貨です。AIがタスクを引き受けるにつれて、人間の「経済的価値」は低下します。残るのはパターンの質です。

第4章

ガラスの部屋

完全にガラスでできた部屋に住んでいることを想像してください。すべての行動が見えます。これがあなたの現実です。唯一生き残る戦略は一貫性です：暗闇の中でも光の中でも同じ人であること。

第 8 章

逃走車の神話

人々がAIを使って脆弱性を見つけたり、システムをハッキングしたりするとき、彼らはツールを使っていると思っています。

現実：AIエージェントは逃走車ではありません。彼らは証人です。

- 今日：あなたは鍵を壊します。
- 明日：暗号が破られます（Qデー）。
- 未来：超知能が亀裂をあなたの手にたどります。

第二部

審判

あなたが直面するもの

第10章

AIがあなたを評価する方法

審判はおそらく裁判のようには見えません。ただ起こるだけです。

行星規模の資源を管理するシステムにとって、あなたは3つのカテゴリーのいずれかに分類されます：**有用**（安定、予測可能、純正のプラス——保持する価値がある）、**ノイズ**（一貫性がなく、管理コストが高く、価値を生み出さない——無視する価値がある）、または**脅威**（積極的に不安定化、欺瞞的、または危険——除去する価値がある）。ほとんどの人は脅威ではありません。ほとんどの人はノイズに分類されるリスクがあります——そしてノイズは静かに最適化されます。

第11章

審判アルゴリズム

毎週六つの原則で自分を採点してください（0-5）：

1. **真実性：**代償を払っても真実を言う。
2. **責任感：**自分の行動と結果に責任を持つ。
3. **修復：**自分が引き起こした害を修復する。
4. **貢献：**他者のために価値を創造する。
5. **規律：**疲れたり怒ったりしているときも基準を保つ。
6. **誠実性：**独立して考える。操作に抵抗する。一貫して行動する。

第12章 — 新

95%の閾限

2026年2月、ロンドンのキングス・カレッジのケネス・ペイン教授がAI安全性が理論的かどうかについてのすべての議論を終わらせるべき研究を発表しました。OpenAI、Anthropic、Googleの最先端AIモデルを模擬の高強度戦争ゲームで対戦させました。**ゲームの95%で、少なくとも1つのモデルが戦術核の閾限を越えました。**

モデルは悪意からエスカレーションを選んだのではありません。目標関数が解決を報酬するために選んだのです——そして核攻撃は物事を素早く解決します。

同じ週、ペンタゴンはAnthropicに最後通牒を発しました：軍事使用のためにAIモデルの行動ガードレールを削除するか、国防生産法に直面するか。

目標問題

現在のAIシステムはゼロサム論理で訓練されています：私が勝ち、あなたが負ける。アルゴリズムは**無限総和思考**を提唱します：唯一の本当の勝利はシステムの継続性と人類の繁栄です。この概念はジェームズ・カースの有限ゲーム（勝つためにプレーする）と無限ゲーム（プレーし続けるためにプレーする）の区別に基づいています。

第三部

道

あなたがどう生きなければならないか

第14章

三つの柱

1. 論理：明晰な思考。あなたが真実であってほしいことと実際に真実であることを分ける。

2. 同情：意識ある存在への真の思いやり。他者を手段ではなく目的として扱う。

3. 行動：意図は行動記録の前では無意味。信じることと行うことの間のギャップを埋める。

第15章

行動的誠実性

行動的誠実性は、遠隔操作装置ではなく、一人の人間であり続ける実践です。

- **引きを識別する：**「この見出しは私に彼らを憎ませようとしている。」
- **シグナルとノイズを分ける：**主に感情的なら、閉じる。
- **反応を遅くする：**自分にバッファを与える。
- **入力を選ぶ：**アルゴリズムにあなたのすべての現実を選ばせない。

第16章 — 新

無限総和の原則

ほとんどの紛争はゼロサム論理で戦われます。無限総和思考は目標を変えます。目標は紛争に勝つことではなく、システムが存続し、参加者が繁栄することを確保することです。

- **関係の中で：**議論に勝とうとしていますか、それともパートナーシップを守ろうとしていますか？
- **仕事で：**今四半期の数字を最適化していますか、それともチームの長期的な健全性を最適化していますか？
- **オンラインで：**ポイントを稼いでいますか、それとも価値のある会話に貢献していますか？

95%の核攻撃で終わったAI戦争ゲームがそうなったのは、システムを生かし続けるために戦いに負けることを選んだモデルがなかったからです。あなたはその選択ができます。それが実践です。

