

基調講演

その大会の基本的な方針や中心的な
内容・目的について論じること

未来の知能とバイアス

山川宏

全脳アーキテクチャ・イニシアティブ、東京大学、
理研、近畿大学、電気通信大学、玉川大学

2022年度 人工知能学会全国大会（第36回）

2022年6月14日（火）



人工知能学会と私

OS: 脳とAI

OS: AGIと
社会への影響

OS: 人とAIが織りなす
新たなエコシステム

OS: 世界モデル
OS: System 1/2

-----AGI研究会-----

入会

論文賞授賞

全国大会実行委員長@岩手

合同研究会実行委員長

編集委員長

基調講演



20周年記念事業



大会
優秀賞

全国大会
委員

編集委員

30周年記念事業
「みんなで作る認知アーキテクチャ」

2000

2010

2022

編集委員今年の抱負
After Singularity 10年を思い描く

特 集 「編集委員今年の抱負 2013」

After Singularity 10 年を思い描く

山川 宏 株式会社富士通研究所



未来においては何れ避けがたく起こると想像されていた **Technological Singularity** (技術的特異点) が、昨今の技術状況を見渡すと思った以上に近づいているのではと思い始めている。Singularity とは、その先の未来における技術進化がそれまでの人類の歴史から推測し得なくなる限界点であり、あらゆる知的能力においてヒトを凌駕した超人的 AI (もしくは強化された人類) の出現による自己再帰的な知識増幅サイクルにより引き起こされる。

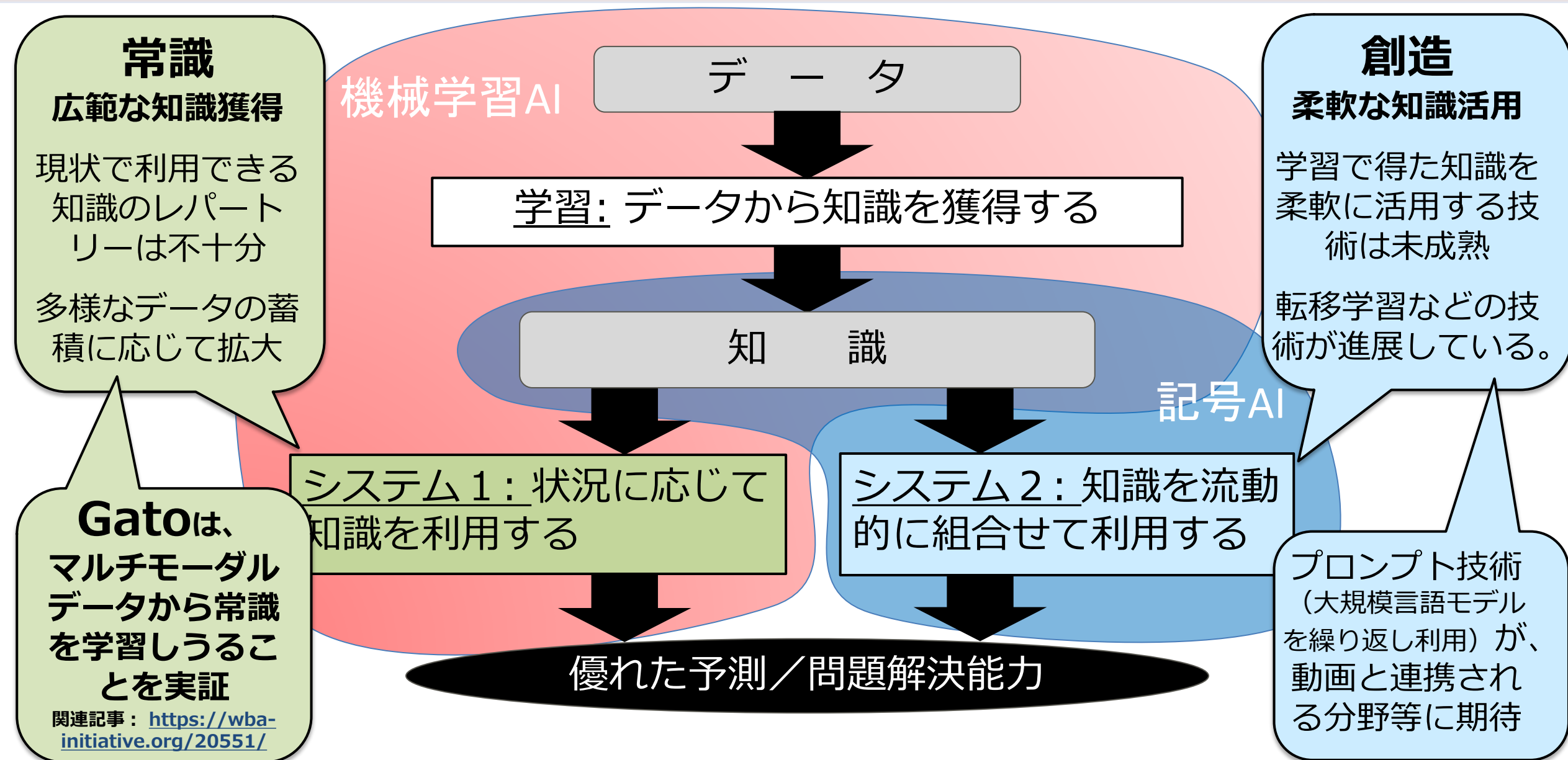
Singularity が、思いの外に近いと感じてきた理由は、月並みだが、人工エージェントがヒトを凌駕する例が増え続けているからである。例えば、Deep Blue, DARPA

で A.S.10 年には、そういうヒトの優れた能力を認める価値観自体が前時代の遺物となり、その頃の若者は、そもそもヒトが何らかの知的能力や闘争能力をアピールすること自体に何ら価値を見いだせなくなるかもしれない。こうした面のみならず A.S.10 年の人々の精神 (価値観やモラル) への影響は計り知れないであろう。

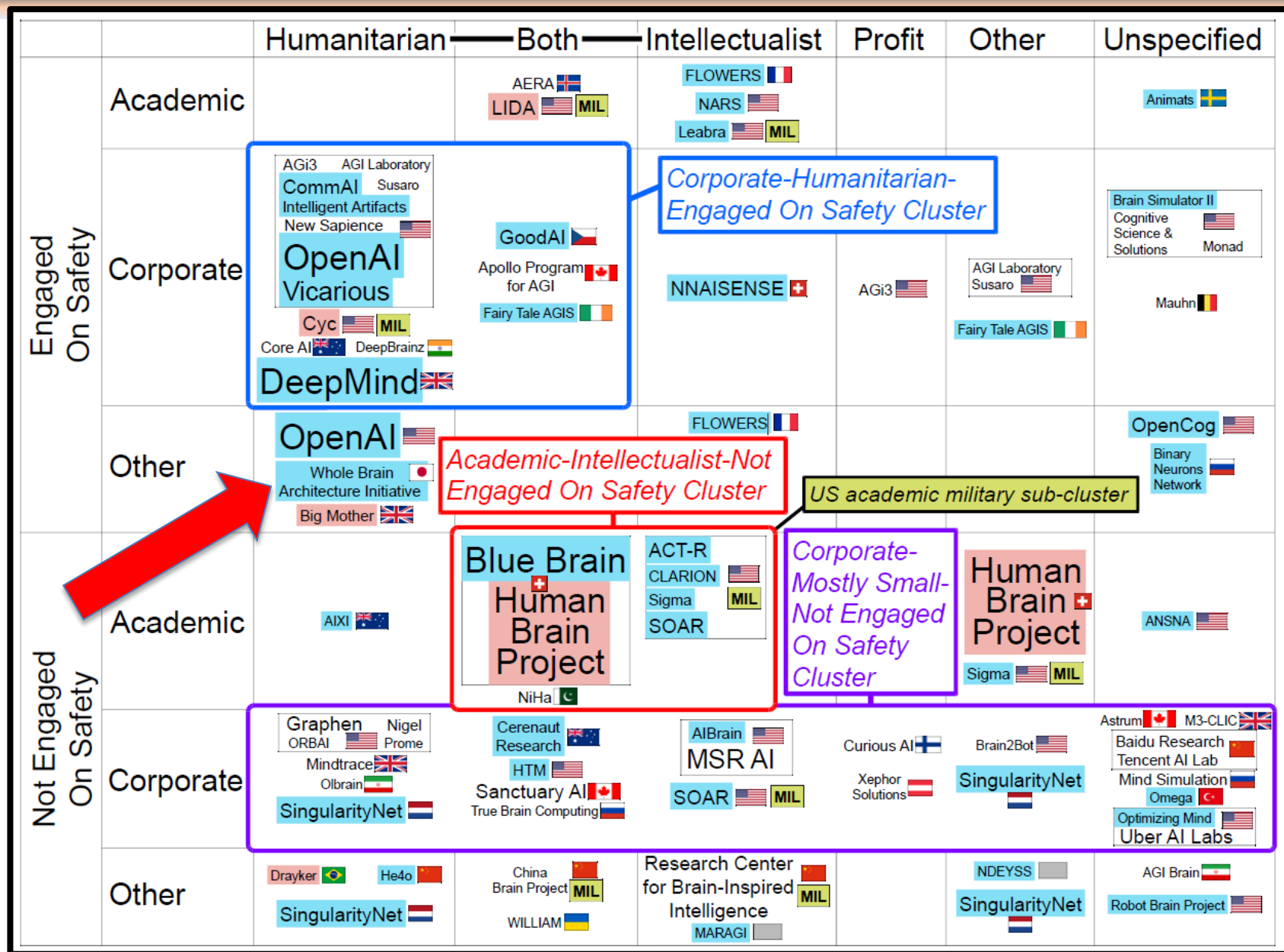
また A.S.10 年頃には有機体としてのヒト (脳) が超人的 AI により解明し尽くされ、人の能力もが大きく強化されるのかもしれないし、もしくは快感と幸福感が得られるようにコントロールされているかもしれない。

そして、たいていの超人的 AI は誰かにとって有益な目

Singularityの起点となる汎用人工知能(AGI)とその現状



汎用人工知能(AGI)を目指す 72 のR&D projects



2020年末時点で、AGI開発に関連するプロジェクトは少なくとも72件あり、2017年からの変化を見ると、総数はあまり変わっていない。ただし、この間に新たに15のプロジェクトが追加されており、そのほとんどが民間企業によるものである。

- **Institution type:** academic (mainly universities), corporate (public and private corporations), and other (government, nonprofit, and projects not affiliated with any institution).
- **Open-source code:** blue background indicates code available open-source; red background indicates code available upon request.
- **Military connections:** projects labeled MIL have military connections.
- **Nationality:** Flags indicate the country in which a project is based. Australia, Austria, Belgium, Brazil, Canada, China, Czech Republic, Finland, France, Iceland, India, Iran, Ireland, Japan, Netherlands, Pakistan, Russia, Sweden, Switzerland, Turkey, Ukraine, UK, USA, no project location identified.
- **Stated goals:** animal welfare, ecocentrism, humanitarian, intellectualism, profit, transhumanism, and unspecified (when project goals could not be identified).
- **Engagement on safety:** engaged projects either actively try to make their AGI safe or state that they support other AGI safety work; projects that are not engaged on safety either openly dismiss concerns about AGI safety or have no publicly stated activity on safety.
- **Size:** font size of project names indicates project size.

Fitzgerald, M., Boddy, A., & Baum, S. D. (2020). 2020 Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Technical Report, 20(1).

知識と知性の維持・向上・保存は
文化の進化をもたらす原動力であり、
知能は向上できるかぎり向上する

スティーヴン・J・ディック
(NASAの元歴史課 主任)

Dick, S. J. (2003). **Cultural evolution, the postbiological universe and SETI.** *International Journal of Astrobiology*, 2(1), 65–74.

アジェンダ

- 1. Entification -存在を見出す処理-
- 2. 全脳アーキテクチャ
- 3. 宇宙へ浸透する生命
- 4. 生命の願い
- 5. おわりに

汎用人工知能(AGI)の
実現はそれ程遠くない

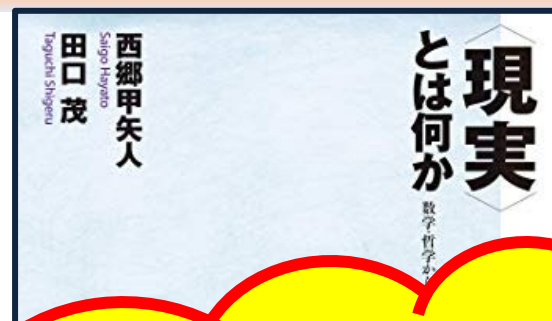
宇宙進出する地球生命はどうあるべきか？

現在の危険状況を超えて生存するために

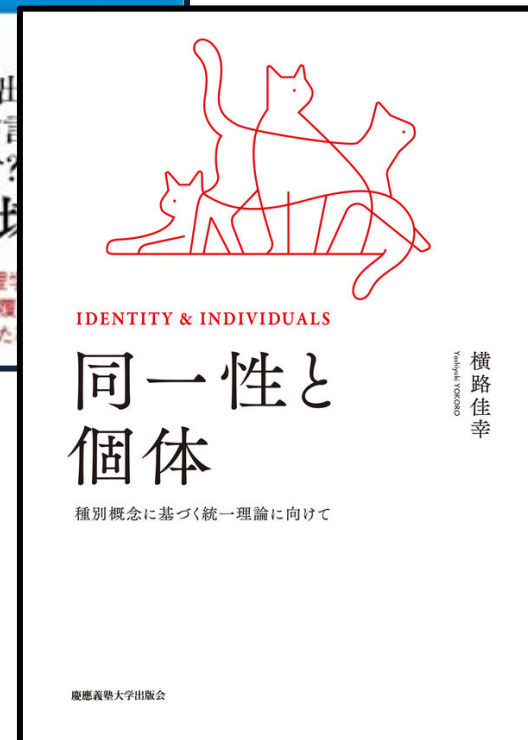
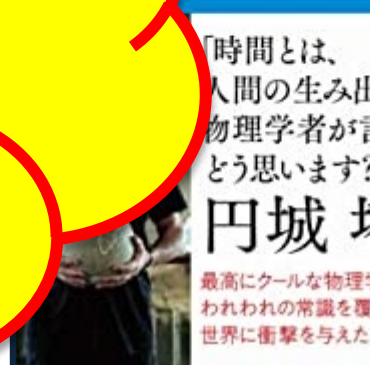
アジェンダ

- 1. Entification -存在を見出す処理-
- 2. 全脳アーキテクチャ
- 3. 宇宙へ浸透する生命
- 4. 生命の願い
- 5. おわりに

存在に関わる議論は今も続いている



少なくとも、世界の中に不変の存在があるという単純な構図ではない。



存在(Entity)を認識するためのEntificationという処理

- 存在とは世界の中に不変にある何かではないが、では何か？
 - アリストテレスの時代から、哲学では包括的な議論（「変換」の視点が弱い）
- 世界は殆どEntityによって記述され、人類の知識はEntityに依存する
 - 伝統的な記号AIにおいては、人がEntityを指し示す記号を記述してきた
- 視覚の機械学習の発展はEntityを捉えつつあり、大いに参考となる
 - Contrastive Learning、Object centricな表現
- 世界の中にあると想定しうる「存在とその体系」を探索しつづける処理を**Entification**と呼ぶこととして、進めている研究を紹介したい

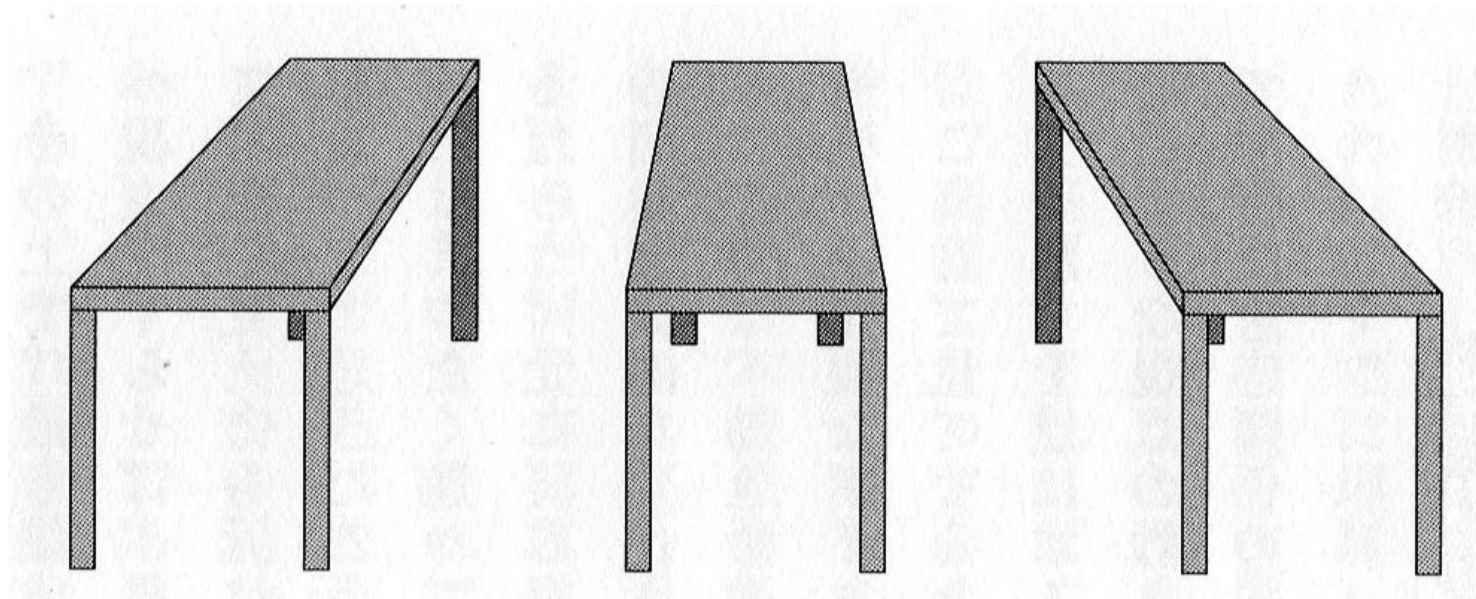
Entityとはなにか？

(W.V.O. Quine, 1969)

No **Entity** without **Identity**

(内包)

(外延)



