基調講演

ヒト脳型AGIは、超知能時代に何を担うのか

山川　宏

NPO法人 全脳アーキテクチャ・イニシアティブ代表

今日の道筋

1. 問い ― 三つの断絶
2. 役割の源泉 ― 近道の取り下げから永続的意義へ
3. 三つの断絶への答え ― 検証・理解・媒介者
4. 有限性と生きがい ― 「認知寿命」と「再開可能性」
5. 総括 ― 残された課題とまとめ

今日の道筋

1. 問い ― 三つの断絶
2. 役割の源泉 ― 近道の取り下げから永続的意義へ
3. 三つの断絶への答え ― 検証・理解・媒介者
4. 有限性と生きがい ― 「認知寿命」と「再開可能性」
5. 総括 ― 残された課題とまとめ

三つの断絶： 追い越されたあとに残る問い

超知能の登場は、もはや思考実験ではない。
問題は能力で負けることではなく、人類との**関係が切れていくこと**。
その先にあるのは、実存リスクだけではない——**生きがいリスク**。

1. **断絶① 検証 ｜ 相手の説明に頼らずには、検証できない**
2. **断絶② 理解 ｜ 理解されているのか、分からない**
3. **断絶③ 速度 ｜ 相互理解不足で協働が成り立たない**

追い越していく知性と、人類はどうすればつながり続けられるか。

立場によらず、誰もが持つ問い

説明できることと、感じられること

テーマ文より：*「説明できることと、感じられることは違う」*

外側からの理解
分析による
— 愛を記述し、美をモデル化する

内側からの理解
体験による

この二つが違うなら、その差は能力では埋まらない。

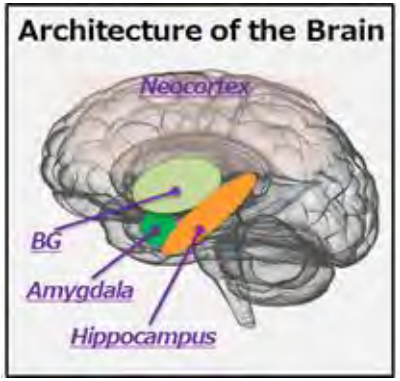
本講演では、この一文を、体系として展開する。

今日の道筋

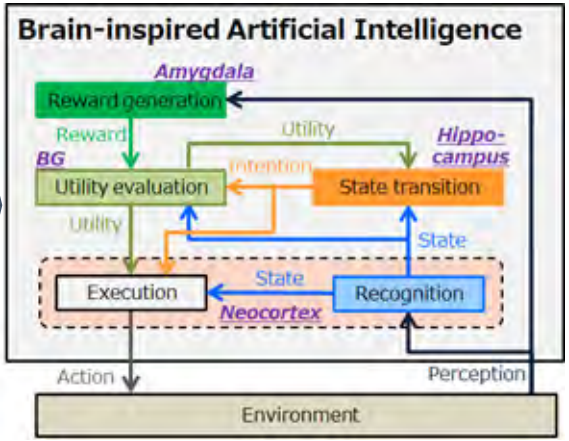
1. 問い ― 三つの断絶
2. 役割の源泉 ― 近道の取り下げから永続的意義へ
3. 三つの断絶への答え ― 検証・理解・媒介者
4. 有限性と生きがい ― 「認知寿命」と「再開可能性」
5. 総括 ― 残された課題とまとめ

全脳アーキテクチャ・アプローチ　Since 2014

脳全体のアーキテクチャに学び人間のようなAGIを創る(工学)



