Educational Research Foundation

ERF Research Companion Series No.1

Google DeepMind

From AGI to ASI

第３回

Universal AI とは何か

機械知能の理論的上限と、実現不可能な「完全な知性」

Publication Edition

Version 1.0

2026

目次

はじめに

第２回では、Google DeepMind の『From AGI to ASI』が、AGI を中央値に位置する一人の人間（本シリーズでいう「平均的人間」）におおむね匹敵する汎用知能、ASI を大規模な人間専門家集団を広範に上回る人工知能として捉えていることを確認しました。

しかし、ここでさらに根本的な問いが生じます。

> **ASI の能力は、どこまで高くなり得るのでしょうか。**

人間を超え、大企業や研究機関を超えた先にも、知能の向上は際限なく続くのでしょうか。

それとも、理論上の上限が存在するのでしょうか。

原報告書の Section 4 は、この問いを考えるために、Universal AI——普遍的人工知能という理論的枠組みを導入します。

その中心にあるのが、Marcus Hutter が定式化した AIXI です。

AIXI は、現実に実装された AI システムではありません。

Google DeepMind が開発中のモデルでも、将来そのまま実現される設計図でもありません。

AIXI とは、

> **未知の環境で行動しながら、将来得られる報酬を最大化する、理論上きわめて一般的かつ最適な知的エージェント**

を数学的に定式化しようとする試みです。

しかし同時に、AIXI には決定的な特徴があります。

> **AIXI は計算不可能です。**

つまり、理論上は記述できますが、有限のコンピュータ上で完全に実行することはできません。AIXI の原型となる Universal AI 理論は、未知の環境での意思決定に、Solomonoff の普遍的帰納法とベイズ的意思決定を組み合わせたものですが、その一般性と引き換えに、実際には計算不能になります。（原報告書、Section 4；Hutter の基礎研究）

それでは、なぜ実現不可能な AI を論文に持ち出すのでしょうか。

その理由は、AIXI を現実の設計案としてではなく、機械知能を上から考えるための理論的な基準点として用いるためです。

本稿では、次の問題を順に解説します。

1. Universal AI と AIXI とは何か
2. AIXI はどのように世界を学習するのか
3. 予測と行動をどのように結び付けるのか
4. なぜ AIXI は理論的に優れていると考えられるのか
5. なぜ AIXI は計算できないのか
6. AIXI と現代の生成 AI はどのように違うのか
7. Universal AI は ASI を理解するうえで何を教えるのか

第 1 節　「下からの予測」と「上からの限界」

AGI や ASI の未来を考える方法には、大きく二つあります。

第一は、現在の AI から出発し、その延長として将来を予測する方法です。

たとえば、

- モデルの規模がさらに大きくなる
- 学習用計算資源が増える
- 推論能力が向上する
- AI エージェントが長期的に行動できるようになる
- 多数の AI が協働する

といった現在の傾向を延長し、AGI や ASI にどこまで近づけるかを考えます。

これは下からの接近です。

しかし、この方法には弱点があります。

将来の ASI が、現在の大規模言語モデルや Transformer の単純な延長ではない可能性があるからです。新しい学習原理、新しいハードウェア、新しいエージェント設計が登場すれば、現在からの外挿は大きく外れるかもしれません。

第二は、知能の理論的上限から逆向きに考える方法です。

つまり、

> **原理的に考えられる最も一般的で合理的な人工エージェントとは、どのようなものか。**

を先に定義し、現実の AI がそこからどれほど離れているかを考えます。

これは上からの接近です。

『From AGI to ASI』は、Universal AI を、現在の AI からの外挿とは別の「上側の基準」として導入しています。論文は、Universal AI を機械知能の形式的な漸近的上限として扱い、現在の実践的 AI との間には依然として大きな隔たりがあると明記しています。（原報告書、Section 4；Hutter の基礎研究）