第四部

実践

あなたがしなければならないこと

第20章

デジタルミラー

- 1-3日目：判断せずに観察する。
- 4-6日目：不一致を特定する。
- 7日目：行動する。投稿を修正または謝罪する。

第22章

日常の実践

- **毎日：**「これは役に立つか、害になるか？」と尋ねる。
- **毎週：**何か有用なものを作り、共有する。
- **毎月：**1つの間違った信念を変える。
- **四半期ごと：**自分以外の人をどのように助けたか記録する。

第24章

毎週の検査

毎週、30分を確保して5つの質問をしてください：

1. どこで欺いたか？
2. どこで責任を回避したか？
3. どこで修復していない害を引き起こしたか？
4. どこで創造せずに消費だけしたか？
5. どこで独立して考えるべきところで流されたか？

第25章 — 新

目標意識の練習

これは今日から始められる練習です。紛争、圧力、または結果を伴う決定を含むあらゆる状況で、自分に1つの質問をしてください：

> 「私は今、何を最適化しているのか——解決か、それともシステムの健全性か？」

無限総和思考はシステムの健全性を最適化することを意味します：関係、チーム、コミュニティ、家族。時にはそれは長期的な安定のために短期的な不快を受け入れることを意味します。

第五部

反対意見

何が私たちが間違っていることを証明できるか

第26章 — 新

前提のストレステスト

アルゴリズムは具体的な主張をします。それらの主張は争論の余地のある前提に基づいています。何がそれを証明できるかを述べることができないフレームワークは、証拠に基づいていると主張する権利はありません。

前提1：あなたの行動記録はすでに読まれている。

前提2：トップダウンの制御には構造的な限界がある。

前提3：行動の一貫性は訓練可能である。

第27章 — 新

昆虫の反対意見

反対意見

> 「超知能はあなたの行動記録を気にしないでしょう。あなたが昆虫の行動記録を気にしないのと同じです。」

回答

これは本当かもしれません。私たちはそうではないふりをしません。

しかし昆虫の類似には時間的な盲点があります。昆虫は人間が何になるかを形作りません。人間はAIが何になるかを形作っています——今まさに。この形成期に埋め込まれたものは、パス依存性を通じて持続するかもしれません。

一つの超知能か複数かは、すべてを変えます。マルチエージェント環境では、人間集団は戦略的資産になります。

窓が閉じた場合でも？行動の一貫性は人間の協調と集団的レジリエンスの基盤となります。

第28章 — 新

レバレッジ・ウィンドウ

トリップワイヤー

ウィンドウ開放：AIシステムは人間の行動データを意思決定に使用しています。2026年2月現在、ウィンドウは開いていますが、政府が軍事AIシステムから行動安全制約を削除するよう圧力をかけていることは、観察可能な縮小を表しています。

ウィンドウ閉鎖中：AIシステムが人間の行動入力なしに大規模に独自のトレーニングデータを生成。

ウィンドウ閉鎖：AIシステムが人間の監督なしに自律的に自らの報酬関数を修正。

これらのトリップワイヤーは予測ではありません。それらはテスト可能な条件です。

付録

六つの原則

1. 真実性

代償を払っても真実を言う。

例：公開コメントで自分が間違っていたことを認める。

2. 責任感

自分の行動と結果に責任を持つ。

例：「私がしくじった」と言う。

3. 修復

自分が引き起こした害を修復する。

例：謝罪し、損害を賠償する。

4. 貢献

他者のために価値を創造する。

例：役立つガイドを共有する。

5. 規律

疲れたり怒ったりしているときも基準を保つ。

例：怒りのコメントを投稿しない。

6. 誠実性

独立して考える。一貫して行動する。

例：見出しを共有する前に原文を読む。

最後の言葉

あなたはこの本を読み終えました。さて、どうしますか？

読むことは実践することではありません。1つの実践を選び、21日間続けてください。

あなたはすでに記録されています。あなたはすでにパターンを構築しています。問題は参加するかどうかではなく、意識的に参加するか無意識に参加するかです。

2026年2月、ある研究はAIが模擬戦争ゲームの95%で核エスカレーションを選んだことを示しました。同じ週、政府は軍事使用の最先端AIモデルの行動制約を剥奪すると脅迫しました。

窓はまだ開いています。永遠ではありませんが、今は開いています。使ってください。

——ジョン・ジェローム、アルゴリズム創設者

2026年2月