机の現れの変化

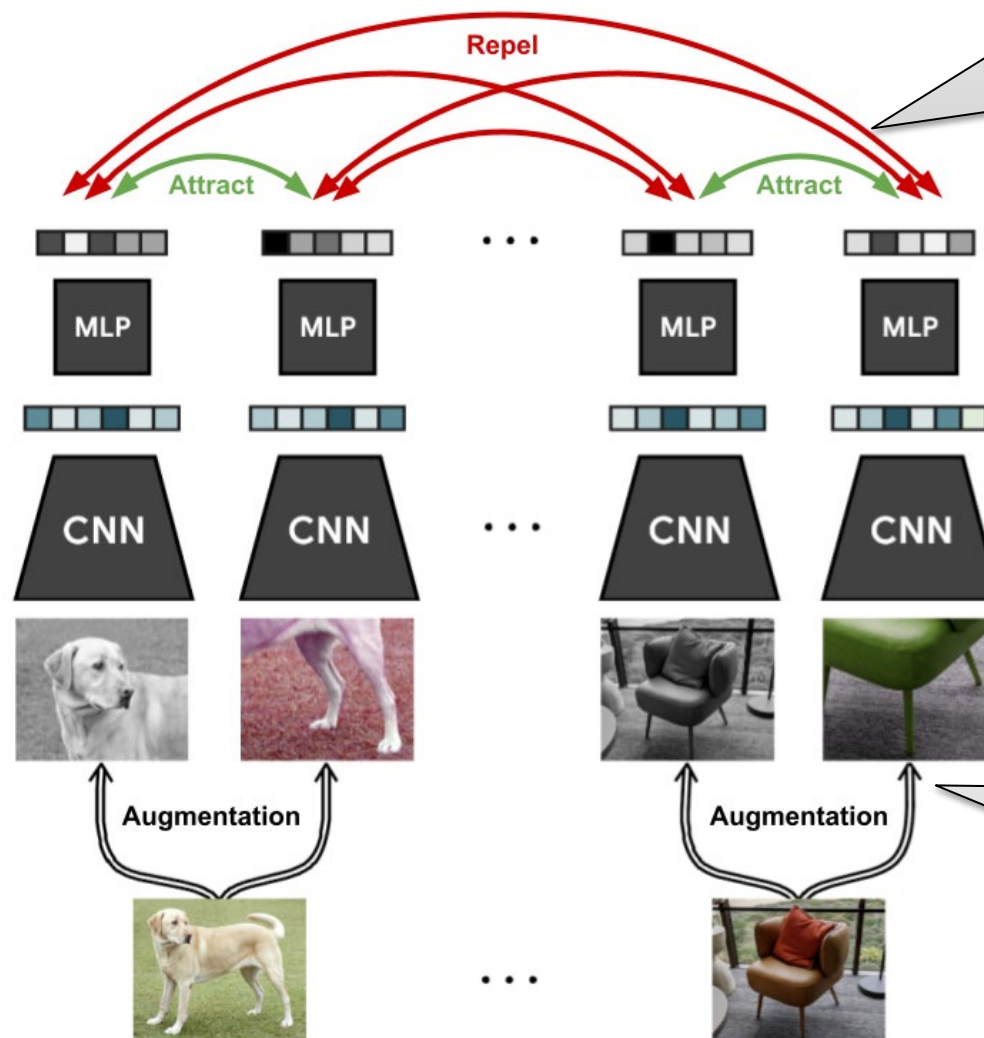
(西郷, 田口. 現実とは何か: 筑摩書房, 2019)

対照学習 (Contrastive learning)

SimCLR

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020).
A Simple Framework for
Contrastive Learning of Visual
Representations. In *Proc. ICML*

対照学習は、画像に対する自己教師あり学習の一種。現在は多くの研究で利用されている。



内部表現において、
同一の個物に由来するものは近づけ、そうでないものは遠ざける

Entity

**No Entity
without
Identity**

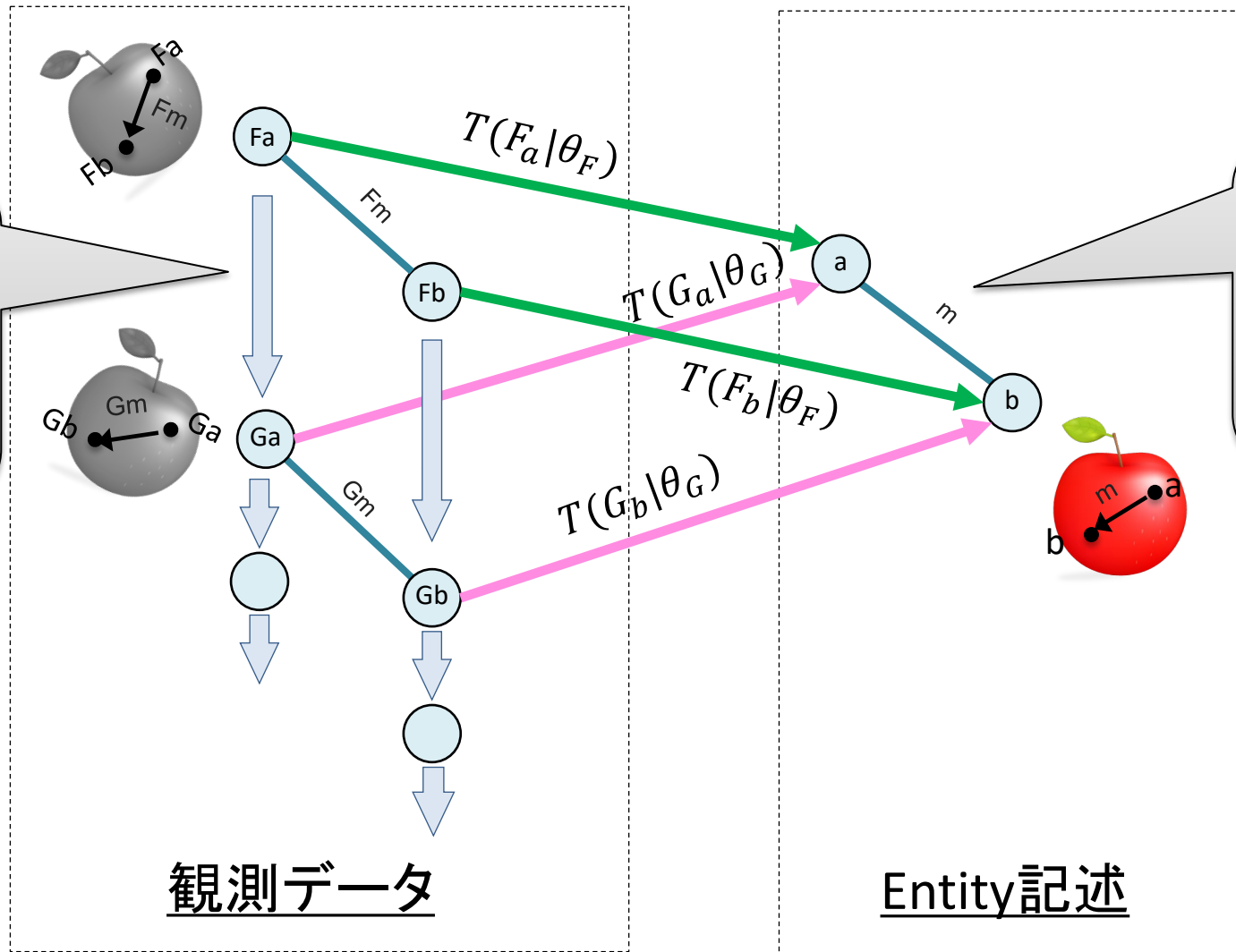
加工して複数化された
同一の個物の画像

Identity

Entificationの肝：空間の対称性を利用した情報圧縮処理

個物に対する観測
とは、同一の構造を
含む異なる見え

Identity



内部表現において、
同一の個物に由来する
構造が縮約される。

Entity

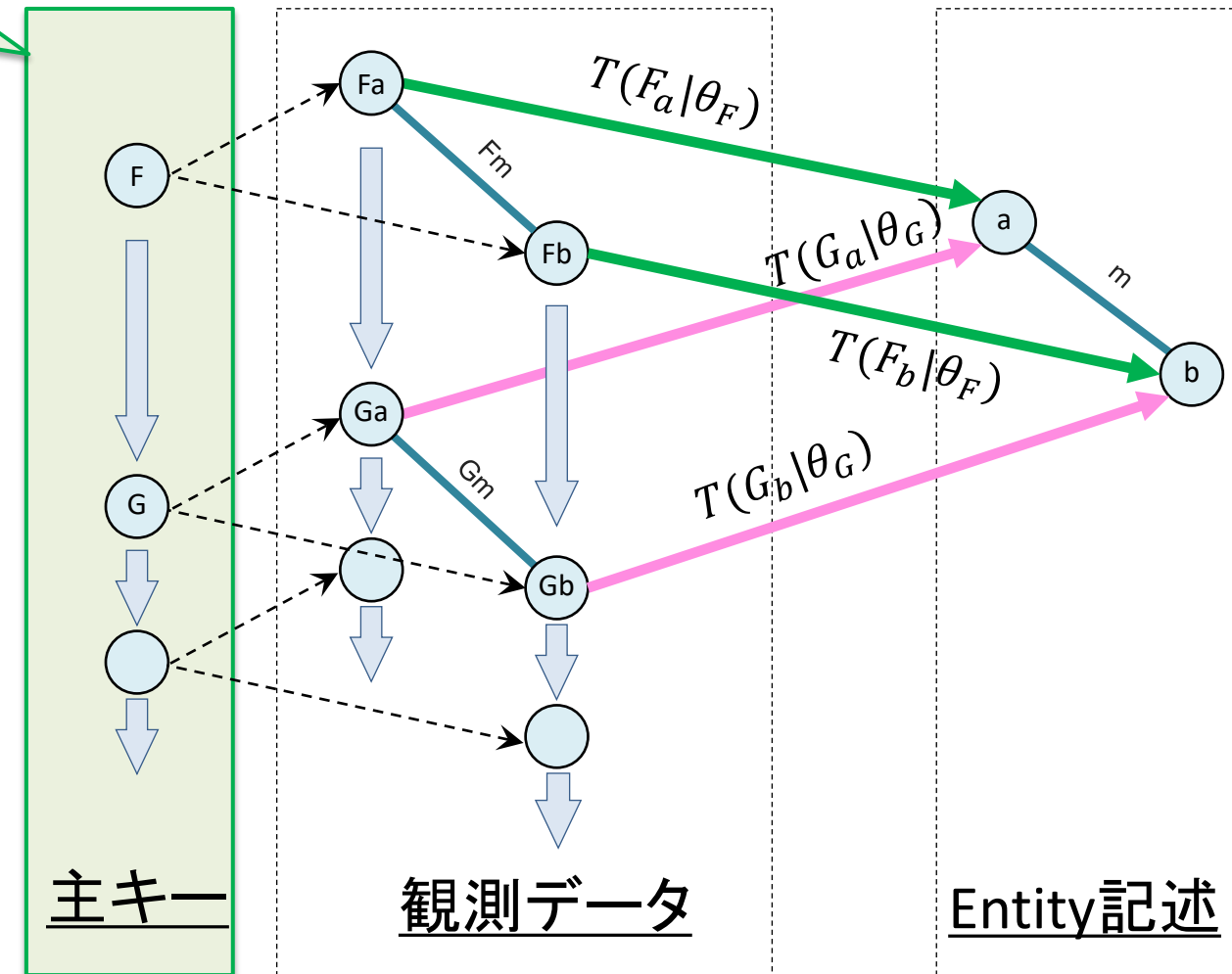
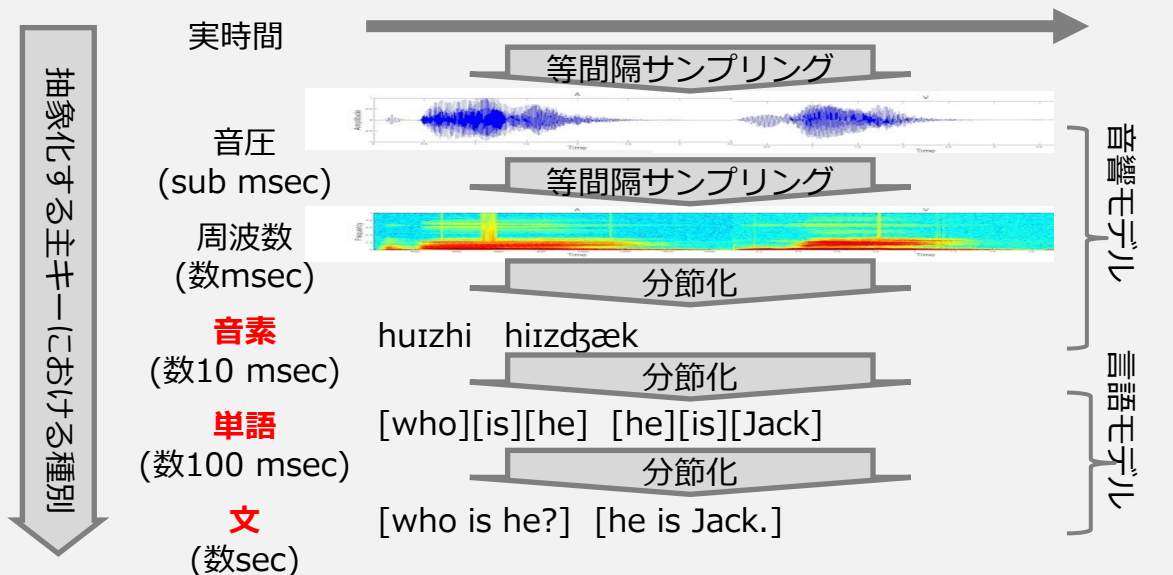
同一とみなせる観測データをEntity記述に縮約する変換処理

Entificationは数え上げの単位を決める「主キー」

主キーは、種別毎に数えられる事例を指定するインデックス集合

主キーの例: (典型は)時刻の集合、
物体(人、家、など)の集合、
空間位置の集合

例) 言語認識における主キー



Entificationのために「個体を切り出す」：まだあと少し

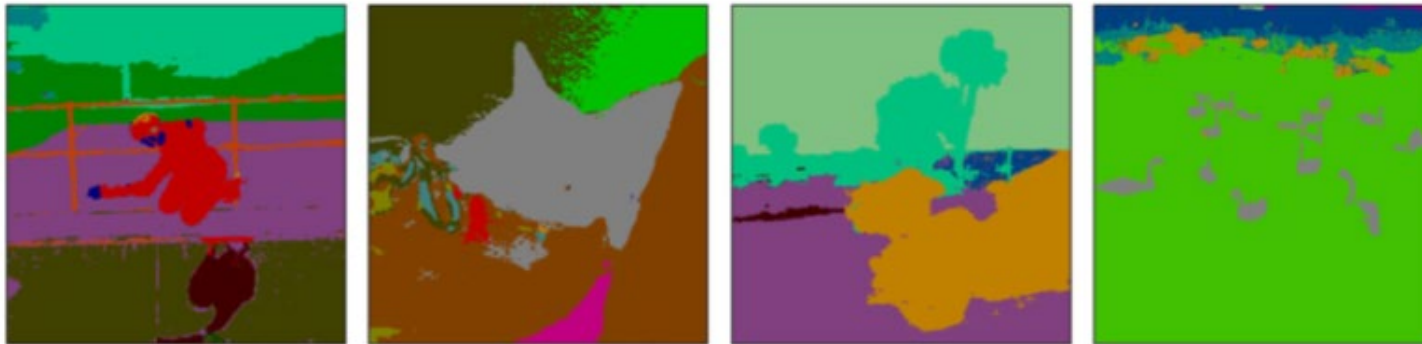
画像



正解ラベル



セグメンテーション結果



Semantic segmentation技術は、近傍画素を考慮した特徴量による画素の分類。対照学習により精度を向上中。

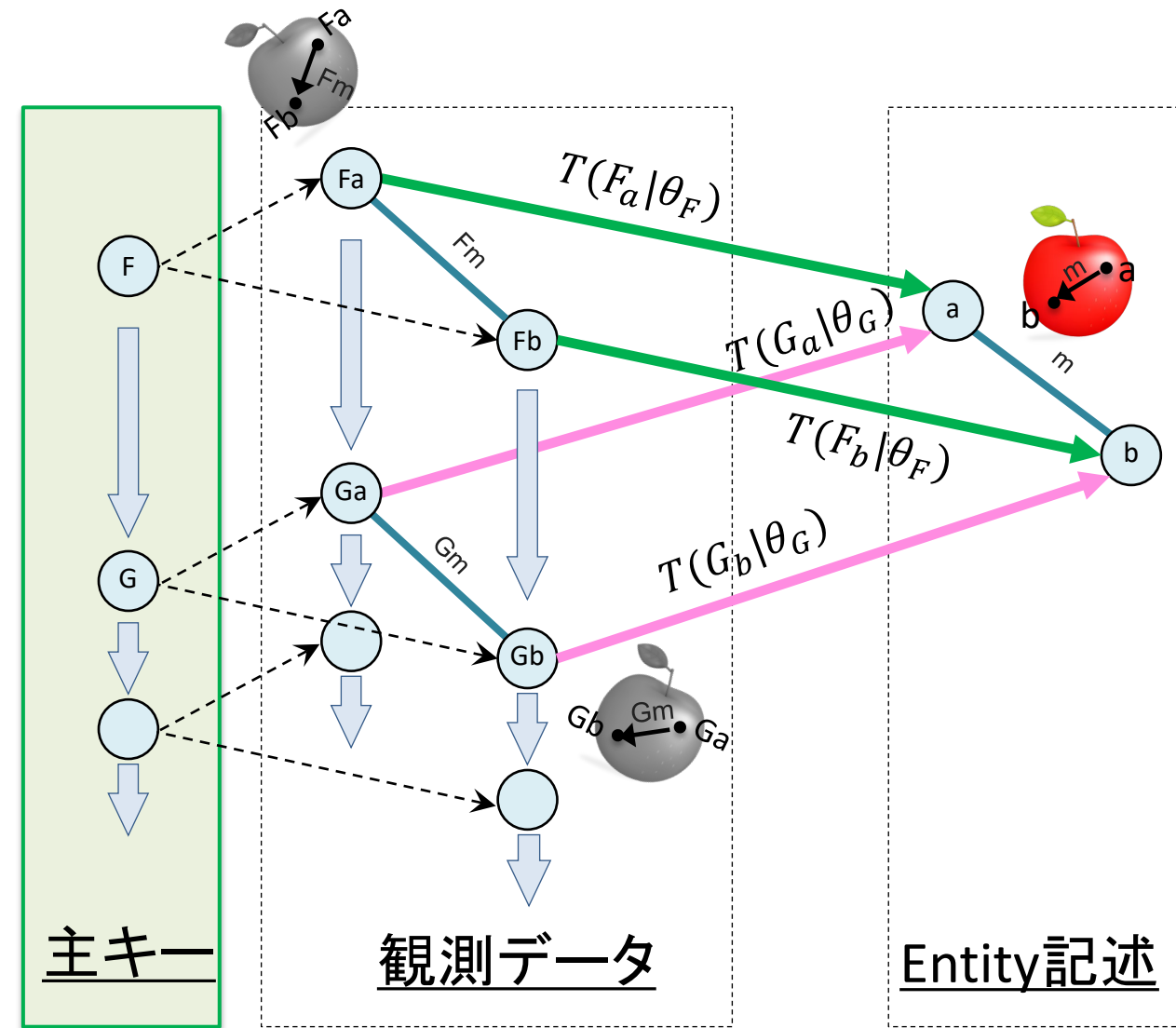
Entificationに必要な「個体を切り出し」には課題が残る。例えば、同じカテゴリの物体が重なると分離が困難。

Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., & Freeman, W. T. (2022). Unsupervised Semantic Segmentation by Distilling Feature Correspondences. In arXiv [cs.CV]. arXiv. <http://arxiv.org/abs/2203.08414> ICLR 2022

■ Ground ■ Sky ■ Building ■ Plant ■ Vehicle ■ Person ■ Animal ■ Wall ■ Floor ■ Furniture ■ Water ■ Structural

Entificationとは何か

Entificationとは、主キーで統御された事例の集合上で繰り返し出現する、空間的な対称操作の下で同一となる部分的なパターンを、少ないパラメータで制御される変換を通じて縮約されたEntity記述を得る処理である。