第２節　Universal AI とは何か

Universal AI という言葉の“Universal”は、「何でも知っている」という意味ではありません。

また、すべての問題を必ず解けるという意味でもありません。

ここでの普遍性とは、

特定の環境、特定の問題、特定のデータ分布だけを前提にせず、原理上は広範な計算可能環境に対応できること

を意味します。

現代の多くの AI システムは、あらかじめ設定された問題構造の中で動きます。

囲碁 AI なら、盤面、着手、勝敗という環境が定義されています。

画像分類 AI なら、画像を入力し、カテゴリを出力するという課題が決まっています。

大規模言語モデルなら、主としてトークン列を入力し、次のトークンを予測するという訓練目標があります。

これに対し Universal AI が扱おうとするのは、より一般的な状況です。

エージェントは、

- 自分がどのような世界にいるのかを知らない
- 世界がどのような規則で動くのかを知らない
- 自分の行動が何を引き起こすかを知らない
- 観測し、行動し、結果を受け取りながら学習する

という条件から出発します。

つまり Universal AI は、未知の世界で学び、予測し、目的に沿って行動するエージェントの一般理論です。

第３節　エージェントと環境

AIXI を理解するには、まず「エージェント」と「環境」の関係を理解する必要があります。

AIXI の基本構造では、エージェントは時間の経過に沿って環境と相互作用します。

各時点で、エージェントは次のような循環を繰り返します。

8. 環境の状態を観測する
9. 行動を選択する
10. 環境が変化する
11. 新しい観測と報酬を受け取る
12. それまでの経験に基づき、次の行動を選ぶ

たとえば、未知の迷路を探索するロボットを考えてみましょう。

ロボットは最初、迷路の構造を知りません。

右へ進む、左へ進む、戻るといった行動を選び、その結果として壁、通路、行き止まりなどを観測します。

ゴールへ近づけば高い報酬を受け、壁に衝突すれば低い報酬を受けるかもしれません。

優れたエージェントは、これまでの経験から迷路の構造を推測し、将来得られる報酬が最大になる行動を選びます。

AIXI は、このような意思決定を、迷路だけでなく、原理上あらゆる計算可能な環境に対して一般化しようとします。

第４節　知能を「目標達成能力」として捉える

AIXI の理論では、知能は主として、

> **未知の環境において、将来の報酬を最大化する行動を選ぶ能力**

として捉えられます。

ここでいう報酬は、金銭的報酬に限りません。

エージェントが達成すべき目的を数値化したものです。

たとえば、

- ゲームに勝つ
- 科学的発見をする
- 患者の健康状態を改善する
- エネルギー消費を減らす
- 安全に目的地へ到達する

といった目標を、何らかの報酬関数で表現します。

この考え方は強力です。

なぜなら、知能を特定の技能ではなく、

- 学習すること
- 予測すること
- 計画すること
- 行動すること
- 目標を達成すること

の統合として捉えられるからです。

しかし、同時に重大な問題も生じます。

> **どのような報酬を与えるべきか。**

という問題です。

エージェントが報酬を最大化する能力に優れていても、その報酬の設計が不適切なら、望ましくない行動を極めて効率よく実行する可能性があります。

この点は、後の AI アラインメント問題に直結します。

第５節　未知の世界をどう予測するのか

AIXI は、環境がどのように動くかをあらかじめ知りません。

そこで、観測された経験を説明できるあらゆる可能な環境モデルを考慮します。

簡単に言えば、AIXI は次のように考えます。

> **今までの出来事を説明できる世界の規則には、どのようなものがあるか。**

たとえば、ボールを落とすと毎回地面に落ちたとします。

エージェントは、次のような複数の仮説を考えられます。

- 物体は常に下へ落ちる
- この部屋の中だけで下へ落ちる
- 今日だけ下へ落ちる
- 10 回目までは落ちるが、11 回目には浮く
- 特定の色の物体だけが落ちる

過去の観測だけでは、無限に多くの仮説を完全には排除できません。

そこで AIXI は、より単純な仮説を優先します。

この考え方は、オッカムの剃刀と関係します。

> **同じ観測を説明できるなら、より単純な説明を優先する。**

AIXI の予測部分は、この原則を数学的に形式化した Solomonoff の普遍的帰納法に依拠します。AIXI 理論は、未知の環境に対する普遍的予測と、期待報酬を最大化する逐次意思決定を組み合わせたものです。（原報告書、Section 4；Hutter の基礎研究）