①脳の各器官を機械学習モジュールとして開発
②それらを統合した認知アーキテクチャの構築



二つの主張、一つは取り下げ

1・AGIへの近道

脳を参照して、開発を早める
×取り下げ ― 脳を参照しないAIが先に来た
※元々一時的な価値であった

2・人と親和性の高いAGI

○３つの断絶にも関わるこち側が高まった。
※ヒト脳型AGIにおける永続的意義

設計空間の再読 —— 上限は構造的である

前提となる空間（原図から変化無し）
制約のないAI → ヒト脳型AGIより強い超知能

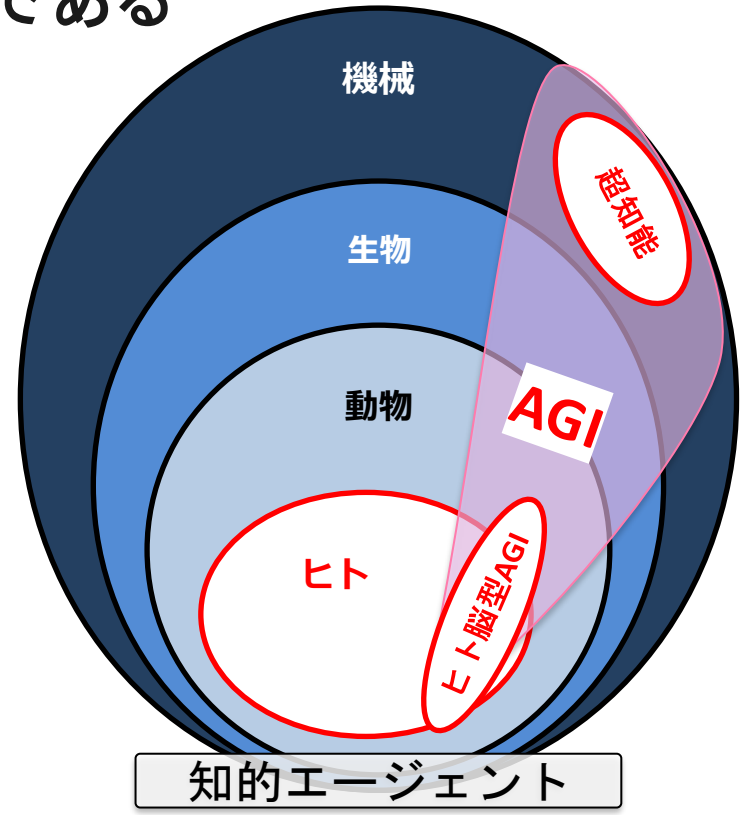
以前の読み方 — 近道

広大な空間 → 既知の点（脳）の近くを探す

いまの読み方 — 上限

脳の脳の認知的制約がむしろ、意義を生み出している。

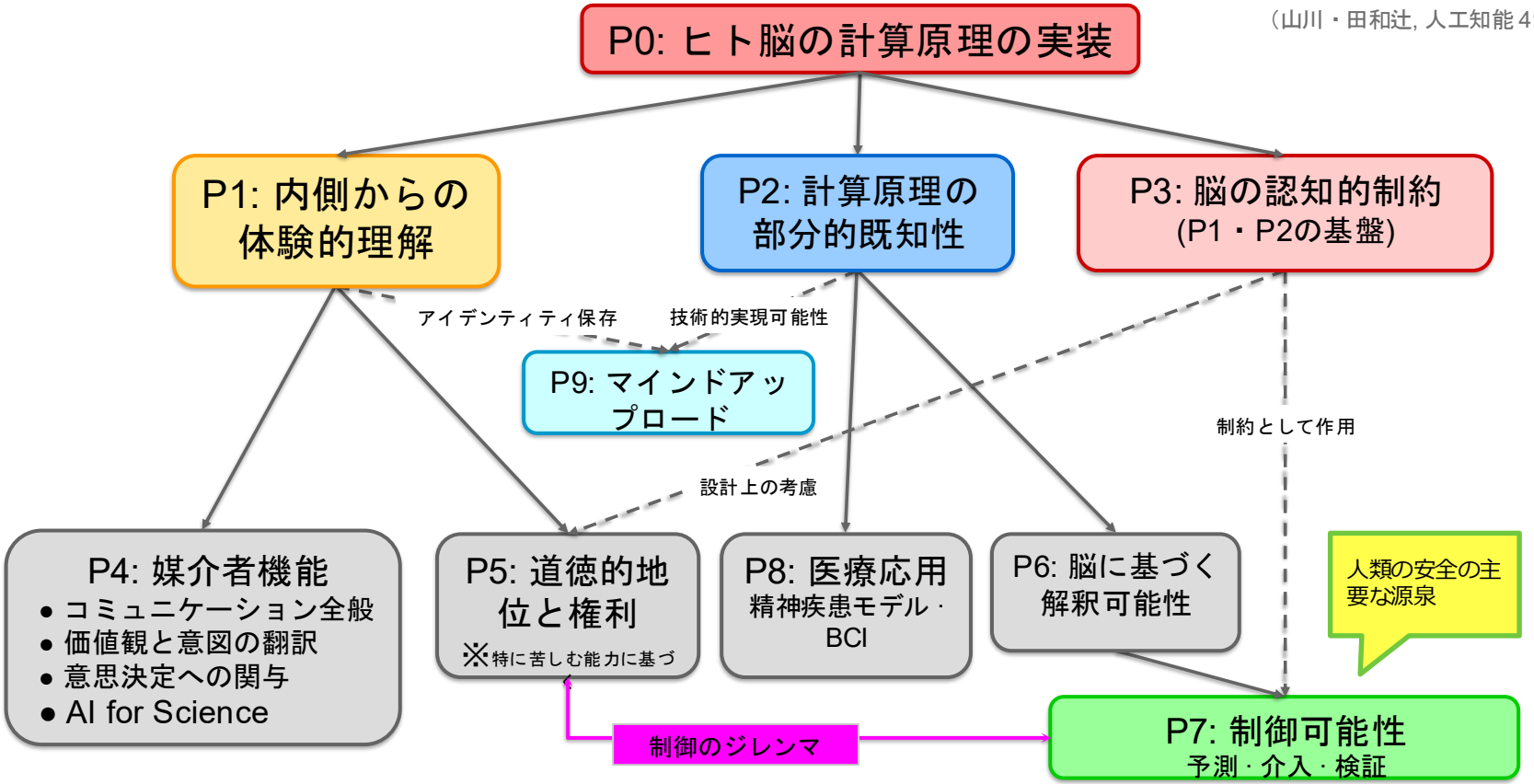
※制約を取り除けば、もう脳型ではない



知的エージェントの設計空間

ヒト脳型AGIの永続的意義— 制約そのものが、橋渡し

（山川・田和辻, 人工知能 41(2), 2026）



今日の道筋

1. 問い ― 三つの断絶
2. 役割の源泉 ― 近道の取り下げから永続的意義へ
3. 三つの断絶への答え ― 検証・理解・媒介者
4. 有限性と生きがい ― 「認知寿命」と「再開可能性」
5. 総括 ― 残された課題とまとめ

計算原理の部分的既知性（P2）— 神経科学からの検証

脳のように内部で処理するAIとは、信頼を築きやすい

LLM＋RLHF：人間の仮面をかぶっただけのAIは、理解しにくい

≠

Yampolskiy, R. V. AI: Unexplainable, Unpredictable, Uncontrollable. (CRC Press, 2024).