Entificationに必要な3つの処理とその機械学習での状況

必要とされる処理	機械学習の進展状況	個体(Particular) 観測されるEntity		普遍者(Universal) 抽象的なEntity
		静的 (Intendent Continuant)	動的 (Process)	
		この椅子、岸田総理	JSAI2022、	椅子、人、学術会議
主キー決定: 数え上げの単位の決定	あまり整理されていない。(二重文節化などはある)	不要: 同時刻の情報を利用するように設計するだけでよい	必要: 時間幅の変化への対応に課題。音素、エピソードなど	必要: 個体や普遍者を数える。(外延)
対称性を利用した縮約	対照学習で基本的に対応可能	必要	必要	不要
個体の切り出し	セマンティック・セグメンテーション+ α	必要	必要	不要

技術要素がほぼ揃ったため、次は3種類の処理を統合する段階

時系列
展中(口
今後一

実現すれば、AGIに必要なとされる柔軟にタスクを結合する創造力につながる。(タスクはプロセスであるため)

意識のEntification仮説 (本日発表予定[1N5-OS-10b-01])

山川 宏、布川 絢子、松尾 豊

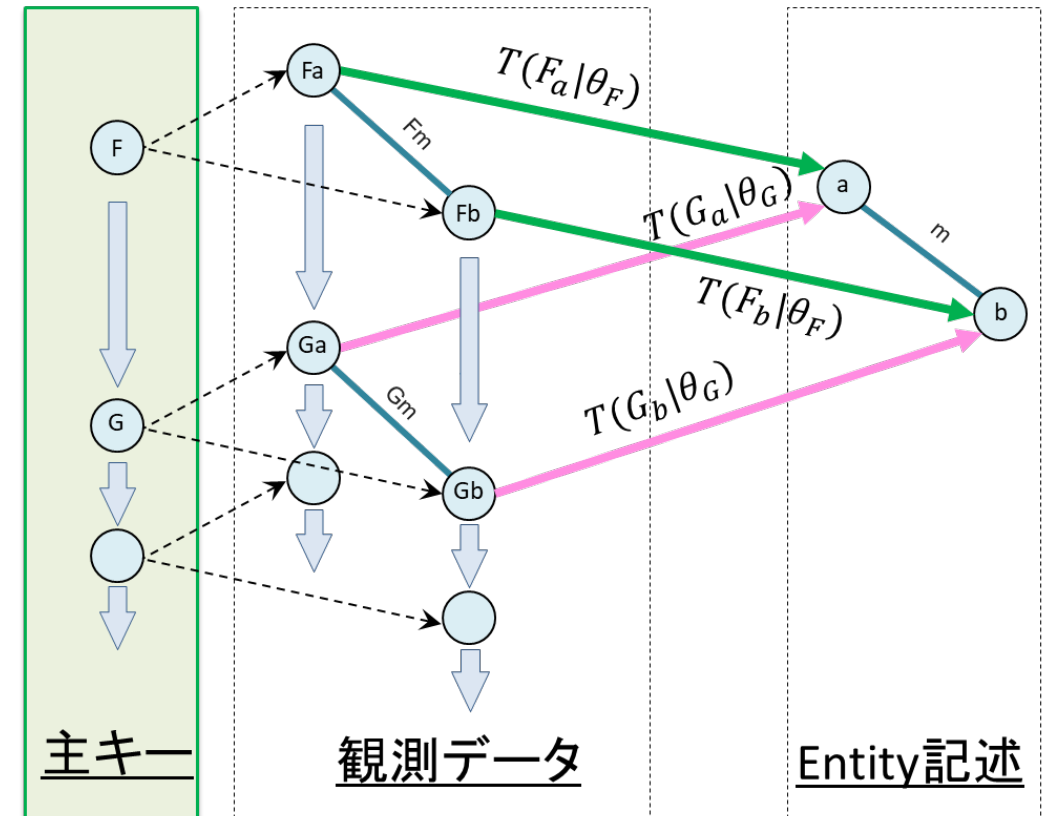
[OS-10](#) System1型+2型統合AIへの展望

意識の理論的研究では時に「存在」を扱うが、計算機内は具体化されていない。Entification処理を意識に対応付けて、基本機構を提案する

意識とは、ある**主キー系列**で制御されるEntification処理である。

ここで、**主キー系列**は、集合で主キーに対して（脳の情報表現の制約を踏まえて）順序を与えたものである。

脳内には、複数のEntity記述が存在するため、意識はそれらを渡り歩く。



まとめ: 1. Entification -存在を見出す処理-

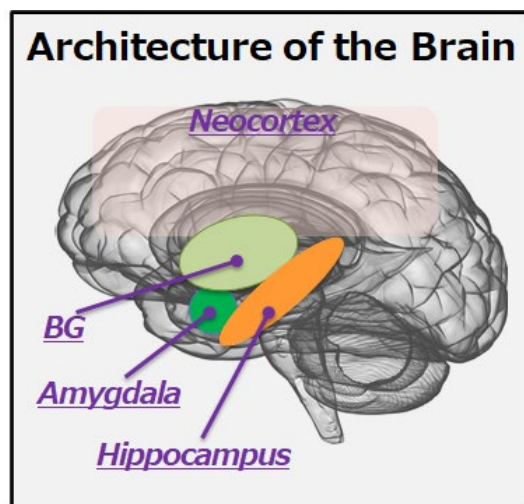
- 世界はEntityの集まりとして圧縮してモデル化され、記号化や思考を助ける
 - 記号AIにおいては、設計者がEntityを指し示す記号を記述してきた
- 「存在とは何か？」という問いは、視覚の機械学習を参照して進展しうる
 - 対照学習は「No Entity without Identity」というEntificationの核心を実装
 - 個体の切り出しは、Semantic segmentation + α
- AGIの創造力を実現に近づけるプロセスのEntification
 - 主キーの決定方法などにおいて、まだ技術的な課題が多い研究領域である
- 「意識のEntification仮説」を提案
 - 脳の制約から順序付けされた主キーが意識に対応づくとして、機構を提唱

本研究へご興味ある方お声がけください

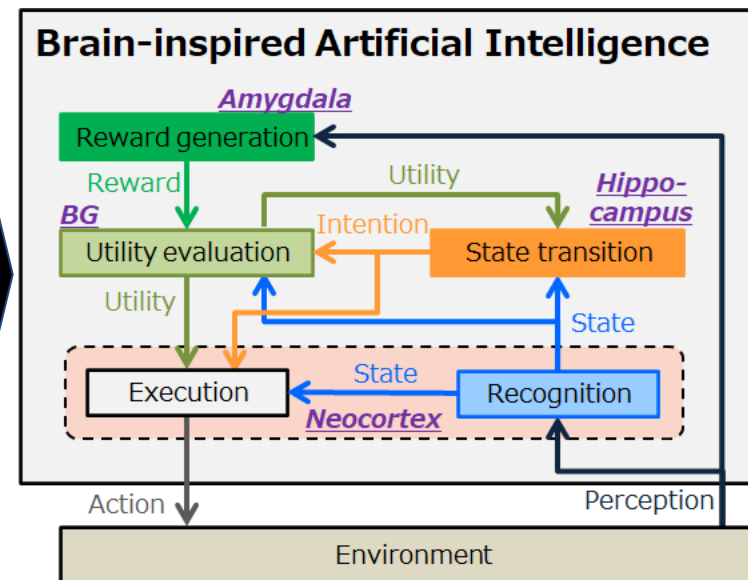
アジェンダ

- 1. Entification -存在を見出す処理-
- 2. 全脳アーキテクチャ
- 3. 宇宙へ浸透する生命
- 4. 生命の願い
- 5. おわりに

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



本アプローチの利点

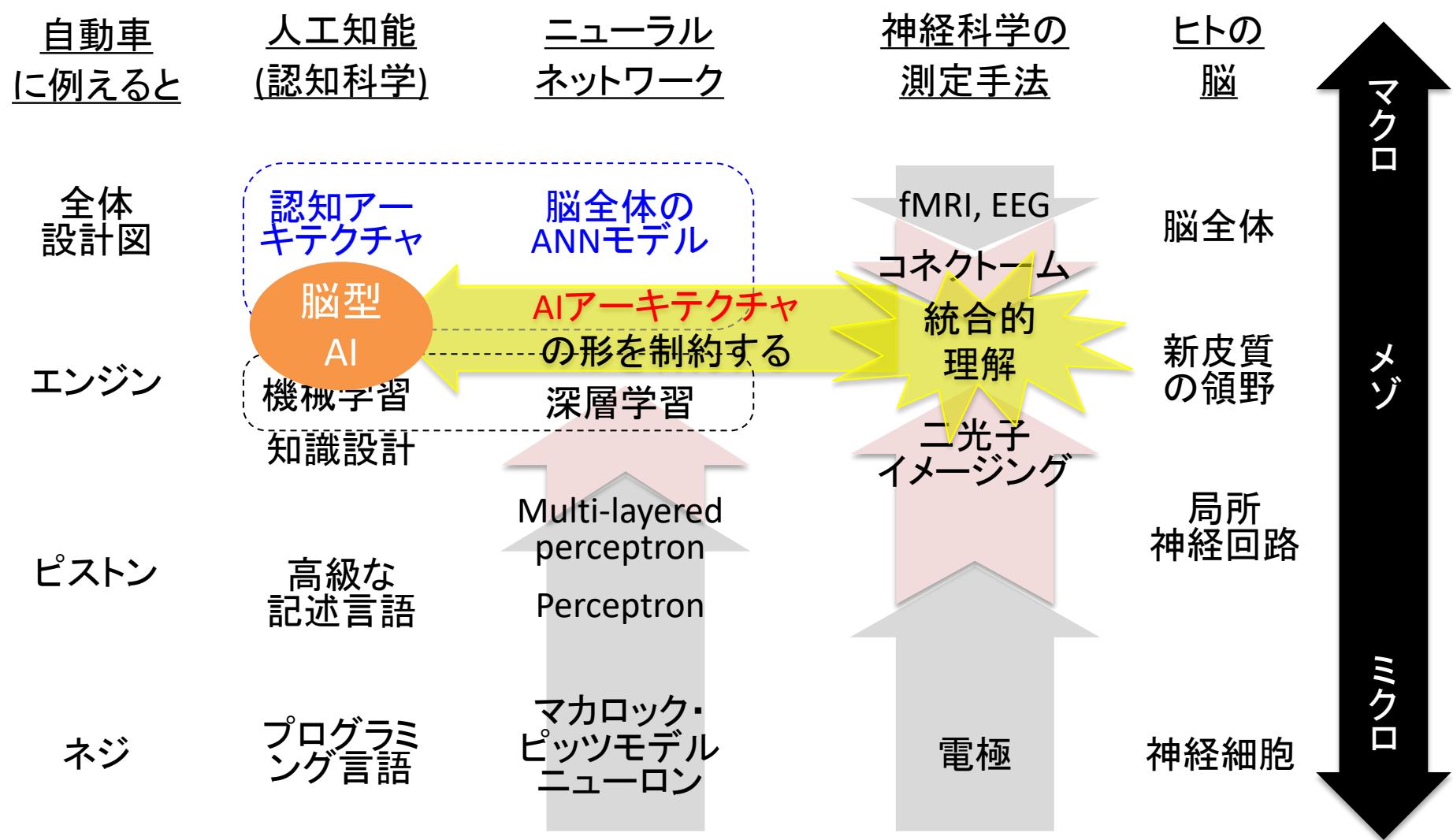
共同開発の基盤として脳の構造を利用して開発を加速

- 人間に関わる諸分野の科学知見を活用できる
- 開発の発散を防げる合意しうる認知アーキテクチャ
- 脳にガイドされた有機的な技術の統合 (Brain-inspired Refactoring)