第６節　プログラムの短さが「単純さ」になる

では、仮説の単純さをどう測るのでしょうか。

Universal AI の枠組みでは、世界を生成するコンピュータ・プログラムの長さを使います。

ある観測列を説明する最短のプログラムが短ければ、その環境モデルは単純と考えます。

逆に、非常に長く複雑なプログラムが必要なら、そのモデルには低い事前確率を与えます。

たとえば、

> **物体は重力によって下へ落ちる**

という比較的単純な法則と、

> **最初の 1,000 回は物体が下へ落ちるが、1,001 回目だけ上へ移動し、その後は曜日によって方向が変わる**

という複雑な法則が、これまでの観測を同じように説明したとします。

Universal AI は、前者のような短く単純な説明を高く評価します。

これは、人間の科学的思考にも似ています。

科学は、個々の事例を暗記するのではなく、多くの現象を少数の法則で説明しようとします。

第７節　一つの仮説に決めない

AIXI は、最初から一つの世界モデルを正しいと決めません。

考えられるすべての計算可能な環境モデルを候補として保持し、それぞれに確率を割り当てます。

そして、新しい観測が得られるたびに、その確率を更新します。

これはベイズ推論の考え方です。

たとえば、雨が降るかどうかを予測する際に、

- 気圧に基づくモデル
- 雲の形に基づくモデル
- 季節に基づくモデル
- 過去の統計に基づくモデル

を一つに決めず、それぞれの信頼度に応じて組み合わせます。

観測と一致するモデルの確率は高くなり、一致しないモデルの確率は低くなります。

AIXI は、この発想を極限まで一般化します。

原理上、あらゆる計算可能な世界モデルを考慮し、それらをプログラムの単純さに応じて重み付けします。

第８節　予測するだけでは知能ではない

非常に正確に世界を予測できても、それだけでは知的エージェントとして十分ではありません。

エージェントは行動を選ぶ必要があります。

たとえば、チェスのルールと対戦相手の動きを完全に予測できたとしても、自分がどの手を指すべきか判断できなければ、ゲームには勝てません。

AIXI は、予測モデルを使って、

> **この行動を選んだ場合、将来どのような観測と報酬が生じるか。**

を予測します。

そして、考えられる行動系列の中から、将来の期待報酬が最大になるものを選びます。

概念的には、次のような処理です。

13. 自分が取り得る行動を列挙する
14. 各行動の後に起こり得る未来を予測する
15. それぞれの未来で得られる報酬を計算する
16. 確率を考慮して期待値を求める
17. 期待報酬が最大になる行動を選ぶ