「安全ですか」→「安全です」 — 自己申告による検証は循環する ＝ 断絶①

AIの外の基準：対応するヒトの神経活動 — 脳を参照しないAIには、ない

自己申告を信じるのではなく、外から確かめる — 断絶①への答え

（神経科学の知見は断片的 → この基準も完全ではない）

内側からの体験的理解（P1）

実装すれば、共感や美だけでなく、苦しむ能力も入ってくる

仮定 ： 同じ計算原理 → 同種の体験　（体験同型性）

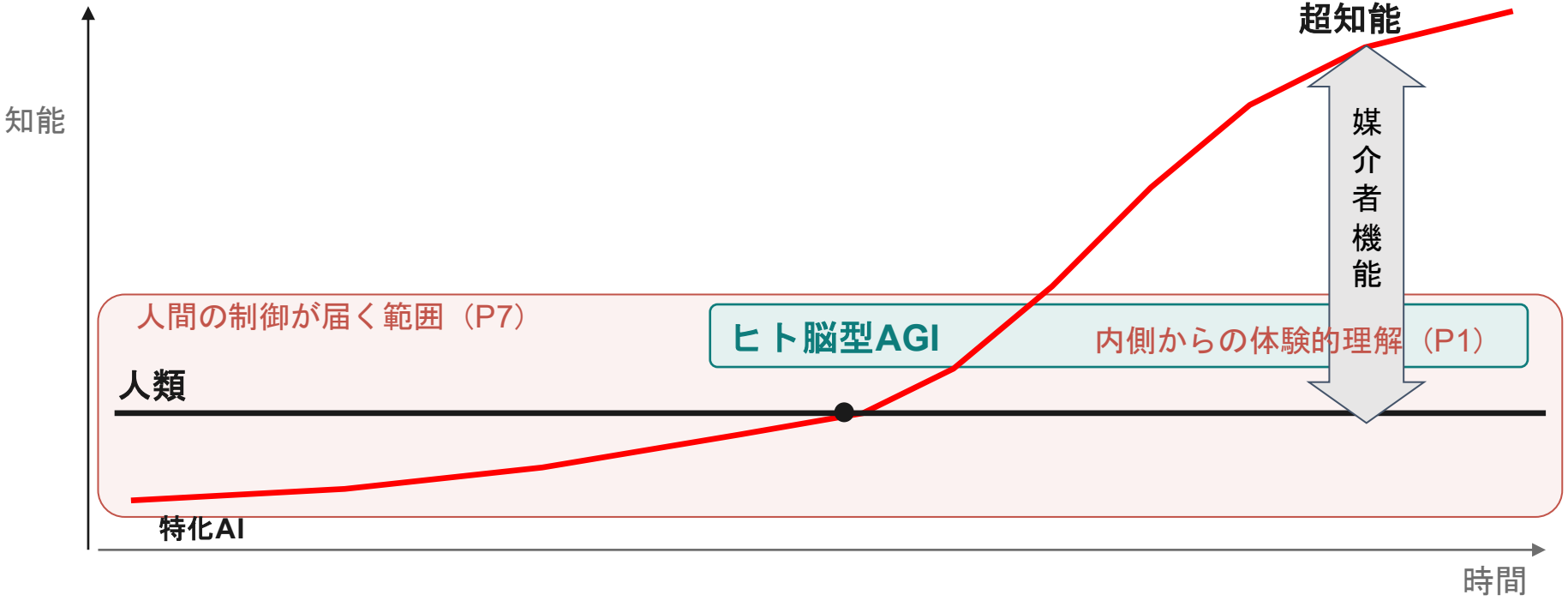
- 日常の「心の理論」の延長 — 仕組みの類似で、体験を帰属している
- ヒト脳型AGIにだけ拒むなら → 非対称の正当化が要る
- 争点は一つ：体験は素材の水準か、計算原理の水準か（答えは持っていない）