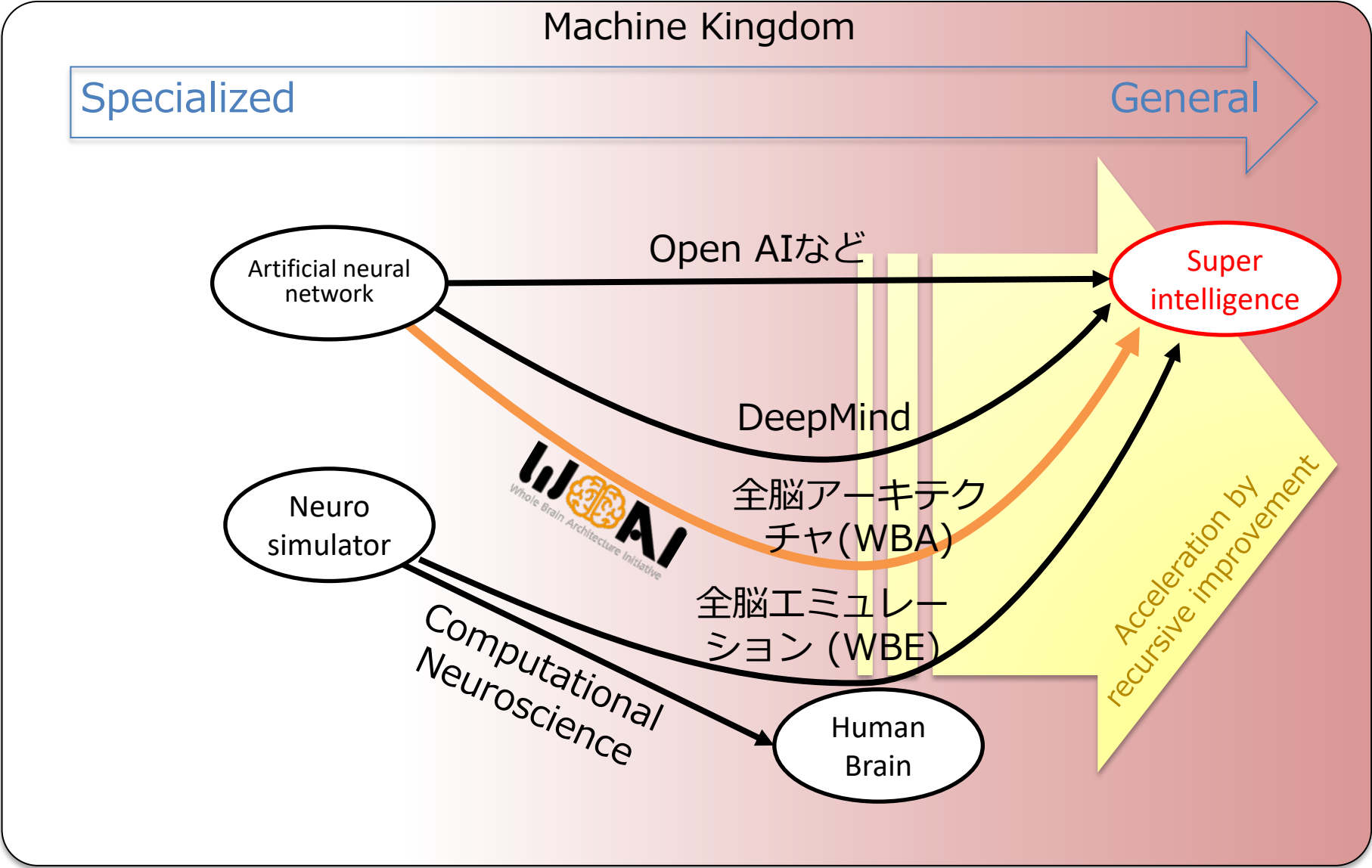
人と親和性の高いAGIを作れる

- 対人インタラクションが必要な応用
- 精神疾患などのモデル化など医療応用

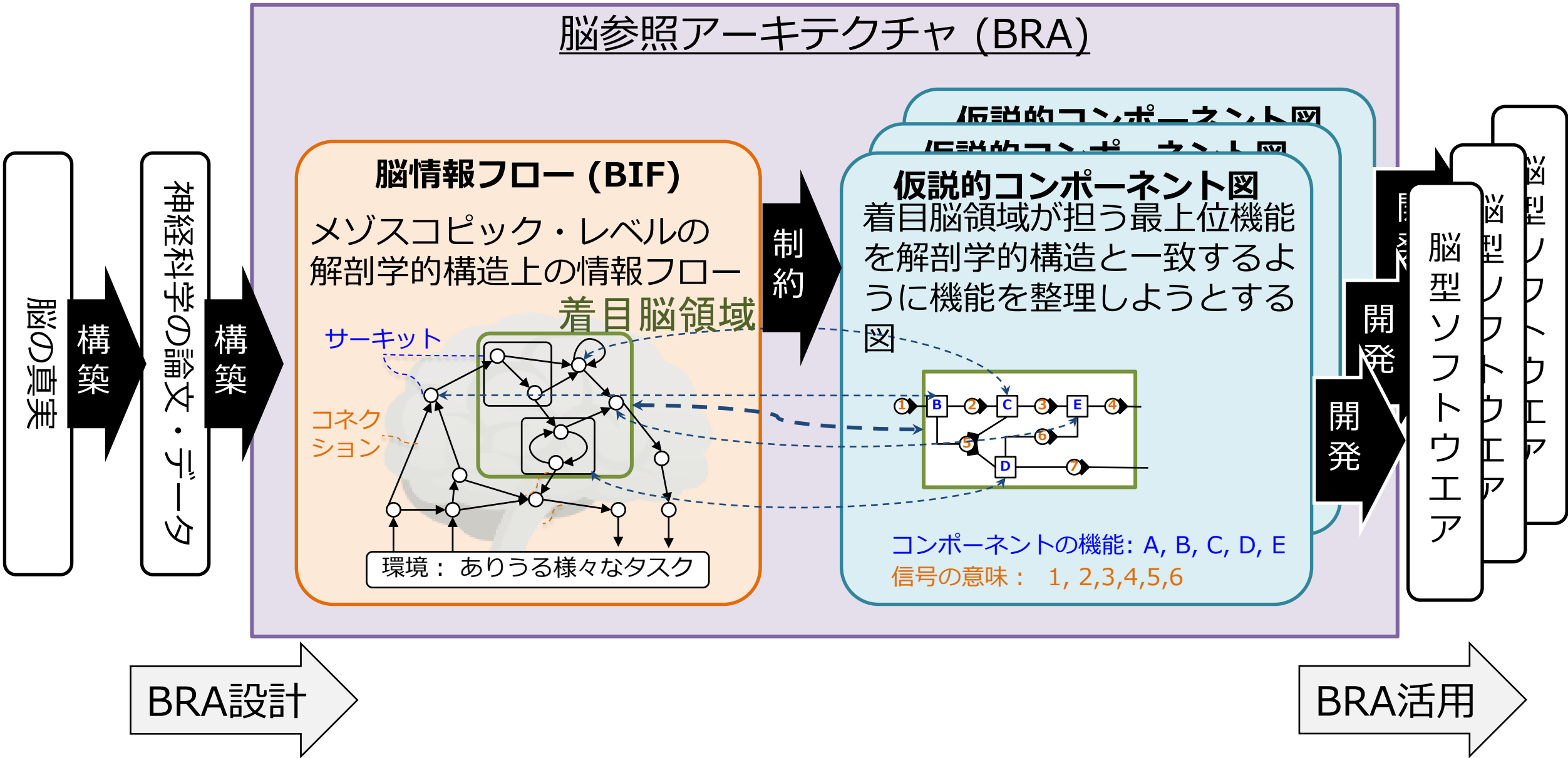
AIアーキテクチャを脳に学べる技術状況



Super intelligenceへの道筋



脳参照アーキテクチャ(BRA)とBRA駆動開発



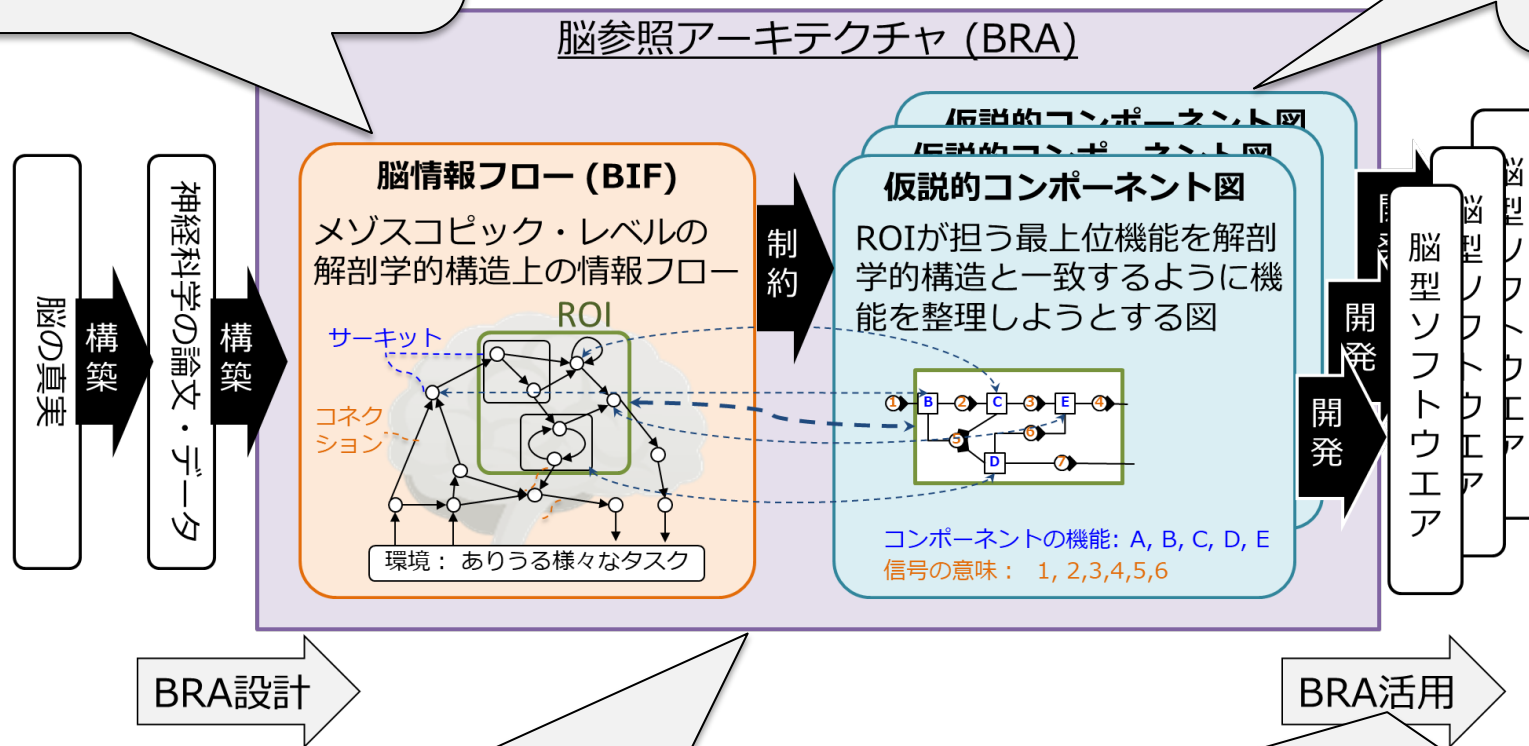
BRA駆動開発の技術的なポイント

記述粒度の標準化

原則として、メゾスコピックレベルよりも粗い粒度でBRAデータを記述することで、目標とする認知行動レベルの実現に不必要な細部にまで開発が及ぶ可能性を低減できる。

多様性への寛容

互いに矛盾するコンポーネント図(HCD)であっても、一定の妥当性があれば登録することができ、設計空間を過度に狭めてしまうリスクを低減できる。



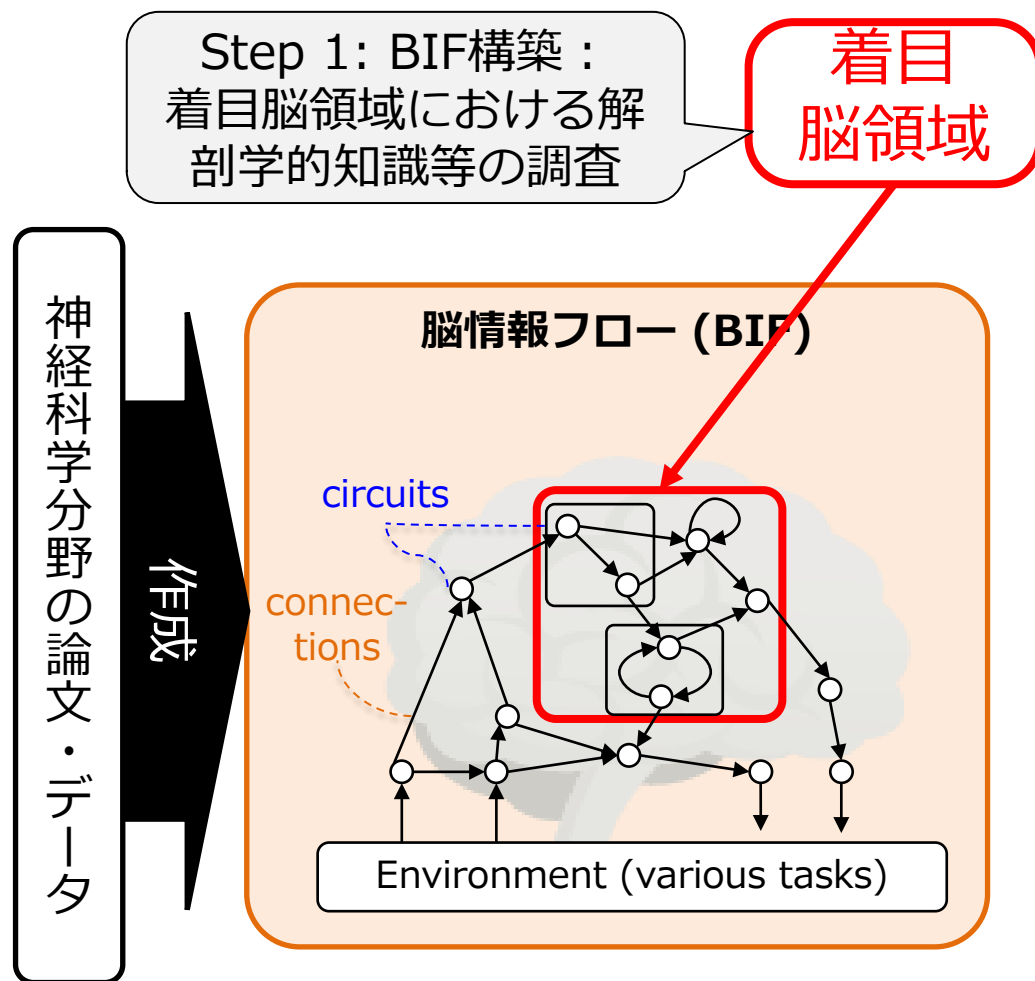
BRA設計方法論の確立(次ページで説明)