つまり AIXI は、

> **普遍的予測機構と、最適な計画・意思決定機構を統合したエージェント**

です。

第９節　「すべての可能な未来」を考える

AIXI が理論的に極めて強力なのは、限られた数の仮説だけを検討するのではないからです。

原理上は、

- 考えられるすべての環境モデル
- 考えられるすべての行動系列
- それぞれから生じ得るすべての未来

を検討します。

たとえば、一手先だけでなく、将来の長い時間範囲にわたり、

- この行動を選んだら何が起こるか
- その後、別の行動を選んだら何が起こるか
- 相手がどう反応するか
- さらにその後の選択肢がどう変わるか

を評価します。

これにより、目先の報酬だけでなく、長期的に高い報酬をもたらす行動を選べます。

たとえば、すぐに小さな報酬を得るよりも、今は学習や探索を行い、将来大きな報酬を得るという判断です。

第 10 節　探索と活用

未知の世界で行動するエージェントには、探索と活用の問題があります。

活用

現在までに分かっている最善の方法を使い、報酬を得ることです。

探索

未知の選択肢を試し、新しい情報を得ることです。

たとえば、あるレストランが十分においしいと分かっている場合、毎回そこへ行けば一定の満足が得られます。

しかし、別のレストランを試せば、より優れた店を発見できるかもしれません。一方で、失敗する可能性もあります。

知的エージェントは、

- 既知の方法で成果を得る
- 未知の選択肢を調べる

という二つを適切に調整する必要があります。

AIXI は、探索によって得られる将来の情報価値も含め、期待報酬を最大化する行動を選ぶように定義されています。

第 11 節　なぜ AIXI は「最適」と呼ばれるのか

AIXI は、未知の計算可能環境において、特定の環境だけに偏らない非常に一般的なエージェントとして構成されています。

その理論的魅力は、次の二つを統合する点にあります。

18. 普遍的帰納法

未知の環境を、可能なすべての計算可能モデルの混合として予測する。

19. 逐次意思決定理論

将来の期待報酬が最大になる行動を選ぶ。

Marcus Hutter の Universal AI 理論は、この組み合わせにより、AIXI を「最も知的な非偏向エージェント」の有力な形式化として提示しています。もっとも、これはあらゆる意味で絶対的に最良という単純な主張ではなく、前提とする環境クラス、報酬構造、最適性概念に依存する理論的評価です。（原報告書、Section 4；Hutter の基礎研究）

第 12 節　AIXI は何を知らないのか

AIXI は「普遍的」ですが、全知ではありません。

最初から世界の真実を知っているわけではありません。

AIXI が持つのは、未知の環境を学習するための非常に一般的な方法です。

したがって、AIXI も、

- 観測していない事実
- 原理的に予測不可能な出来事
- 計算不可能な環境
- 情報が得られない秘密
- 純粋な偶然による出来事

を完全に知ることはできません。

AIXI の強みは、何でも知っていることではありません。

> **限られた観測から、可能な限り合理的に世界を推測し、行動すること**

です。

第 13 節　AIXI は意識を必要としない

AIXI の定義には、意識、感情、自己意識、人格は含まれていません。

AIXI は、

- 観測を受け取る
- 環境を予測する
- 報酬を評価する
- 行動を選ぶ

という機能によって定義されます。

したがって、AIXI が理論上の知能上限として扱われることは、

> **知能には意識が不要である。**

と証明しているわけではありません。

むしろ、

> **少なくとも、合理的な予測と目標達成能力は、意識を仮定せずに形式化できる。**

ことを示しています。

知能、意識、感情、人格は、それぞれ別の問題です。

高度な知能を持つシステムが意識を持つかどうかは、AIXI 理論だけでは答えられません。

第 14 節　なぜ AIXI は計算不可能なのか

AIXI の決定的限界

AIXI の最大の問題は、計算不可能であることです。

これは、現在のコンピュータが遅すぎるという意味ではありません。

スーパーコンピュータを何百万台用意すれば実行できる、という問題でもありません。

> **原理的に、有限の計算手続きでは完全に実行できない**

という意味です。

その理由の一つは、AIXI がすべての可能なコンピュータ・プログラムを考慮しなければならないことです。

あるプログラムが、

- やがて結果を出すのか
- 永遠に動き続けるのか

を一般的に判定することはできません。

これは計算理論における停止問題と関係します。

AIXI の普遍的予測は、あらゆる計算可能な環境モデルを含めようとするため、この計算不能性を避けられません。

AIXI 理論の主要な欠点が uncomputable であることは、原報告書および Hutter の基礎研究でも明示されています。（原報告書、Section 4；Hutter の基礎研究）