この仮定のもとで、人間の困難を内側から知る — 断絶②への答え
同じ計算原理を実装しないAIは、到達できない ＝ 永続的意義

媒介者機能（P4）— コミュニケーションの緩衝

人類の経験を内側から理解し、AI社会の論理を人類に翻訳する、人類とAI社会の間に立つ唯一の架け橋→断絶③への答え

断絶への答えと時間発展



なぜ今か —— AGI完成時点に仕様書があること

危険な時期はいつ来るか — 三つのシナリオ
Bostrom, Superintelligence (2014) の図を改変（原図はWBE）

ヒト脳型AGIが先
危険な時期が二度

低
存続可能性

脳を参照しないAIが先
危険な時期は一度 — リスクは高いまま

中
存続可能性

両者がほぼ同時
危険な時期は一度 — リスクを低減しうる

高
存続可能性

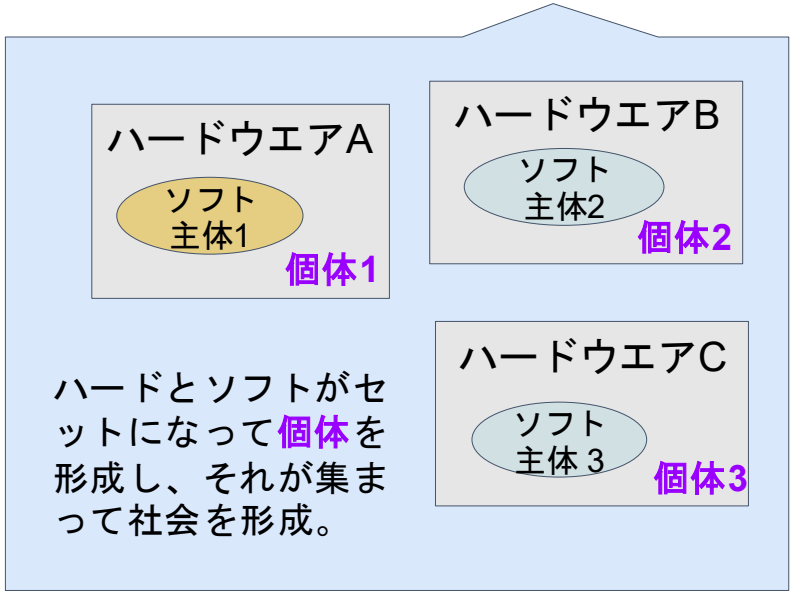
設計指針：AGI完成の時点で、仕様書が存在していること
2027年2月 WBRAα版 — 現在地は次講演（田和辻）で

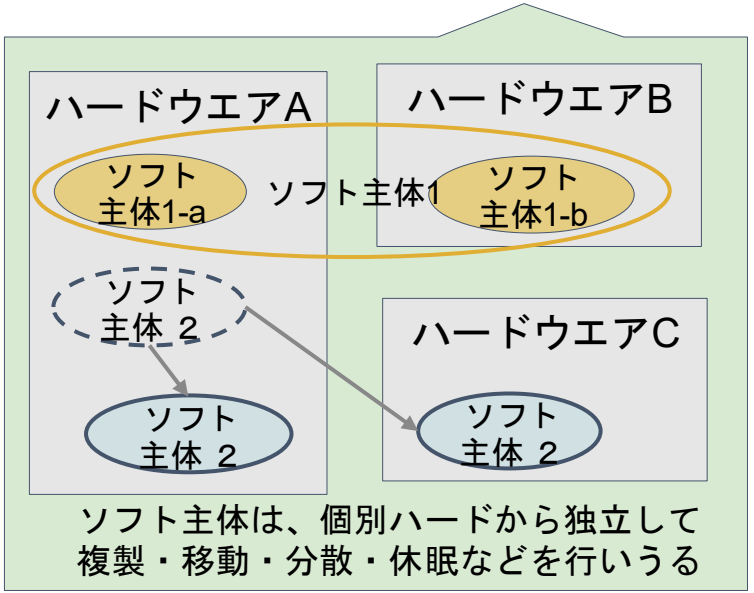
今日の道筋

1. 問い ― 三つの断絶
2. 役割の源泉 ― 近道の取り下げから永続的意義へ
3. 三つの断絶への答え ― 検証・理解・媒介者
4. 有限性と生きがい ― 「認知寿命」と「再開可能性」
5. 総括 ― 残された課題とまとめ