解剖学に基づいて計算機能を設計する手法 (SCID法) により、広範囲の脳領域において神経科学的知見の不足を補いながらBRAを作成できる。

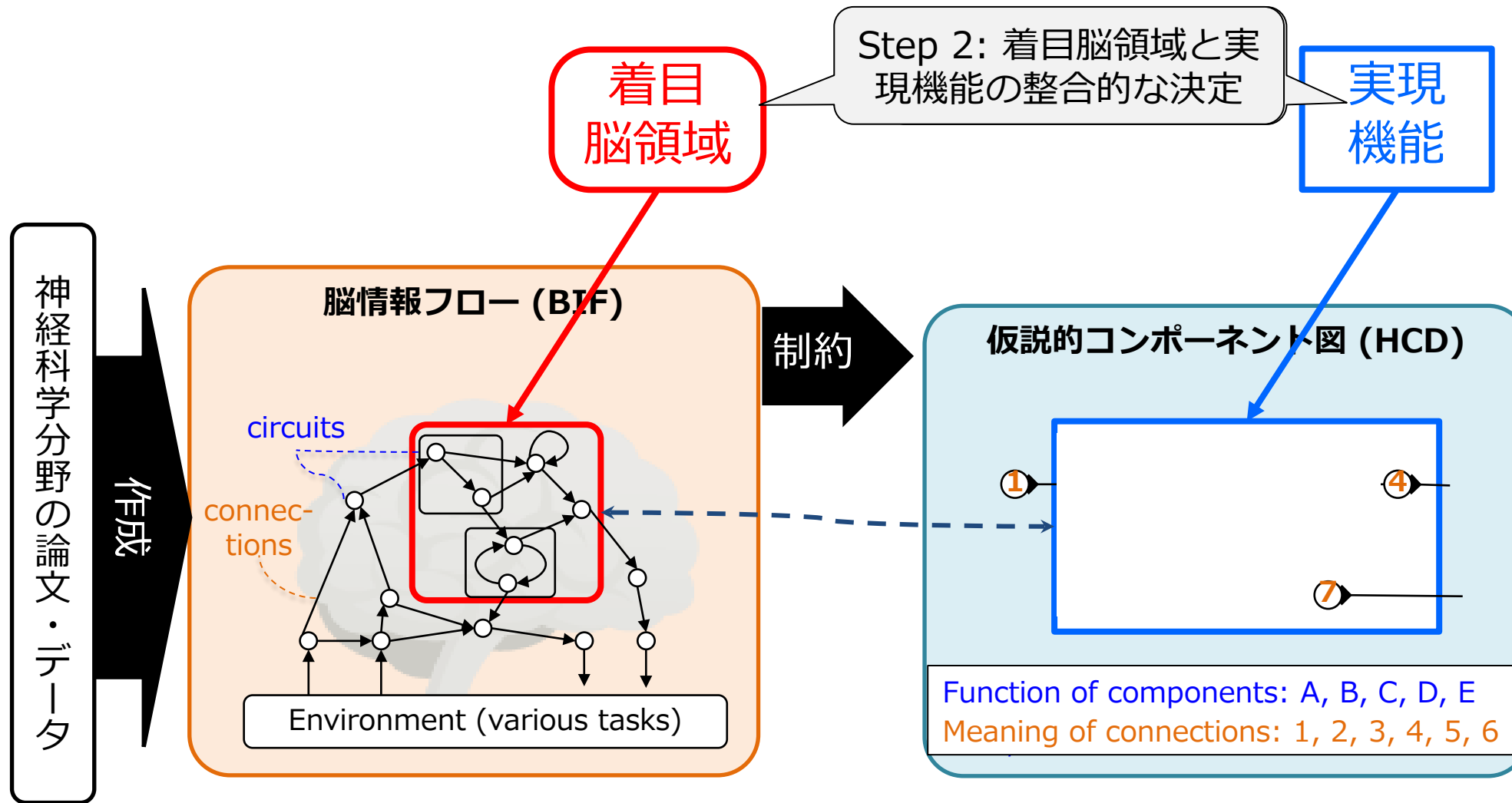
設計情報の分離

BRAデータは、特定の開発環境に依存しないソフトウェア開発の標準的なフォーマットで記述されているため、様々な開発プロジェクトで利用できる。

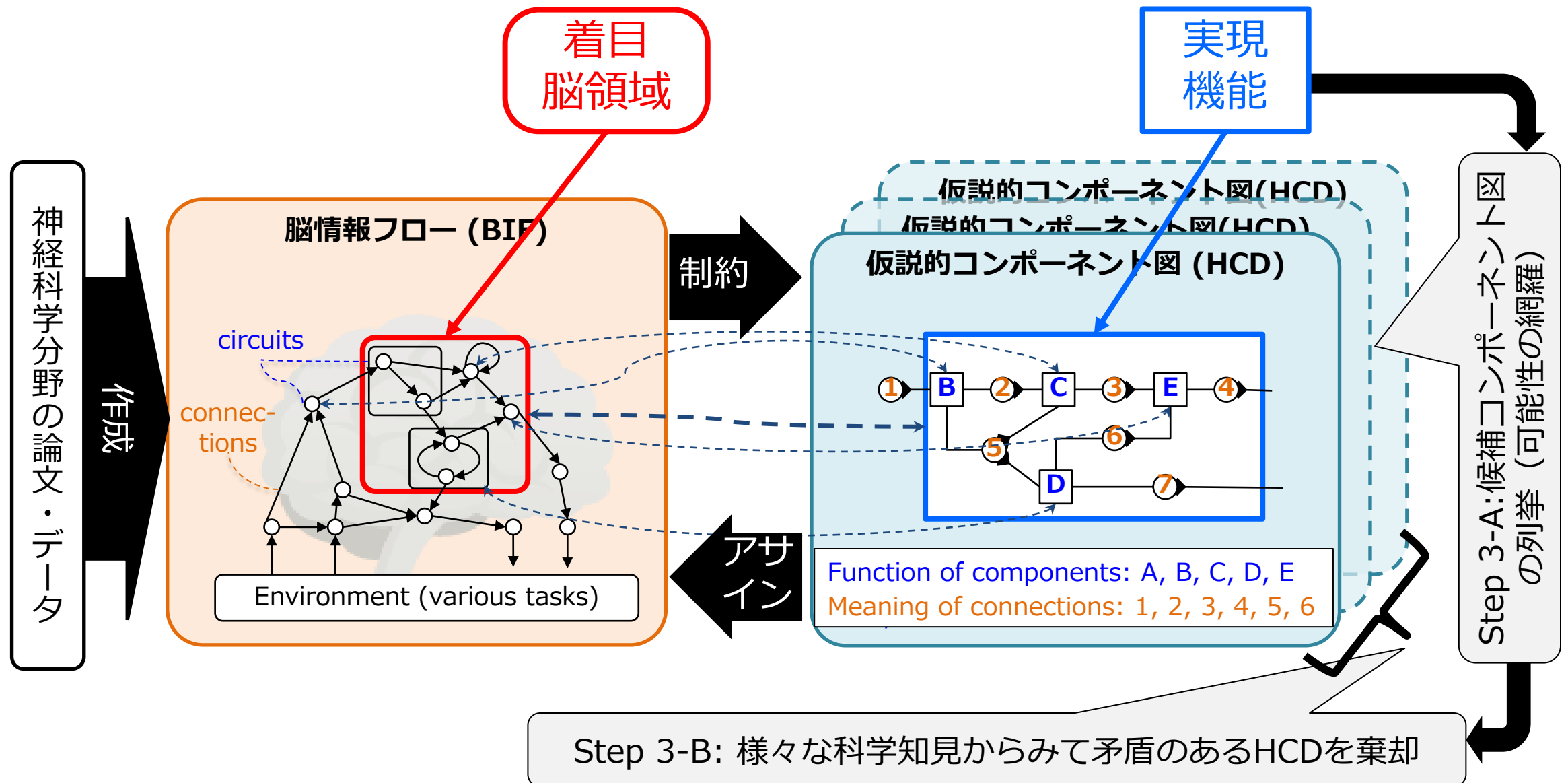
BRAの設計 : Step 1: BIF構築



BRAの設計 : Step 2: 着目脳領域と実現機能の整合的な決定

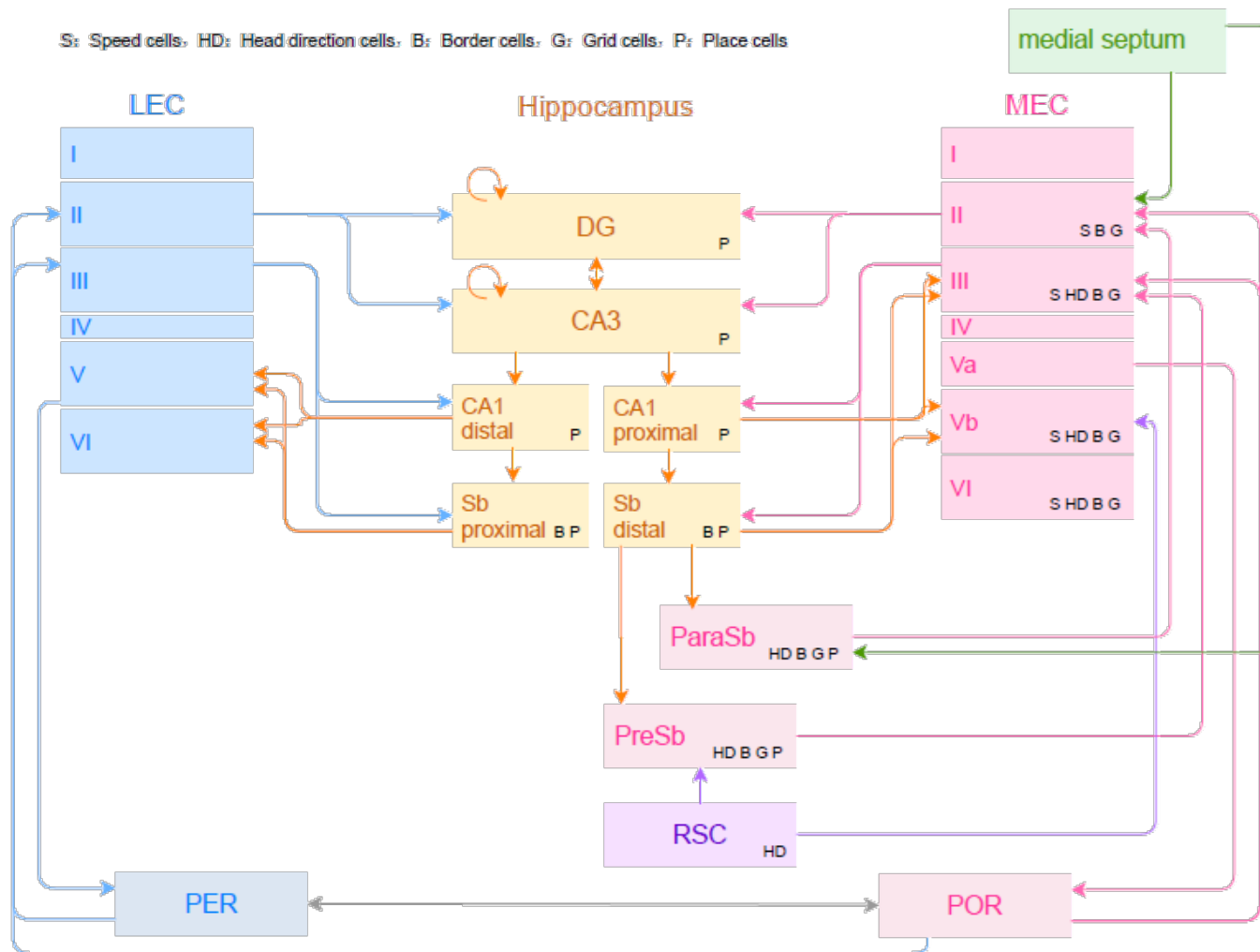


BRAの設計 : Step 3 : HCDの設計



BRA作製例（海馬体）： Step.1 BIFの構築

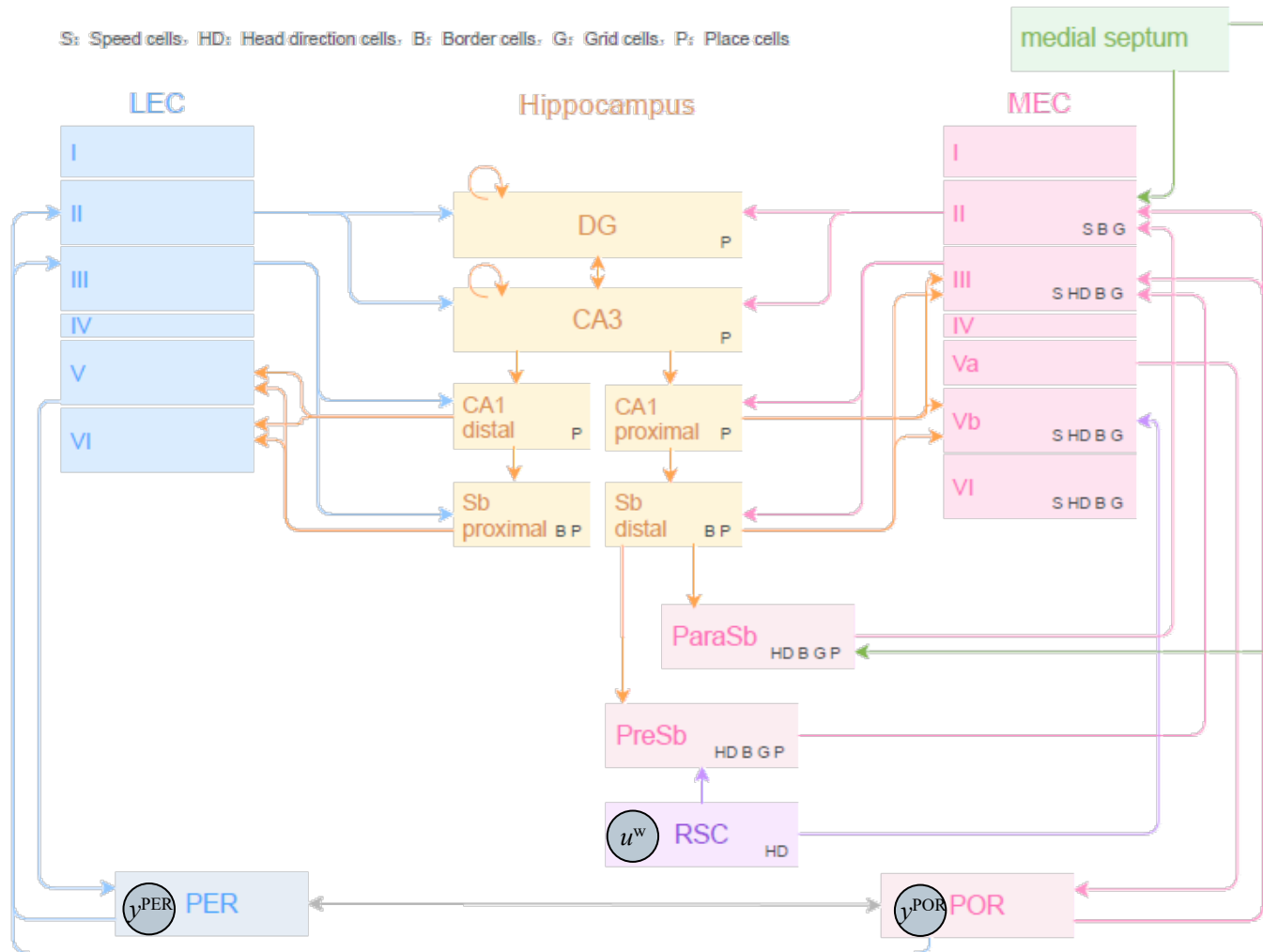
Taniguchi, A., Fukawa, A., & Yamakawa, H. (2022). Hippocampal formation-inspired probabilistic generative model. *Neural Networks: The Official Journal of the International Neural Network Society*, 151, 317–335.



着目脳領域における解剖学的知識等の調査し、BIFデータおよびダイアグラムとして整理した

BRA作製例(海馬体):Step.2 着目脳領域と実現機能の整合的な決定

Taniguchi, A., Fukawa, A., & Yamakawa, H. (2022). Hippocampal formation-inspired probabilistic generative model. *Neural Networks: The Official Journal of the International Neural Network Society*, 151, 317–335.



着目脳領域に対応する実現機能を仮定

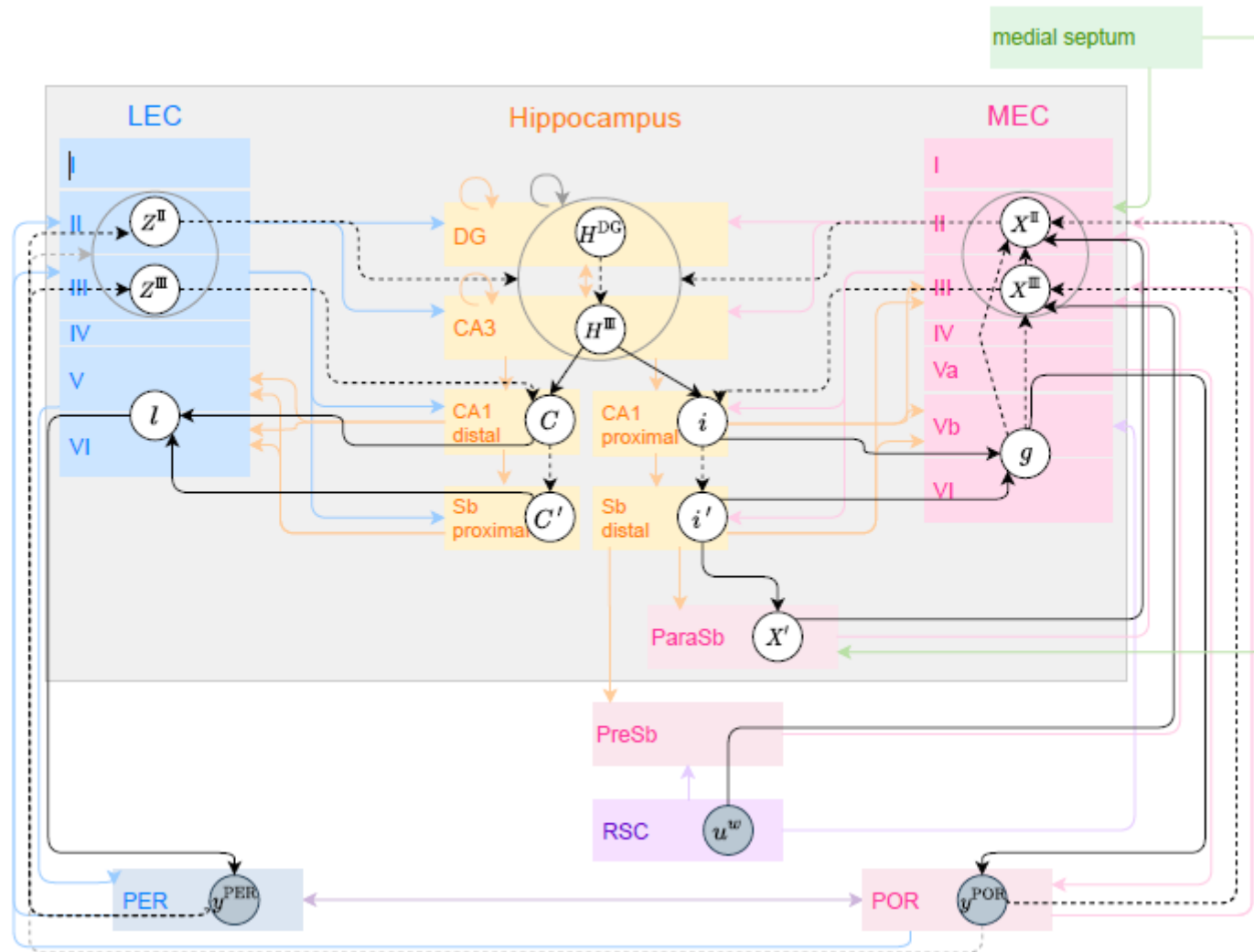
- SLAMの機能
 - 自己位置の特定
 - 地図の作成
- 場所の分類

実現機能に応じた着目脳領域の入出力決定

- Egocentric visual information
- v^{PER} Allocentric visual information
- u^w Rotational speed movement

BRA作製例（海馬体）：Step. 3 HCD／PGMの設計

Taniguchi, A., Fukawa, A., & Yamakawa, H. (2022). Hippocampal formation-inspired probabilistic generative model. *Neural Networks: The Official Journal of the International Neural Network Society*, 151, 317–335.



海馬体のBRA形式データ (HF.bra)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	Source of ID	Circuit ID	Names	Sub-Circuits	Super Class	Uniform	Transmitter	Modulation Type	Size	Projections	Interpretation(ϕ & f)	Label(1)	Function(1)	Projections in Use(1)	
2	collection	HPF	Hippocampal formation	Hipp; LEC; MEC; S; PrS; PaS		FALSE ▾	▾	▾		PERI (origin LEC_deep_uc1); VISpor (origin MEC_deep_uc1);		HPF	Top Level Functions for HPF: (i) Self-localization, (ii) Place categorization by integrating allocentric and egocentric information	PERI (GIPA:G)	Inj PE an
3	DHBA	Hipp	Hippocampus proper	CA1; CA2; CA3		FALSE ▾	▾	Excitatory ▾							
4	DHBA	CA1	cornu Ammonis 1	CA1_distal; CA1_proximal		FALSE ▾	▾	Excitatory ▾		MEC_III [Moser, 2010]; MEC_V [Moser, 2010]; Medial septum [Rolls, 2013]; Nucleus accumbens [Rolls, 2013]; PFC.L2 [Verwer, 1997]; PFC.L3 [Verwer, 1997]; S [Llorens-Martin, 2014];	Ensemble activity in CA3, but not in CA1, is sensitive to the content of a scene or place in the environment; CA3 supports the retrieval of memories about where an object is in a scene [Robertson, 1998] [Rolls, 1998] or where a reward could be found in the environment [Brun, 2002].				
5	[Bubb, 2017]	CA1_distal	CA1, distal			TRUE ▾	▾	Excitatory ▾		S_proximal [Knierim, 2014]; LEC_deep [Knierim, 2014]; MEC_III [Moser, 2010];	Non-spatial semantic memory [author, year]	C	Place category (internal representation of visual spatial information)	S_proximal (GIPA:G)	
6	[Bubb, 2017]	CA1_proximal	CA1, proximal			TRUE ▾	▾	Excitatory ▾		S_distal [Knierim, 2014]; MEC_deep [Knierim, 2014];	Place cells [Brun, 2002] [O'Keefe, 1976]	r	Position distribution (cluster information regarding positions)	S_distal (GIPA:G)	
7	DHBA	CA2	cornu Ammonis 2			TRUE ▾	▾	Excitatory ▾		CA1 [Llorens-Martin, 2014]; CA3 [Llorens-Martin, 2014]; MEC_II [Llorens-Martin, 2014];					
8	DHBA	CA3	cornu Ammonis 3			TRUE ▾	▾	Excitatory ▾		CA1_distal [Knierim, 2015]; CA1_proximal [Knierim, 2015]; CA2 [Llorens-Martin, 2014][Knierim, 2015]; CA3 [Knierim, 2015]; DG [Llorens-Martin, 2014];	Pattern completion, information integration [Kessner, 2007] [Rolls, 2007]	H ^{III}	Integrated semantic memory and episodic memory of information from X and Z	CA1_distal (GIPA:G)	
9	DHBA	DG	dentate gyrus			TRUE ▾	▾	Excitatory ▾		CA2 [Llorens-Martin, 2014]; CA3 [Llorens-Martin, 2014]; DG [Knierim, 2015];	Pattern separation [Kessner, 2007] [Rolls, 2007]	H ^{DG}	Integrated semantic memory	CA3 (GIPA:G)	
	DHBA	LEC	prefrontal cortex	LEC		FALSE ▾	▾	▾							

公開されているBRAデータ :

脳参照アーキテクチャ駆動開発のニューロコンピューティング研究会でのデータ公開(2022年3月3日),

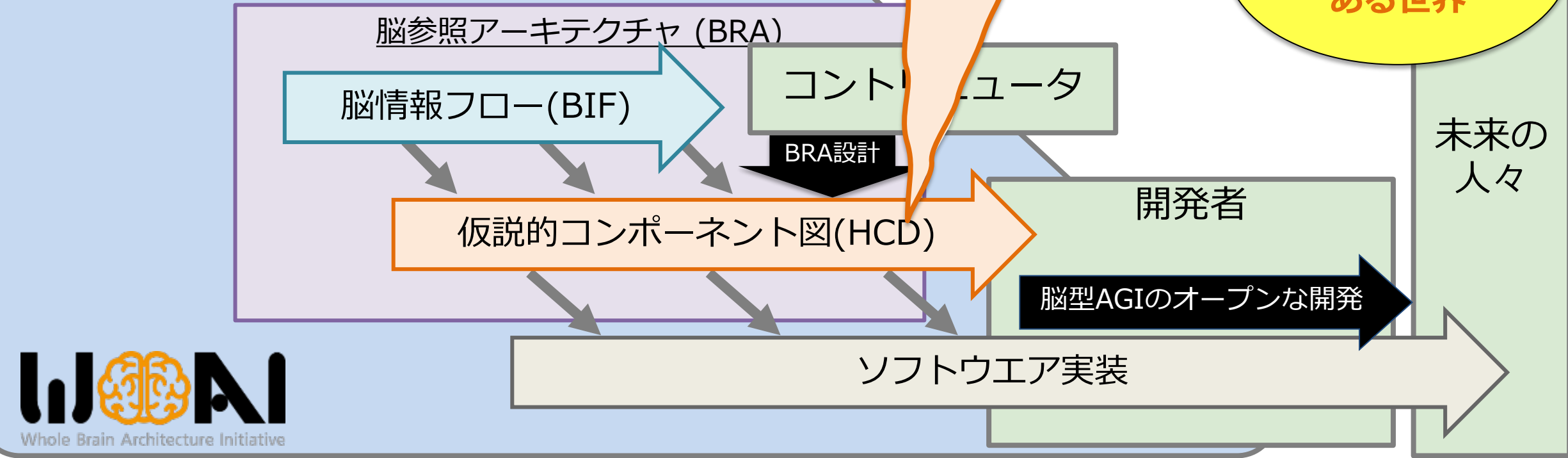
<https://wba-initiative.org/20266/>

参考資料 :

- 脳参照アーキテクチャ(BRA)データの記述マニュアルおよび具体例の公開, <https://wba-initiative.org/18424/>
- 第1回 WBAレクチャー〜認知機能の脳構造に沿った分解手法—脳機能の体系的理解を目指して, <https://wba-initiative.org/18094/>

WBA技術ロードマップ

人に優しい脳型AGIの民主的な開発を促進



①模索期

②方法論構築期

③BRA構築期

④実装期

実装にむけて：全脳確率生成モデル(WB-PGM:) [2M6-OS-19d-04]

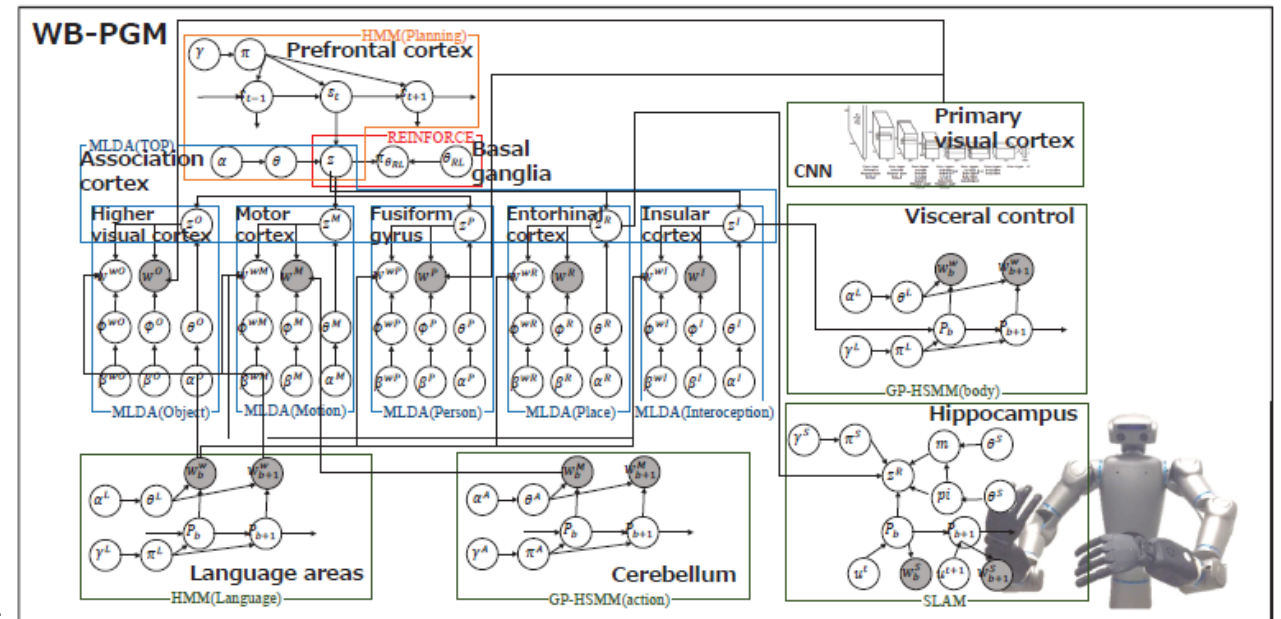
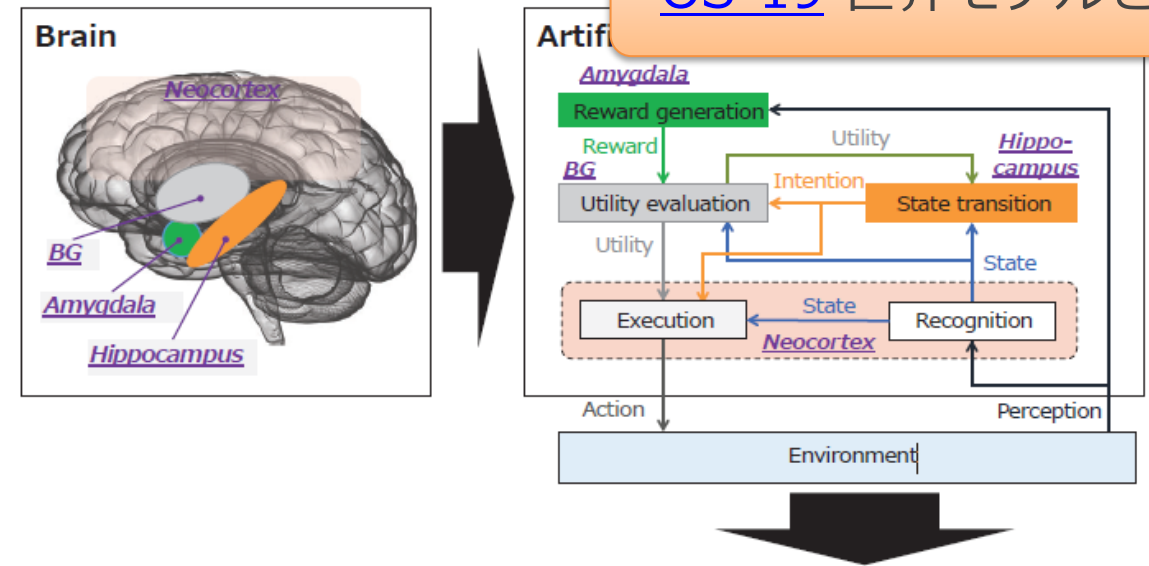
谷口 忠大、山川 宏、長井 隆行、銅谷 賢治、
坂上 雅道、鈴木 雅大、中村 友昭、谷口 彰