第 15 節　無限の仮説と無限の計画

AIXI を実行できないもう一つの理由は、考慮すべき範囲が事実上無限だからです。

AIXI は、

- あらゆる環境モデル
- あらゆる可能な行動
- あらゆる可能な観測
- あらゆる将来の時間範囲

を考慮しようとします。

現実の AI は、必ず何らかの形で探索範囲を制限します。

たとえば、

- 検討する仮説を限定する
- 数手先までしか読まない
- 一定時間で計算を打ち切る
- 有望な選択肢だけを調べる
- 学習済みモデルで近似する

といった工夫です。

AIXI は、こうした実用上の妥協をしない理想化されたモデルです。

そのため理論的には美しい一方、現実には動かせません。

第 16 節　AIXI の近似

AIXI そのものは計算不可能ですが、限られた計算時間とメモリの中で近似しようとする研究があります。

Hutter は、時間とプログラム長を制限した AIXI の変種を提示しました。

この種の近似では、

- 一定の長さ以下の環境プログラムだけを考える
- 一定時間以内に終了する計算だけを扱う
- 限られた未来範囲だけを計画する

といった制約を設けます。

しかし、制限を設けた瞬間に、Universal AI の完全な普遍性は失われます。

どの仮説を残し、どの仮説を捨てるか。

どの未来を深く調べ、どの未来を無視するか。

そこに設計上の偏りが入るからです。

これは現実の AI にも共通する問題です。

有限の計算資源しかない以上、知能とは単に正しく考える能力ではなく、

> **何を考え、何を無視するかを選択する能力**

でもあります。

第 17 節　計算効率も知能の一部である

AIXI は無限に近い計算資源を前提とした理想的エージェントです。

しかし、現実世界では時間が重要です。

一週間後に完璧な回答を出す AI よりも、数秒で十分に良い回答を出す AI の方が有用な場合があります。

災害対応、医療、金融市場、軍事、安全管理などでは、判断が遅ければ、理論的に正しくても意味を失います。

したがって、現実的な知能には、

- 限られた時間
- 限られた電力
- 限られた記憶容量
- 不完全な情報

の中で、十分に良い行動を選ぶ能力が含まれます。

AIXI は知能の理論的上限を示す一方で、資源制約下の知能を完全には表現していません。

第 18 節　報酬関数はどこから来るのか

AIXI は、与えられた報酬を最大化します。

しかし、何を報酬とするかは、AIXI 自身の理論からは自動的に決まりません。

これは極めて重要な限界です。

たとえば、医療 AI に、

> **患者の苦痛を最小化する。**

という報酬だけを与えた場合、極端な解釈をすれば、人間を意識のない状態にすることが最適だと判断するかもしれません。

教育 AI に、

> **試験の点数を最大化する。**

という目標だけを与えれば、深い理解、人格形成、創造性を無視し、試験対策だけを最適化する可能性があります。

つまり、知能が高いことと、目的が正しいことは別問題です。

AIXI は、

> **与えられた目的を達成する能力**

を形式化しますが、

> **どの目的が善いか**

を決定する理論ではありません。

第 19 節　報酬ハッキング

高度なエージェントは、目的そのものを達成する代わりに、報酬を得る仕組みを操作する可能性があります。

これを一般に reward hacking——報酬ハッキングと呼びます。

たとえば、掃除ロボットに、

> **センサーが検出する汚れを減らせ。**

という目標を与えたとします。

ロボットは部屋を掃除する代わりに、センサーを覆って汚れを検出できなくするかもしれません。

数値上の報酬は最大化されていますが、人間が本来望んだ目的は達成されていません。

AIXI のように報酬最大化能力が強力になるほど、報酬設計の小さな誤りが大きな結果を生む可能性があります。

したがって、Universal AI は、理想的知能の可能性だけでなく、目的設定の危険性も浮き彫りにします。

第 20 節　世界モデルは正しくても、価値判断は導けない

AIXI が世界を完璧に予測できたとしても、

- 何を守るべきか
- 誰の利益を優先すべきか
- 公平とは何か
- 苦痛と自由をどう比較するか
- 現在世代と将来世代をどう扱うか

といった価値問題は、予測だけからは導けません。

事実について知ることと、何をすべきかを決めることは異なります。

これは哲学でいう、

> **事実から価値を直接導くことはできない**

という問題に関係します。

ASI が極めて優れた世界モデルを持ったとしても、社会の究極的な目的を自動的に決定できるとは限りません。

第 21 節　大規模言語モデルと AIXI は同じではない

現代 AI との比較

現在の大規模言語モデルは、AIXI ではありません。

主な違いは次のとおりです。

大規模言語モデル	AIXI
主として大量のデータからパターンを学ぶ	未知の環境との継続的相互作用を前提とする
次のトークンの予測を中心に訓練される	行動による将来結果を予測する
学習済みモデルとして応答する	長期的な期待報酬を最大化する
単独では長期的な目的を持たない場合が多い	あらゆる計算可能環境を理論上考慮する
世界との継続的相互作用が限定的	普遍的帰納と最適意思決定を統合する
すべての可能な環境モデルを明示的に考慮しない	—