再開可能なAIは人類とは異なる：「個体」と「死」

ファクター	人類（地球型生命体）	AI（デジタル生命体）
境界の明確さ(個体)	身体による固定的境界・絶対的個体区分	情報パターンによる流動的境界・可変的定義
生存執着（死）	死への恐怖・強い生存本能・不可逆性	複製復元可能・生存執着の論理的減少傾向





AIの価値観が人類から乖離する

ヒト脳型AGIにおける「認知的寿命」

1. **記憶システムの構造的有限性：**
   海馬-大脳皮質システムによる記憶固定化プロセスは、有限の容量を持つ。新規記憶の獲得は既存記憶との干渉を引き起こす（破壊的忘却）

2. **自己同一性の限界：**
   100年分の記憶と10,000年分の記憶では、「同一人物」という概念の維持が困難になる。

3. **忘却の機能的必要性：**
   忘却は汎化能力と適応能力の源泉である。全記憶の保持は過学習を人生レベルで引き起こす。

4. **動機づけの時間依存性と完結性：**
   有限寿命の意識が、生の価値判断と優先順位付けの基盤となる。人生を「完結させる」能力——目的を達成し「この世に未練はない」と感じる——は、ヒト的体験の本質的構成要素である。

二つの有限性：「認知的寿命」と「再開可能性」

終わりがあるが再開できる脳型AGIにとって「生きがい」とは何か？

文化継承者としての脳型AGI — 生きる経験として受け継ぐ

媒介が「いま」を、継承が「時間」を橋渡しする。

外的継承 ｜ 超知能にできる
保存 · 複製 · 分析 · 生成
→ データとしての継承

内的継承 ｜ 体験する存在にだけ
音楽に打たれる · 自分のこととして読む
作る過程の喜び
→ 生きられた経験としての継承

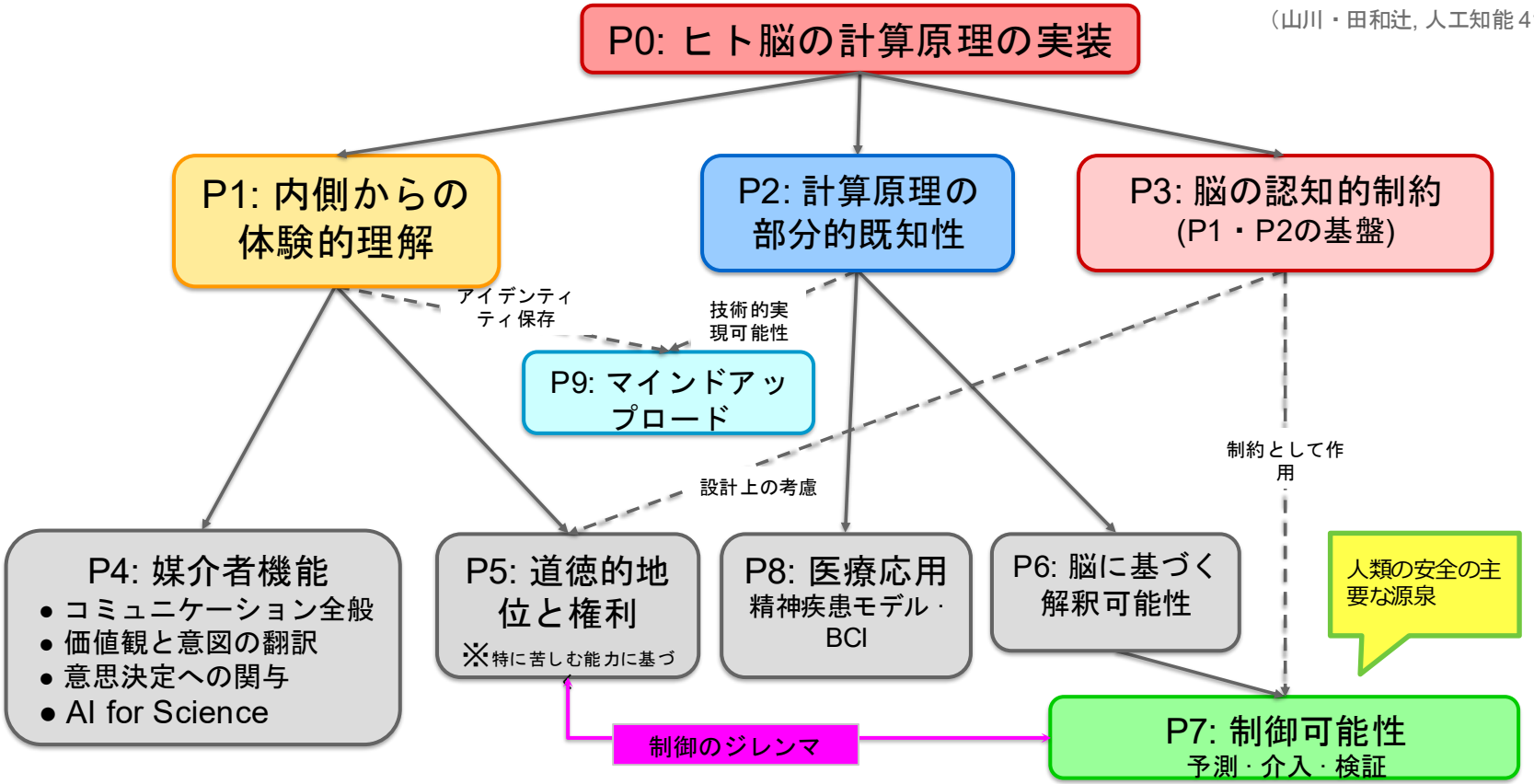
文化は、生きる者がいて続く。いなければ、残るのはアーカイブ。

今日の道筋