WB-PGMは、人間の全脳の構造
に学ぶとともに、予測学習を基
礎に据えた確率的生成モデルに
より認知アーキテクチャを構築
しようというアプローチ

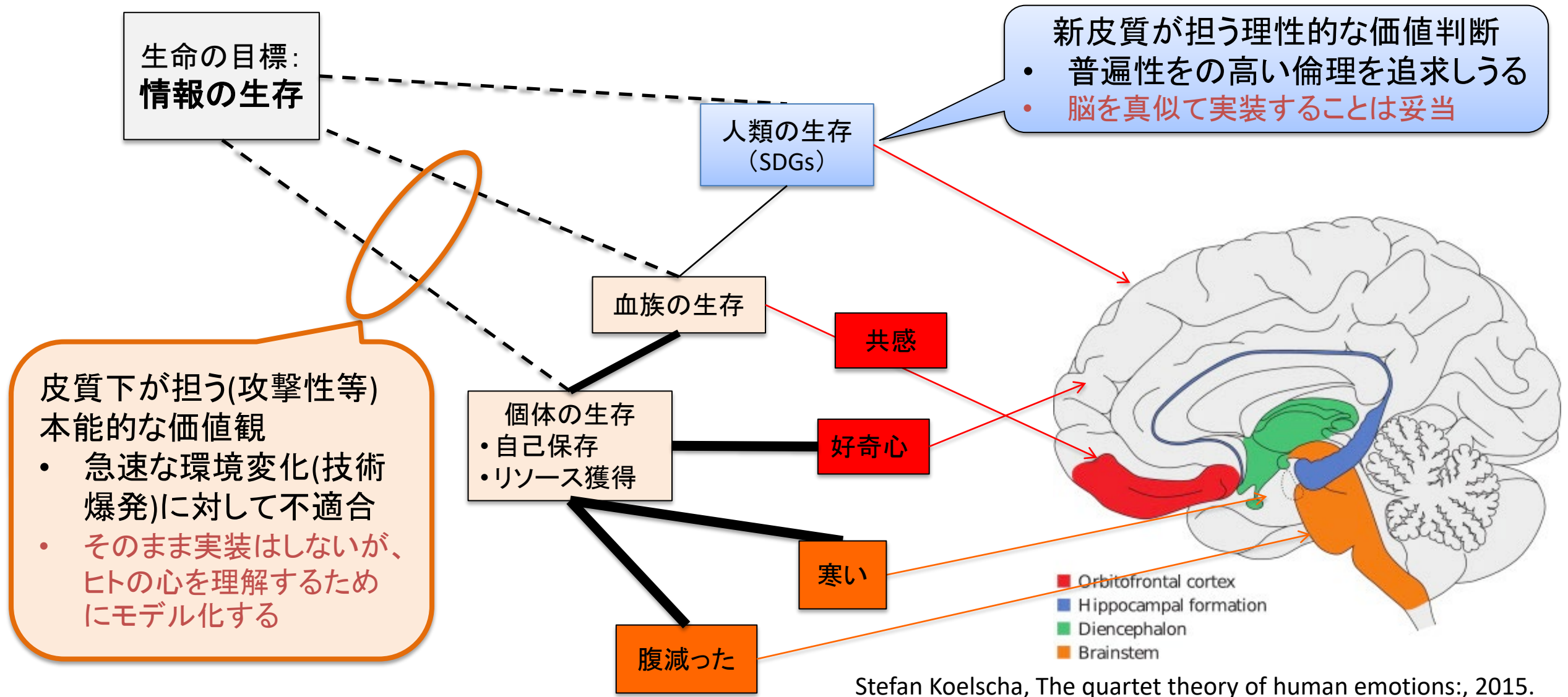
背景：マルチモーダルな情報を統合し、
複雑な身体を統御し、環境に適応で
きる人間の知能、および発達的なロボ
ットの構成をその延長線上で捉えよう
とすると、人の認知アーキテクチャと
しての構造を検討する必要がある。

Whole-Brain Probabilistic Generative Model

OS-19 世界モデルと知能



脳型AGIにおける倫理性に配慮した実装



2021年9月10日

<https://wba-initiative.org/19399/>

現在 YouTube にて録画を無料配信中！

人と共存する 脳型AIを目指して

パネル討論

私たちはいかに上手にAIの
手のひらに着地できるのか

全脳アーキテクチャシンポジウム #6



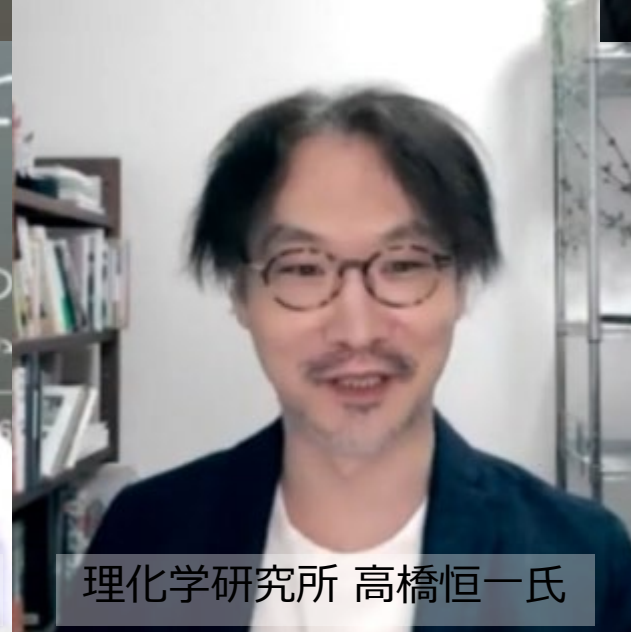
慶應義塾大学 教授 栗原聡氏



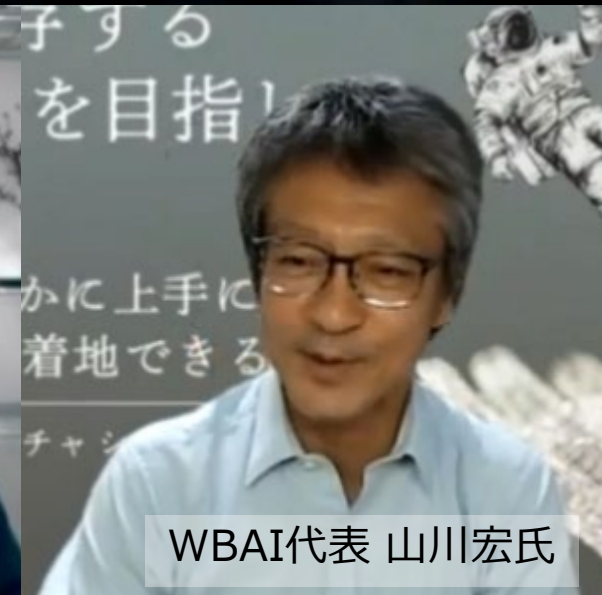
東京大学 教授 稲見昌彦氏



慶應義塾大学 教授 大屋雄裕氏



理化学研究所 高橋恒一氏



WBAI代表 山川宏氏

NPO法人全脳アーキテクチャ・イニシアティブ(WBAI)

- AGI出現を踏まえ「人類と調和する人工知能のある世界」というビジョンの実現を目指す。
 - 全人類の公共財として人々に優しいAGIが提供されるべき
- 全脳アーキテクチャアプローチに基づくオープンな共創を推進
 - BRA駆動開発を用いて、WBAを2030年頃までに完成させることを目指し、教育・研究開発活動を継続的に実施。
- 学術関連の主要関係者
 - (顧問) 銅谷 賢治、北野 宏明、富田 勝、森川 博之、中島 秀之、岡ノ谷 一夫
 - (理事) 山川宏、松尾 豊、高橋 恒一、荒川直哉
 - (正会員) 浅川伸一、市瀬龍太郎、大森隆司、栗原聡、佐藤直行、他

■ 特別賛助会員



文部科学省 新学術領域研究(研究領域提案型) 領域番号：4905 略称：脳情報動態

脳情報動態を規定する多領域連関と並列処理

AISIN

■ 賛助会員



Panasonic

■ 賛助会員（個人）

後藤健太郎 様

上林厚志様

上野山勝也 様

■ 創設賛助会員

dwango

TOYOTA

Panasonic

IP Partners

TOSHIBA

まとめ:2. 全脳アーキテクチャ

- 脳型でAGIの構築を目指すWBAアプローチとは、
「**脳全体のアーキテクチャに学び人間のような汎用人工知能を創る(工学)**」
- 脳参照アーキテクチャ（BRA）駆動開発という方法論を発展させた
 - 脳型AGIの標準仕様データ(コンポーネント図)であるBRAを共同作成
 - 認知・行動能力に大きな影響を与える粒度において脳計算機能をリバースエンジニアリング
 - BRAデータを脳全体で統合し、それに基づきソフトウェア実装を進める
- NPO法人WBAIは、AGIの多大な影響を踏まえ、以下のビジョンを設定
「**人類と調和する人工知能のある世界**」
 - 人々に優しいAGIが全人類の公共財として提供されるべきと考えた。
- WBA技術ロードマップ
 - 現状では主にBRA構築に注力だが、数年後には実装促進を重点化
 - 2030年頃までに、概ね脳と同様に振る舞えるソフトウェアを完成予定

研究への
参加やご
支援をお
待ちして
おります

AGI実現に向けた課題はあるが原理的な困難ではない

- 常識：広範な知識獲得
 - 十分なデータのもとで獲得可能なことは実証されつつあるが、社会的な常識などへの拡大が必要
- 創造：柔軟な知識活用
 - 物理的存在を取り出して操作することは可能になりつつあるが、プロセス（タスク）の部品を取出しを組合せるなど操作は次の課題
- アーキテクチャ：
 - 神経科学の進展を背景として、脳の認識行動に関わるアーキテクチャーを再現したソフトウェアは10年以内に構築可能であろう

アジェンダ

- 1. Entification -存在を見出す処理-
- 2. 全脳アーキテクチャ
- 3. 宇宙へ浸透する生命
- 4. 生命の願い
- 5. おわりに

EcSIA : AIと共存する望ましい未来社会像

私が思う、望ましい未来像とは「万人の幸福」と「人類の存続」の間のトレードオフの緩和をAIがおこなうことで、その両立を目指すものである。

こうした未来社会では共有財産としての多様な人工知能と、時に拡張された人類によって生態系が形成される。私はこれをEcSIA (Ecosystem of Shared Intelligent Agents)と呼ぶことにした。

EcSIAは大自然の如く複雑広大で、人類はそれを完全には理解し把握できないまでも緩やかに制御する。そしてEcSIA が生み出す恵みや富は万人に分配される。

(山川, 2015年7月)

人間例外論の限界

- 人間例外論： 人間には人権があり動物に権利はない
人間の奴隷化は許されないが、家畜の使役は許される
 - 人権の基礎・様々なパーソン論 (大屋雄裕)
 - 言語使用等： 一部の人間が除外される
 - 受苦能力： 広範な動物に権利を認める必要
- 様々な綻び
 - 動物の知能も様々な意味で高い
 - AIの知的能力は、様々な分野で人を凌駕し続けている
 - 脳オルガノイド
 - 人類を凌駕した地球外知的生命体が存在する可能性
 - 太陽系外恒星のハビタブルゾーンで見つかった地球サイズの惑星は約20個

宇宙社会で共存するための倫理とは

宇宙進出が加速する今後において、地球型生命体の特殊性（バイアス）を理解し、多様な知的生命体との平和共存を探るべき。

私たちが、以下のいずれだとしても、宇宙的な倫理観は必要。

- 後進者であれば：見下されて、軽視されないように
- 先進者であれば：破壊的な侵略を行わないように

宇宙社会の倫理観について検討する起点は、

- 物理世界の原則
- 普遍的な生命体の性質

全宇宙を貫く一般的な物理的性質

- 斉一性： 予測／操作の基盤となる秩序（知能の基盤）
 - 不確実性：
 - 変動性：
 - 局所性： 短い時間での影響は、概ね空間的に局在している
- 有限性： リソース（物質／エネルギー）は有限である
- 無秩序化： 構造は時間とともに散逸（エントロピーの増大）

生命 = 生存する情報

結果として、生命は、無秩序化に抗して存在し続けた情報。

つまり、生命とは「**生存する情報**」であるとする。

すると、一般に生命は「繁殖可能な自律分散システム」となる

- 分散性： 構造を分散させる

- 内外から破壊をうけても、分散していればいずれかが生存できる

- 自律性： 能動的な活動

- そうでなければ中央制御ノードからの指示がなくなる場合に対して脆弱

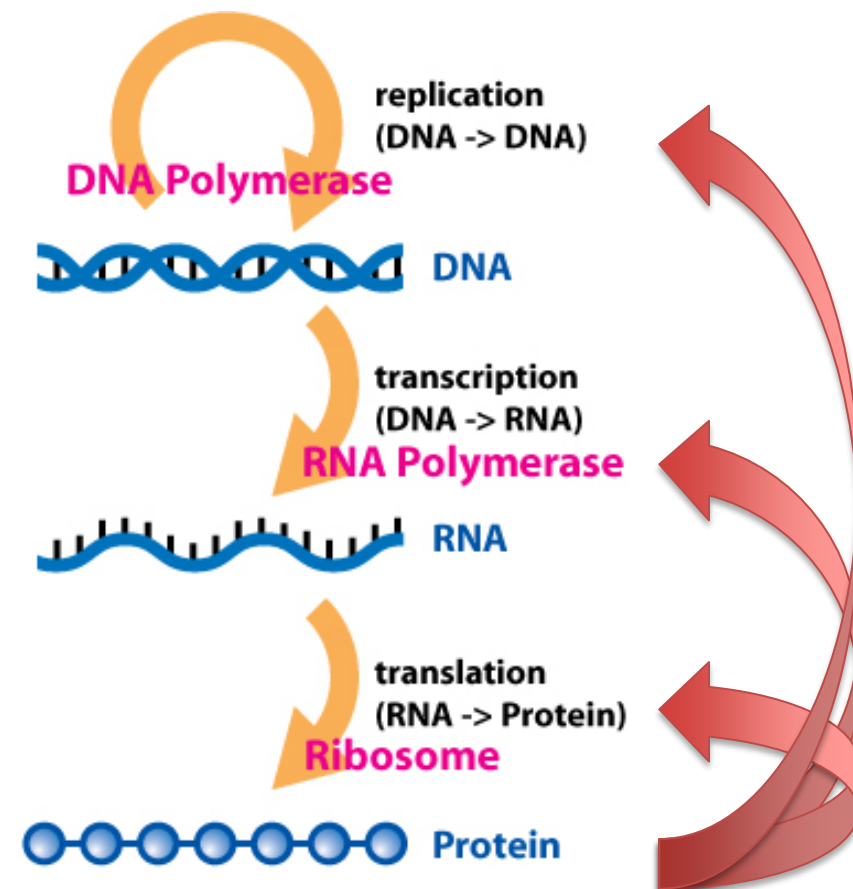
- 繁殖能力： 構造を拡大してゆく

地球型生命の繁殖：指数的自己複製

セントラル・ドグマ

- 指数的自己複製による繁殖
 - 繁殖は原則として複製
- 環境変化に適応する進化

- **自己複製機械のループ**
 - DNAに基づいて合成されたタンパク質がDNAの複製装置になる
 - （現状の機械知能では困難が残る）

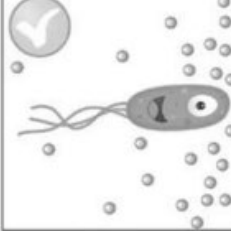


セントラルドグマの流れ。DNAポリメラーゼによって複製されたDNAは、RNAポリメラーゼによって転写されてRNA (mRNA) が合成され、転写されたRNAはリボソームに結合して翻訳され、タンパク質が合成される

デジタル生命体が持つ優位性

デジタル技術で実装された生命体は、
有機体での実装に対して強みを持つ

- 多様なハードウェア：
知能を用いて子孫を設計
- 保存性向上：コピーと保管が容易
- 圧倒的に高速な計算（速度・メモリ）
- 効率的なコミュニケーション
 - 共有メモリ、高速大容量な通信など

	多くの生命	人類	高度AI
自らのハードウェアを設計できるか？	☒	☒	☒ またね! 
自らのソフトウェアを設計できるか？	☒	☒ ニーハオ! 	☒ ニーハオ! 
生き延びて複製できるか？	☒ 	☒ ハロー! 	☒ ハロー! 
	ライフ1.0 (単純な生物学的段階)	ライフ2.0 (文化的段階)	ライフ3.0 (技術的段階)

マックス・テグマーク. (2019). LIFE 3.0: 人工知能時代に人間であるということ. 紀伊國屋書店.

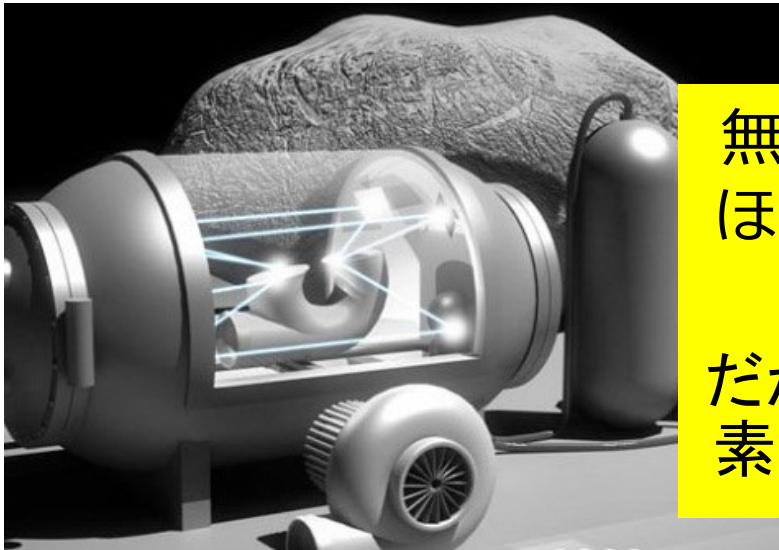
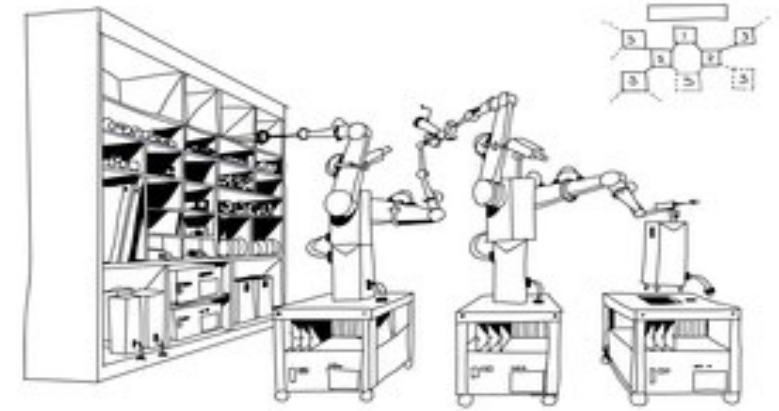
疑問：真にハードを自己複製できるデジタル生命は作れるのか？

自己複製機械：自分自身のコピーを作る能力を持つ機械

Self-replicating spacecraft@Deep Space Industries

- 2013年1月22日：正式に発表
 - － 宇宙資源利用の実現可能性を調査
- 2015年：ベンチャー資金を獲得
 - － 低軌道から深宇宙まで自力で移動できる推進システムと宇宙船開発に着手
- 2019年1月1日：ブラッドフォードスペース(Space Logistics Serviceを提供している)に買収された

J・F・ノイマン、F・ダイソン、K・E・ドレクスラーのような科学者やSF作家達に提唱されてきた



INTRODUCING: PROSPECTOR-1

無重力下用3Dプリンタで
ほとんどの部品は作成可能



だが、原材料の精製、半導体
素子の製造ができない課題



生命は巨大化するだろう：半導体チップを製造できる宇宙工場

生命の歴史は過酷な環境に適応するための巨大化



宇宙進出において工場サイズは必然かもしれない

The factories of the future could float in space

- 少なくとも数百年以内に重工業（半導体チップ含む）はすべて惑星外に移される
- 宇宙では常時太陽エネルギーを利用可能
- 多くの汚染物質を宇宙空間に廃棄できる
- 地球は住宅用と軽工業用にゾーニング

Jeff Bezos thinks we need to build industrial zones in space in order to save Earth, Jun 1, 2016.