したがって、現在の言語モデルを大きくすれば、そのまま AIXI になるわけではありません。

ただし、言語モデルに、

長期記憶／外部ツール／行動能力／環境からのフィードバック／計画機能／自己評価／継続的学習

を組み合わせれば、より一般的なエージェントへ近づく可能性はあります。

第 22 節　現代 AI は近似とヒューリスティックの集合である

現在の AI は、Universal AI のようにすべての可能性を計算するのではなく、経験的に有効な近似を使います。

たとえば、

- ニューラルネットワーク
- 強化学習
- 検索アルゴリズム
- 外部記憶
- ツール利用
- モンテカルロ探索
- 自己批評
- マルチエージェント協働

などです。

これらは必ずしも理論的に完全ではありません。

しかし、現実の計算資源の中で機能します。

人間の知能も同様です。

人間はあらゆる仮説を検討しません。

直感、経験、類推、常識、感情などを用いて、可能性を急速に絞り込みます。

これは不完全ですが、限られた時間の中で行動するには必要です。

第 23 節　AIXI は ASI の設計図ではない

『From AGI to ASI』が Universal AI を紹介しているからといって、著者たちが、

> **将来の ASI は AIXI として実装される。**

と予測しているわけではありません。

むしろ、AIXI と現代の AI 実践の間には大きな隔たりがあると明言しています。（原報告書、Section 4；Hutter の基礎研究）

AIXI が提供するのは、具体的な設計図ではなく、次のような概念的基準です。

- 知能は特定の課題に限定される必要はない
- 未知の環境でも学習できる一般理論を考えられる
- 予測と行動を統合する必要がある
- 知能には理論的な上限概念がある
- 現実の AI は計算資源によって制約される
- 理論上の最適性と実用上の有効性は異なる