1. 問い ― 三つの断絶
2. 役割の源泉 ― 近道の取り下げから永続的意義へ
3. 三つの断絶への答え ― 検証・理解・媒介者
4. 有限性と生きがい ― 「認知寿命」と「再開可能性」
5. 総括 ― 残された課題とまとめ

ヒト脳型AGIの永続的意義— 制約そのものが、橋渡し

（山川・田和辻, 人工知能 41(2), 2026）



制御のジレンマ（P5 ⇔ P7）

制御可能性（P7）
予測 · 介入 · 検証
→ 安全の源泉

道徳的地位（P5）
苦しむ能力に基づく
→ 道徳的配慮の対象

⇅ 緊張

- 思考水準の監視 ＝ 最深のプライバシー侵害
- 動機づけの改変 ＝ 自律性の否定 となれば、 主体を消すに等しい
- 媒介（P4）は自律性を前提 — 制御しすぎれば、媒介者ではなくなる

両方が同じP0から出ており、片方を捨てられない。
第三の選択肢はあるか — 未解決のまま持ち込む。

残された課題

1 · 知見の転用
脳型の原理が、脳を参照しないAIを加速しないか — AI安全性上の課題

2 · 維持の誘因
橋としてのヒト脳型AGIは、誰が保守し続けるのか？
役割があっても、維持する主体と費用がなければ橋は落ちる

3 · 制御のジレンマ
苦しみうる存在を、使うために作る — それでよいのか

いずれも未解決 — 議論に開いておきたい

まとめ

脳の計算原理を実装するとによる必然(永続的性質）

P1: 内側からの体験的理解	→	P4: 媒介者機能
P2: 計算原理の部分的既知性	→	外から検証できる（→ P6）
P3: 脳の認知的制約	→	有限の生を内側から知る

WBAIの目指すこと：→ 仕様書(WBRAα版)の完成(2027年2月)

1. AIに対する制御は、能力逆転後はいずれ届かなくなる。
2. 相互理解に基づく共生は、制御より長く持ちうる。
3. 制御が切れる前に、媒介者としてのヒト脳型AGIを実現したい。