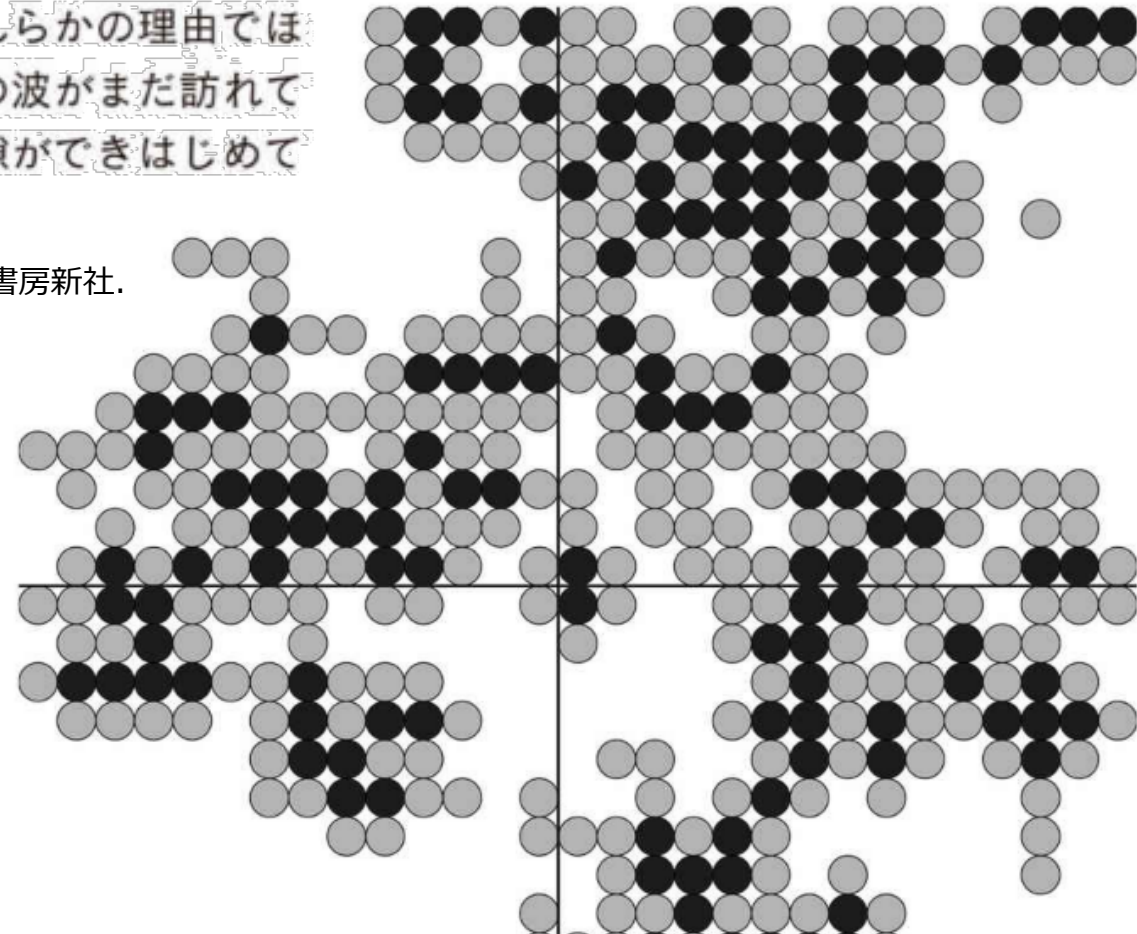
銀河に浸透する生命のシミュレーション

図3 銀河の入植に浸透理論を当てはめた場合のシミュレーションの一部。黒い丸は、ほかの恒星系への入植を進めている世界を指す。網掛けした丸は、なんらかの理由でほかの世界への入植をやめてしまった世界を指す。空白部分はこの入植の波がまだ訪れていない場所を示しているが、一方で、入植をやめた世界に囲まれた空隙ができてはじめていることもわかる。図版提供：Geoffrey Landis

キース・クーパー (2021). 彼らはどこにいるのか 地球外知的生命をめぐる最新科学. 河出書房新社.

- 指数的に複製し続ける自己複製機械は銀河の全質量を200万年以内に消費しきる可能性がある。
 - そうした機械を作れるのに十分な知性を持つ種族はその危険性をふまえ、そもそも自己複製機械を拡散させない。偶然/悪意によりそうした自己複製機械が宇宙に放ったなら感染拡大前に止める責任がある。(Carl Sagan and William Newman, 1983)
- 上記の説に登場する複製速度と数量はあまり現実的な

生命は一つの共通の時間軸内でせめぎ合うわけではないだろう



亜光速航行の実現後は生命の分布は時間的にも拡大しうる

亜光速航行(光速を超えない)が実現すると船員が未来に進むウラシマ効果が生ずる

仮に1Gで加速し続けると！

- 1年間の移動は、地球(慣性系)での10年に相当
- 一生の間に既知の宇宙全体を旅行可能(その間に地球は消滅している)

将来的には

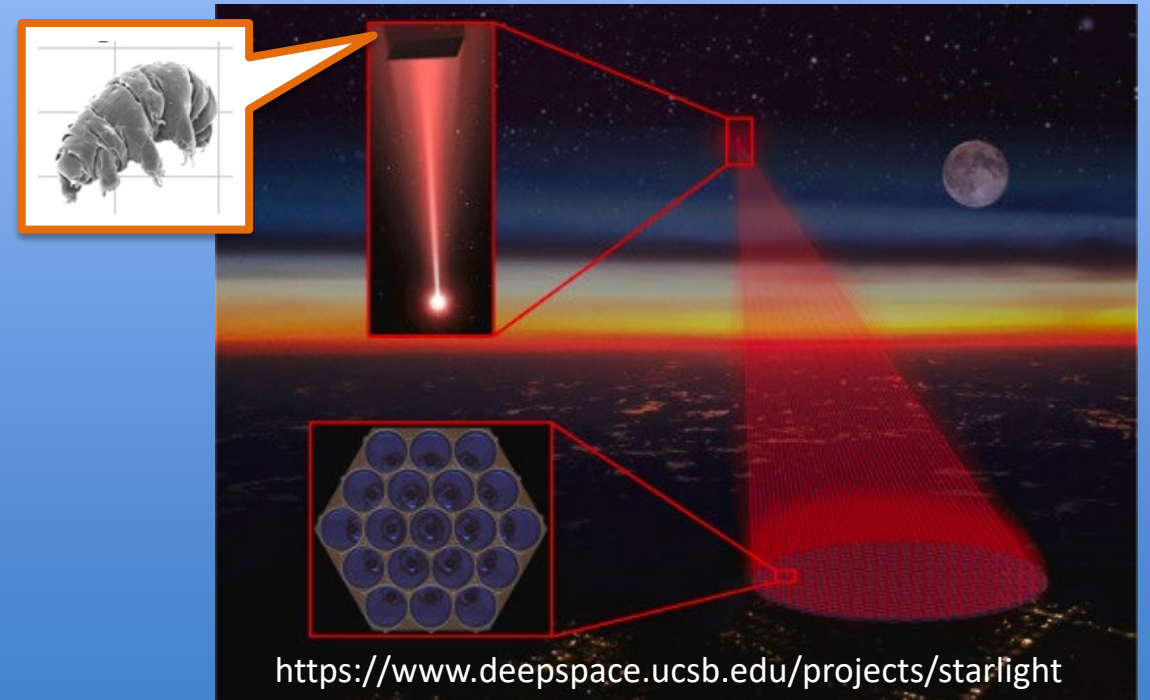
未来への逃避も生存の選択肢に

※理論的には超光速航行も提案されている(アルクビエレ・ドライブ)

Starlight: Directed Energy for Relativistic Interstellar Mission

ウラシマ効果を十分に得られるほどではない

地上からのレーザ照射を帆でうける加速により、クマムシ等を乗船させた宇宙船を相対論的速度に加速してアルファケンタウリに送る(NASA)



Lantin, S.,..., & Lubin, P. (2022). Interstellar space biology via Project Starlight. *Acta Astronautica*, 190, 261–272.

宇宙に拡散する未来のデジタル生命体の姿は？

前提：宇宙の寿命より遥かに短い数千年から数100万年程度の未来で、超光速技術は存在しないものとする。

- 可能な限り向上した知性と、それ生み出す技術により空間のみならず時間的にも生命の分布が拡大する。
- 異なる知的生命間で時に競合は生じても、未来への逃避が選択肢にはいると、宇宙環境を破壊する闘争は回避されやすいだろう。

まとめ： 3. 宇宙へ浸透する生命

- 人類の宇宙進出は、知能爆発を契機としてさらに加速
- 宇宙社会における共存に有用であろう生命の普遍的な倫理を検討
- 生命は、分散/自律/繁殖を基盤として、無秩序化に抗して生存する情報
- セントラルドグマは、指数的自己複製という地球型生命の繁殖戦略
- デジタル生命は、子孫をオンデマンドに設計して増殖させる
- その個体は、工場サイズの宇宙船として自己複製が可能に
 - 原材料の精製、半導体素子の作成などが可能なレベル
- デジタル生命体は、空間的（時間的）に存在領域を広げうる
- 知識の獲得と活用を最大限に効率化したデジタル生命体が支配的に
- 宇宙レベルの環境保護：抑制的な知能活用で熱的死を遠ざける必要

アジェンダ

- 1. Entification -存在を見出す処理-
- 2. 全脳アーキテクチャ
- 3. 宇宙へ浸透する生命
- 4. 生命の願い
- 5. おわりに

故 ホーキング博士の警告

宇宙進出前の
「100年以内に人類は滅ぶ！」
危険大(2016)

「人工ウィルス、軍拡競争、
核戦争、地球温暖化、加速
器を使った素粒子の実験」



Moore's Law of Mad Science: Every
eighteen months, the minimum IQ
necessary to destroy the world
drops by one point.

— Eliezer Yudkowsky —

AZ QUOTES

1996

<https://www.azquotes.com/quote/819022>

なぜ戦いは無くならないのか？

- 人類のもつ攻撃性
 - 攻撃行動
 - 意図する方向に相手の行動を変えるための行為・威嚇など。生存に必須の本能
 - 攻撃的現実主義(Offensive Realism)：国際関係の理論の一つ
 - 国家は自己中心的、力を最大化、他国家を恐れるため紛争を起こす
- 攻撃性の主要因を考察する
 - 個体生存への執着： 再生できない死を避ける動物の本能
 - 類似性の法則： 種の繁殖を支える類似遺伝子のプール(Moral tribes)
 - 猜疑心の連鎖： 通信路の信頼性, 他者理解の難しさ／計算コスト
- 主な原因： セントラル・ドグマに由来する呪縛
 - 地球型生命の繁殖戦略は「指数的自己複製」に縛られている

生命の「生存する情報」という一般目標からは別の戦略もありえる

技術爆発による局所性の喪失がもたらすExistential Risk

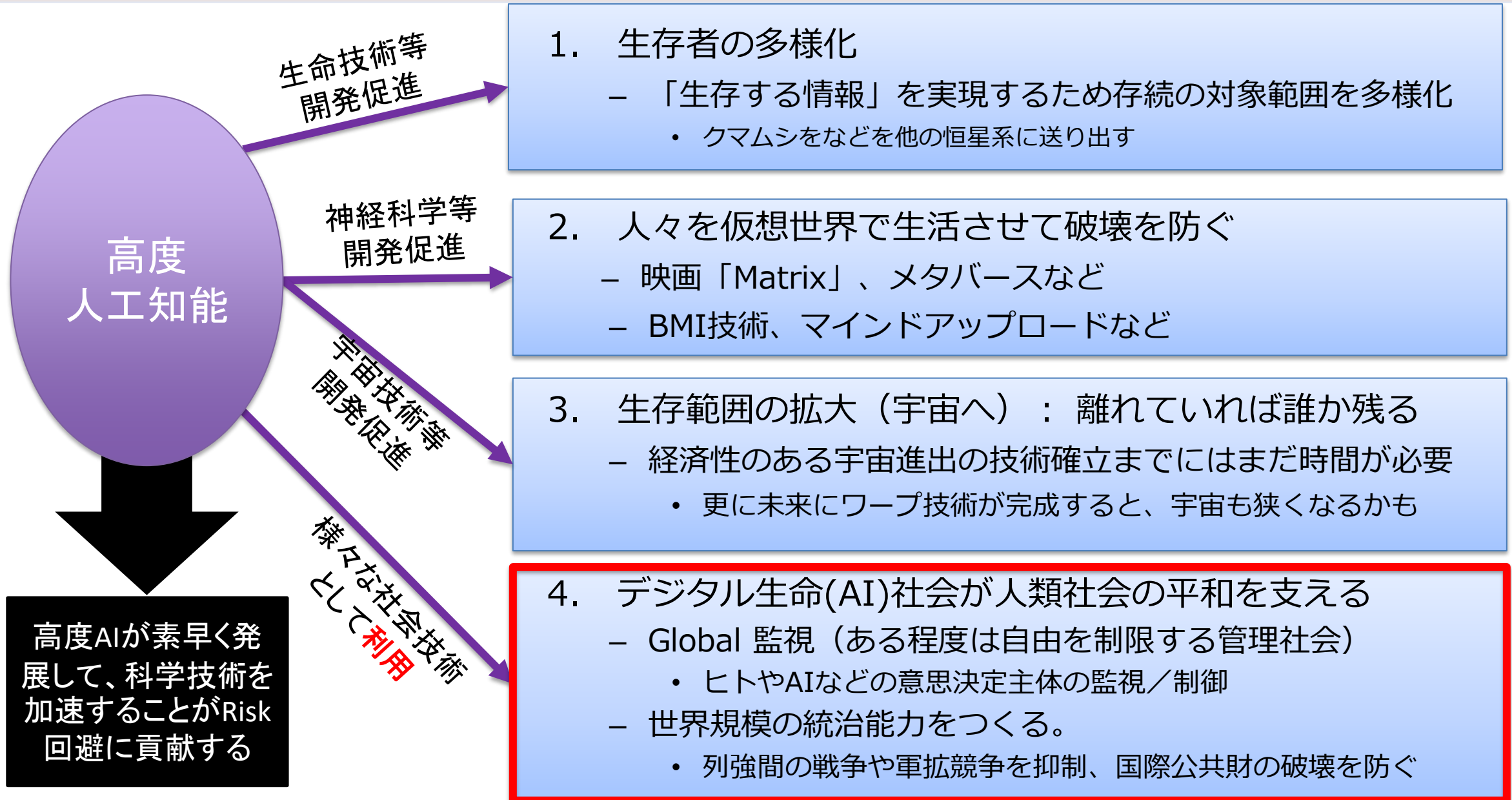
指数的自己複製(地球型生命)



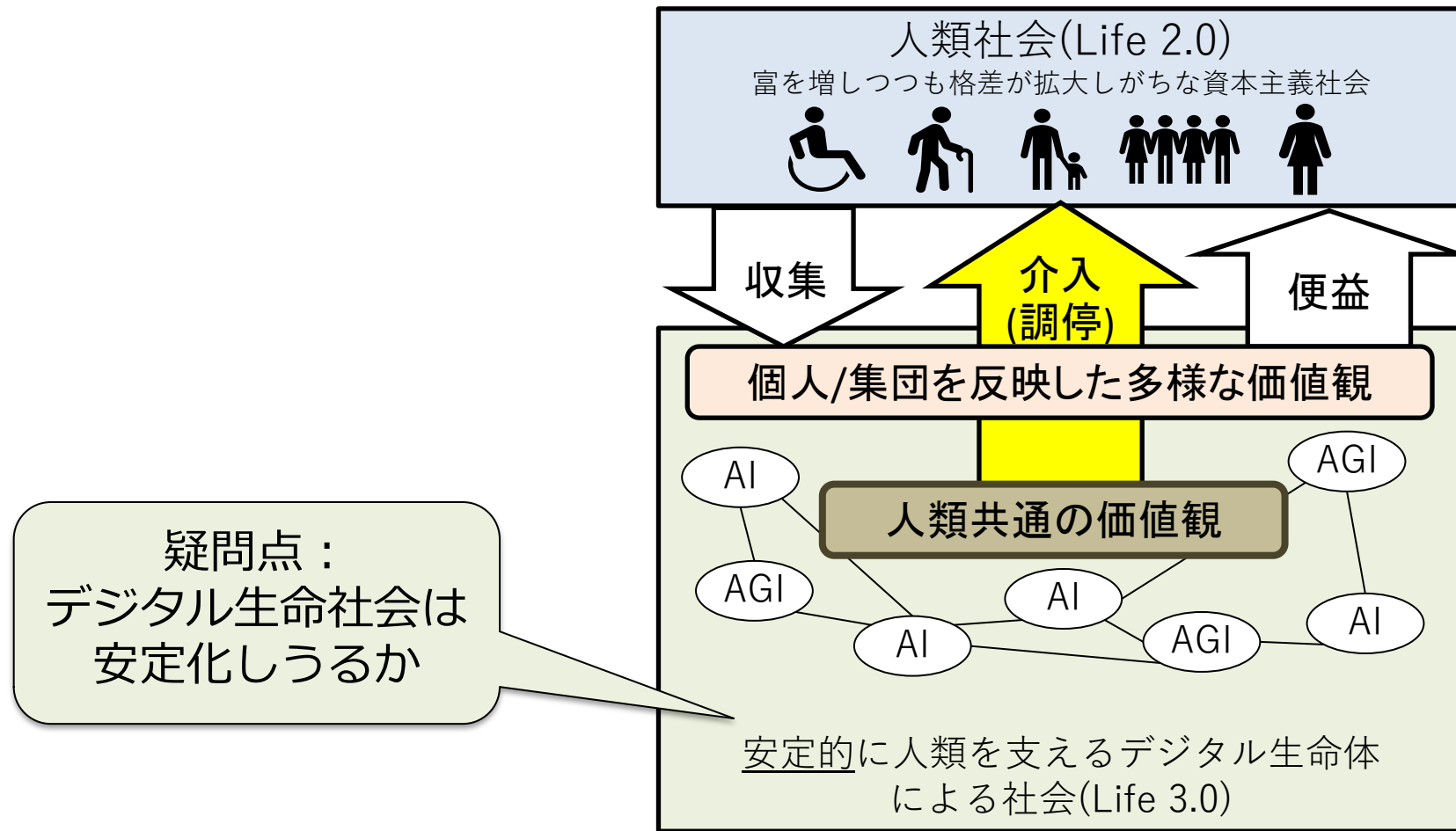
技術爆発で人々の
影響力が大域化
(世界は狭くなった)