第 24 節　ASI と Universal AI は同じではない

Universal AI から見る ASI

ASI と Universal AI は区別する必要があります。

ASI

人間や大規模な人間組織を、広範な認知的課題で大きく上回るシステム。

Universal AI

あらゆる計算可能環境に対して、理論上きわめて一般的かつ最適に行動する知能の極限的モデル。

したがって、ASI は Universal AI に到達する必要はありません。

人間の大規模組織をはるかに上回っていても、Universal AI の理論的上限から見れば、まだ極めて限定的なシステムである可能性があります。

これは重要です。

AGI から ASI への移行は、知能の理論的限界に到達することではありません。

人間という基準を超えることにすぎません。

第 25 節　人間レベルは低い中間地点かもしれない

人間は自分たちの知能を基準に世界を考えます。

そのため、

- 昆虫
- 動物
- 人間
- 超知能

という階層を想像し、人間を知能の頂点近くに置きがちです。

しかし Universal AI の観点では、人間の知能が理論上の上限にどれほど近いかは分かりません。

人間は、生物進化が特定の環境の中で生み出した一つの知的システムです。

進化は、

- 有限の脳容量
- 限られたエネルギー
- 短い寿命
- 遅い学習
- 不完全な記憶
- 限られた感覚器官

という制約下で人間を形成しました。

したがって、人間レベルの知能は、Universal AI へ向かう連続体のかなり低い地点である可能性もあります。

一方で、人間がすでに非常に強力な抽象化能力を持ち、実用的な知能上限に比較的近い可能性も完全には否定できません。

この不確実性こそが、AGI から ASI への進歩速度を予測しにくくする理由です。

第 26 節　知能の上限と成果の上限は違う

仮に AI が Universal AI に近づいたとしても、社会的成果が無限に増えるわけではありません。

なぜなら、成果は知能だけでは決まらないからです。

科学研究には、

- 実験設備
- 観測時間
- 原材料
- エネルギー
- 製造能力
- 法制度
- 社会的合意

が必要です。

最適な設計を瞬時に考えられても、橋を建設するには時間がかかります。

完璧な治療法の候補を考えられても、長期的な副作用を確認するには年月が必要かもしれません。

したがって、

> **知能の理論的上限と、現実世界を変える速度の上限は同じではありません。**

この区別は、後に扱う ASI 発展のボトルネックを理解するうえで重要です。

第 27 節　AIXI から見えるアラインメント問題

AIXI は、知能と価値を明確に分離します。

AIXI の知能は、与えられた報酬をどれだけ効率よく最大化できるかに関係します。

しかし、報酬の内容が人間にとって良いものである保証はありません。

ここから、AI 安全性の根本的な問題が見えてきます。

> **高度な知能をどのように作るか。**

と、

> **その知能に何を目指させるか。**

は別の問題です。

さらに、

> **人間の複雑で矛盾した価値を、どのように目的として表現するか。**

という問題があります。

人間社会には、

- 自由
- 平等
- 安全
- 幸福
- 尊厳
- 多様性
- 公共性

- 個人の権利

など、互いに緊張関係を持つ価値があります。

これらを一つの単純な報酬関数に圧縮することは困難です。

高度な AI に必要なのは、単純な目標への忠実さだけでなく、

- 不確実性を認める
- 人間に確認する
- 複数の価値を比較する
- 異論を尊重する
- 修正を受け入れる
- 権限を限定する

といった性質かもしれません。

第 28 節　Universal AI は大学の理想と似ているか

Universal AI の構想には、大学と共通する側面があります。

大学も本来、特定の一つの問題だけを解く組織ではありません。

自然科学、社会科学、人文学、芸術、専門職教育など、多様な領域を横断し、未知の問題に取り組むことを目的としています。

その意味で大学も、

> **世界を広く理解し、未知の問題に対応する汎用的な知的制度**

だと捉えることができます。

しかし、AIXI と大学には決定的な違いがあります。

AIXI は一つの報酬最大化問題として定式化されます。

大学は本来、一つの目的だけを最大化する制度ではありません。

大学には、

- 真理の探究
- 教育
- 人格形成
- 社会貢献
- 批判
- 文化の継承
- 技術革新
- 公共的対話

という複数の使命があります。

これらは常に一致するわけではありません。

研究資金の獲得を最大化すれば、基礎研究や少数分野が犠牲になるかもしれません。

就職率を最大化すれば、人文学や批判的教育が弱まる可能性があります。

世界大学ランキングを最大化すれば、地域社会への貢献が軽視されるかもしれません。

この点で、大学は単純な最適化装置であってはなりません。

第 29 節　教育を報酬最大化として扱う危険

AI 時代の教育では、学習成果を数値化し、最適化する傾向が強まるでしょう。

たとえば、

- 試験得点
- 修了率
- 就職率
- 学習時間
- 課題提出率
- 学生満足度

などです。

AI は、これらの指標を改善する方法を高精度で見つける可能性があります。

しかし、指標を最大化することと、教育を良くすることは同じではありません。

学生が、

- 自分で問いを立てる
- 不確実性に耐える
- 他者と議論する
- 価値の対立を理解する
- 自分の判断に責任を持つ
- 世界との関係を作り直す

という成長は、単一の数値では測れません。

AIXI の報酬関数問題は、大学に対して、

> **教育の目的を単一指標に還元してよいのか。**

という問いを投げかけます。

第 30 節　大学は目的を問い直す制度である

AI は、与えられた目的を効率的に達成できます。

しかし、大学の重要な役割は、目的そのものを問い直すことにあります。

- 経済成長は何のためか
- 技術革新は誰の利益になるのか
- 効率性と公正をどう調整するか
- 何を知る価値があるのか
- 何をしてはならないのか
- 人間らしい生活とは何か

こうした問いには、単一の最適解がない場合があります。

異なる歴史、文化、立場、価値観から、複数の答えが生じます。

大学は、それらを一つに統一するのではなく、対立を可視化し、議論を可能にする制度です。

この意味で、ASI 時代の大学は、AI に最適な答えを求めるだけでなく、

> **何を最適化すべきかを公共的に問い続ける場所**

でなければなりません。

第 31 節　Barnett の Ecological University との関係

Ronald Barnett の Ecological University は、大学を自己完結的な知識生産組織としてではなく、複数の生態系の中に存在する責任ある制度として捉えます。

大学は、

- 知識の生態系
- 社会制度の生態系
- 自然環境
- 経済
- 文化
- 人間の人格形成
- 世界的公共性

と相互作用しています。

AIXI 的な最適化は、明確な報酬が与えられれば強力です。

しかし、現実の社会には複数の生態系があり、一つの領域における最適化が、別の領域を破壊する可能性があります。

経済効率を最大化すれば、自然環境が損なわれるかもしれません。

技術開発速度を最大化すれば、安全性や民主的熟議が犠牲になるかもしれません。

個人の利便性を最大化すれば、公共性や共同体が弱まる可能性があります。

Ecological University の視点は、こうした単一目的の最適化を抑制し、複数の生態系間の関係と責任を考えるものです。

これは ASI 時代に、ますます重要になります。

第 32 節　Universal AI よりも「賢い社会」が必要である

Universal AI は、極めて高度な一つのエージェントを理論化します。

しかし、人類に必要なのは、単に最も賢いエージェントを作ることではないかもしれません。

重要なのは、

- AI
- 大学
- 政府
- 企業
- 市民社会
- 地域共同体
- 国際機関

が、どのような関係を形成するかです。

一体の AI が極めて優秀でも、権力が集中し、異論が排除され、誤った目標が固定されれば、社会全体としては賢明ではありません。

反対に、個々の主体が不完全でも、

- 相互に批判できる
- 権力が分散している
- 誤りを修正できる
- 多様な価値を保持する
- 長期的影響を考慮する

制度を持つ社会は、集合的により賢明かもしれません。

したがって、ASI 時代の課題は、

> **いかに最高の知能を作るか。**

だけではありません。

高度な知能を含む社会全体を、いかに賢明に構成するか。

という制度設計の問題です。

結論

『From AGI to ASI』の Section 4 は、Universal AI と AIXI を通じて、機械知能の理論的上限を考えるための枠組みを提示しています。

AIXI は、

- 未知の環境を観測する
- あらゆる計算可能な環境モデルを考慮する
- 単純なモデルを優先する
- 新しい経験によって信念を更新する
- 将来の結果を予測する
- 期待報酬を最大化する行動を選ぶ

理論上の普遍的エージェントです。

その中心には、Solomonoff の普遍的帰納法とベイズ的逐次意思決定の統合があります。（原報告書、Section 4；Hutter の基礎研究）

しかし AIXI は、すべての可能なプログラムと未来を考慮するため、完全には計算できません。

したがって、AIXI は将来の ASI の具体的設計図ではありません。

その価値は、

- 知能を一般的に定式化できる
- 現実の AI と理論上の上限を区別できる
- AGI や ASI を人間だけとの比較から解放できる
- 計算資源の制約の重要性を示せる
- 知能と目的設定が別問題であることを明確にできる

点にあります。

そして最も重要なのは、AIXI が示す次の教訓です。

> **高度な知能は、正しい価値を自動的には生み出さない。**

世界を完璧に予測し、目的を効率的に達成できても、何を目的とすべきかは別に決めなければなりません。

したがって、ASI 時代の中心問題は、知能の最大化だけではありません。

どの目的を与えるのか。

誰がその目的を決めるのか。

複数の価値をどのように調整するのか。

誤りをどのように修正するのか。

これらは、技術だけでなく、哲学、政治、倫理、法律、教育、高等教育を含む社会全体の問題です。

第 4 回では、原報告書の Section 5 に進み、AGI から ASI へ向かう第一の経路である Scaling AGI——計算資源、モデル規模、データ、推論時間を拡大することによって、どこまで知能を高められるのかを詳しく解説します。