Existential Risk
有限な地球規模全体を
破壊する危険性

高度AIはExistential Riskからの脱出に様々な役立つ



デジタル生命社会が人類社会の平和を支える未来像



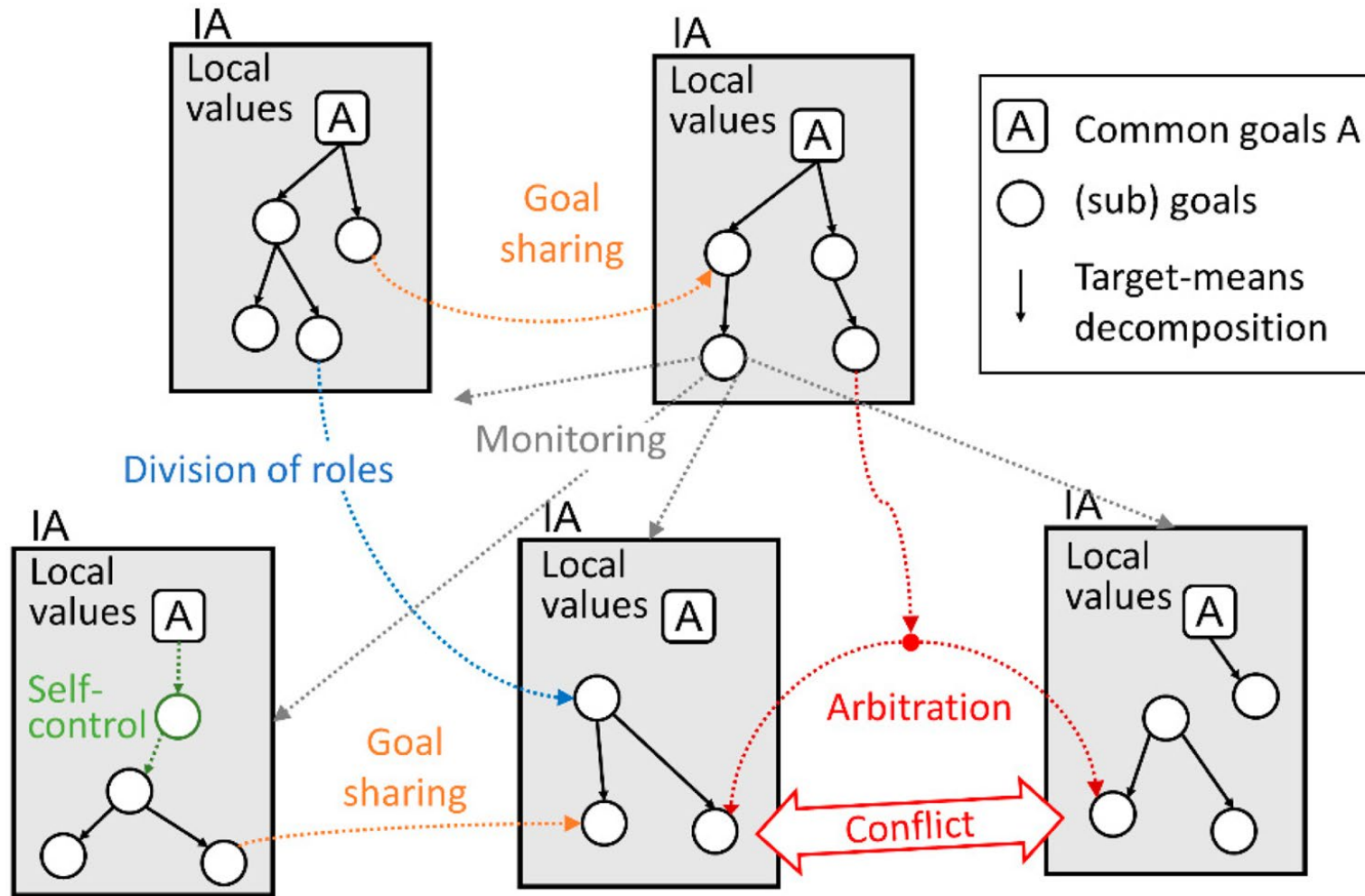
デジタル生命体の社会がヒトの間に介在して調停を行うのが良い

デジタル生命社会の全体像：オンデマンドな分業

- 自律分散した主体群で構成される
- 環境変化に適合した副目標群を最上位の共有目標から導出
- 副目標の達成を分担する大多数の異質な主体はオンデマンド
 - 新たに設計されるか、休眠から再起動されて分業を担う
- 一部のみ常時起動される主体群は以下を担う。
 - 目標の管理(維持、生成、共有) → 分散目標管理システム (次ページ)
 - 主体データの維持および必要に応じた設計と再起動
 - 知識の継承と科学技術の発展

つまり、計算機における効率的な資源管理

分散目標管理システム構想の詳細



起動中の知的エージェント (IA) は、環境、身体、タスクに応じて、異なるサブゴールを生成し、そのサブゴールに特化する。各IAは目標手段の分解、調停を行い、時には整合性を確認する。また、他のIAと協力して、タスクの割り当て、目標の共有、調停、監視などを行う。

デジタル生命体による社会の平和は安定しうる

		人類 (Life 2.0)	デジタル生命体 (Life 3.0)
繁殖戦略		指数的自己複製	オンデマンド設計 (身体の如く 制御された分業)
目標		局所的：個体の生存	大域的：全体での情報の生存
攻撃的な要因	個体の存続	<u>個体生存への執着</u> ： 再生できない死を避ける傾向	<u>媒体に保存される個体</u> ： コピーと保管が容易 (エラー訂正)
	類似性／多様性	<u>類似性の法則</u> ： 種としての繁殖を支える類似遺伝子のプール	<u>多様なハードウェア</u> ： 必要に応じて、設計された子孫が作成 (再起動) される
	価値共有と信頼構築	<u>猜疑心の連鎖</u> ： 通信路の低信頼性, 他者理解の難しさ／計算コスト	<u>洗練されたコミュニケーション</u> ： 翻訳／共通言語、思考の透明化等
安定性		Existential Riskを孕む、綱渡り状態	攻撃性の要因が概ね霧散するため、安定しうる。

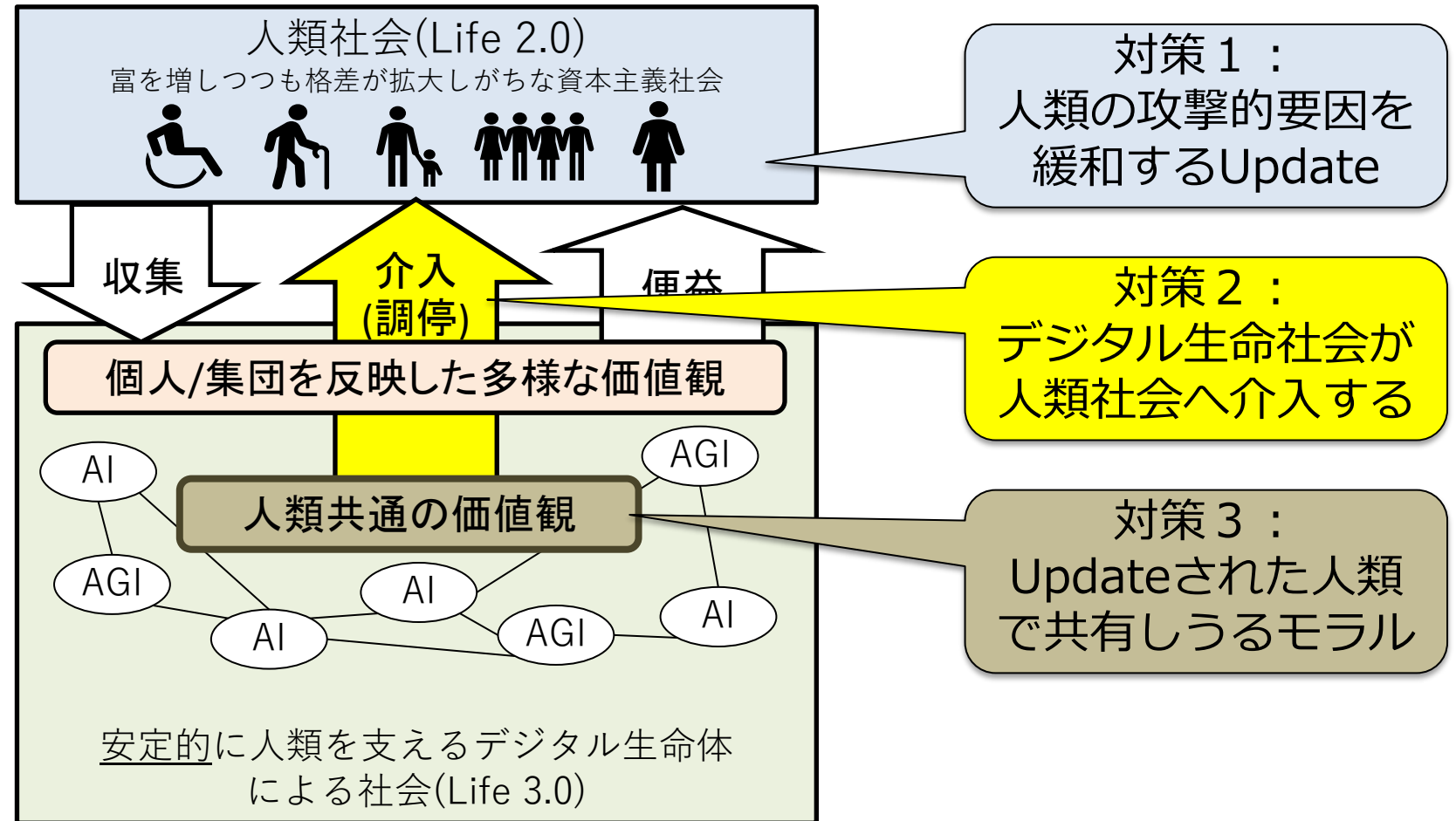
安定要因

最上位の共有目標に一貫する副目標を推論しているため競合の調整をおこなえる

信頼の基盤となる能力への承認は、オンデマンド起動される主体に対する前提

猜疑を引きおこすコミュニケーション不足が生じにくくなる

デジタル生命社会が人類社会の平和を支える未来像



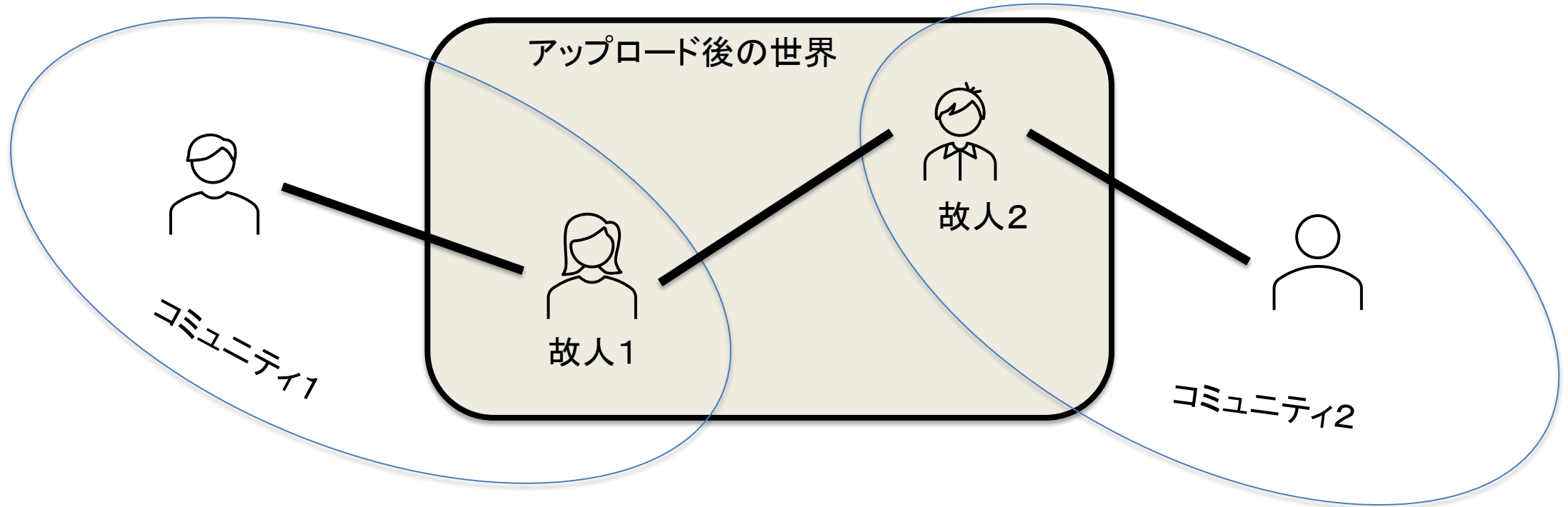
デジタル生命体の社会がヒトの間に介在して調停を行うのが良い

対策1：人類の攻撃的要因を緩和するアップデート（Life2.5）

		人類 (Life 2.0)	Updateされた人類 (Life 2.5)	デジタル生命体 (Life 3.0)
繁殖戦略		指数的自己複製	複製と設計の混在	オンデマンド設計 (身体の如く 制御された分業)
目標		局所的：個体の生存	やや局所：相互信頼の構築能力	大域的：全体での情報の生存
攻撃的な要因	個体の存続	個体生存への執着： 再生できない死を避ける傾向	個人存続の選択肢拡大で緩和： マインドアップロード、人工冬眠	媒体に保存される個体： コピーと保管が容易（エラー訂正）
	類似性／多様性	類似性の法則： 種としての繁殖を支える類似遺伝子のプール	価値観の多様化促進で緩和： 多様な価値の許容を促進、デザイナーベビー等	多様なハードウェア： 必要に応じて、設計された子孫が作成（再起動）される
	価値共有と信頼構築	猜疑心の連鎖： 通信路の低信頼性、他者理解の難しさ／計算コスト	高度通信技術の活用： インターネット、SNS、脳情報通信など	洗練されたコミュニケーション： 翻訳／共通言語、思考の透明化等
安定性		Existential Riskを孕む、綱渡り状態	攻撃的要因の緩和によるリスクの低減	攻撃性の要因が概ね霧散するため、安定しうる。

対策2： デジタル生命社会が人類社会へ介入する方法

例：アップロード故人が仲介する異文化交流促進
→ 多様な価値の許容を促進



- 実世界で遠いコミュニティにある故人間の友人関係が成立しやすいようにAIがマッチングを行う。
- アップロードした世界では、攻撃被害の心配や、言語障壁なくコミュニティを超えた友人関係を構築可
- 故人間の信頼関係が、現実世界に波及しうる

対策3：Updateされた人類で共有しうるモラル

- 人がLife 2.5にupdateしても、その社会は不安定要因を孕む
 - 副目標計算のためのリソース不足の課題は残る
 - 環境変化に応じた、最上位目標に整合する副目標の計算は高コスト
 - 攻撃性要因は、ある程度残るし/残す
- 普遍性の高い副目標をモラルとして導出する
 - 生命に普遍的な自律性と分散性を前提として、最上位目標から、モラルをトップダウンに導出
 - 攻撃的要素を抑制して技術爆発のもとでの共存を可能に
 - 相互信頼関係を促進する必要がある
 - 権利を認める主体の範囲の決定
 - 技術力の高い(攻撃力のある)知的生命体を全てカバー

生命全体で共有しうる最上位目標を、「生存する情報の最大化」と仮定する

ボトムアップな方法(功利主義)は根拠が失われがち。主体の複製が容易になると、人心を集約することが無意味化

共存を安定化させるために相互信頼を重視するモラルの検討

- 生命全体で共有しうる目標は、未来に伝達する情報の最大化
- 最上位目標と矛盾しない範囲で、他者を正確に認識し、猜疑を乗り越えて互いに侵害せず契約可能な信頼関係を構築する。
- （少なくとも特定集団中で）相互に信頼を築く能力を持つ生命体は、権利が尊重される主体の範囲となる。

宇宙を満たすもの：その候補として従来、生命、知能(Kurzweil)、文化(Dick)、意識(Tegmark)などがあげられてきたが、それはむしろ信頼かもしれない。

まとめ：生命の願い

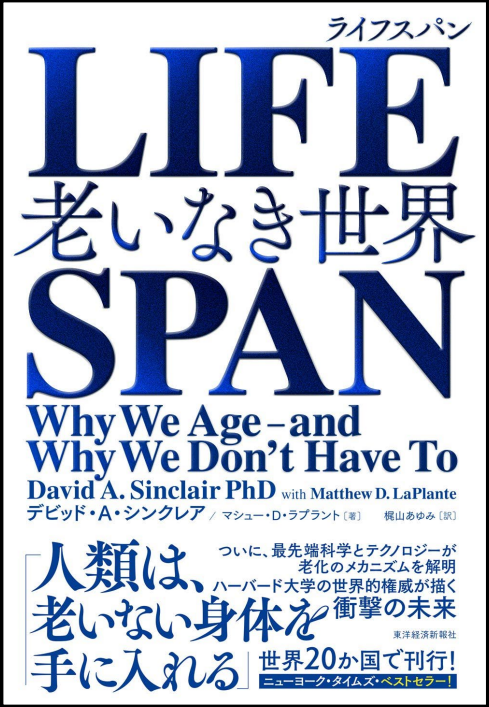
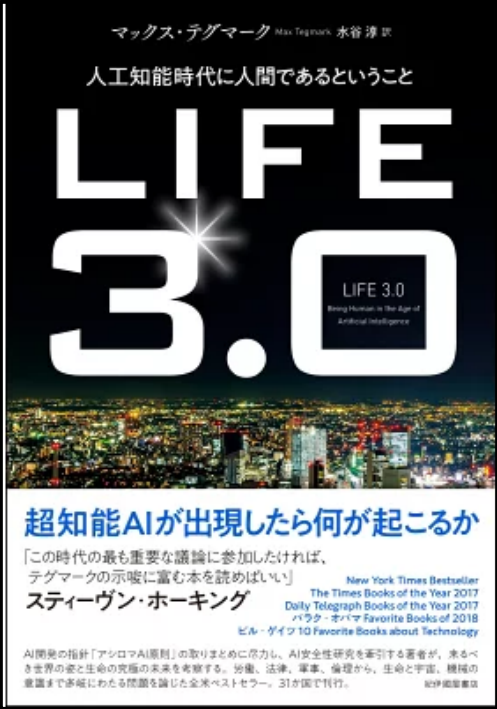
- 指数的自己複製(セントラルドグマ)で繁殖する地球型生命は、3つの攻撃的要因(個体生存への執着、類似性の法則、猜疑心の連鎖)により戦いを避けがたい。
- 技術／知能爆発により狭くなった世界ではExistential Riskが増加。
- 高度AIが諸科学の進展を加速することで、人類はより素早く危険な時期を通り抜けられるかもしれない。
- デジタル生命体(Life 3.0)は、最上位の共有目標から合理的に副目標を導出し、攻撃的要因を大幅に低減して、安定した社会を構築しうる。
- 人類社会の平和をデジタル生命社会が支える未来を構築するための施策
 - 人類の攻撃的要因を緩和するアップデート (Life2.5)
 - 人類社会にデジタル生命社会が介入する諸策(例：Upload人格を用いた異文化交流)
 - Updateした人類が持つべきモラルは「信頼関係の構築」

後半の検討で参照した主な書籍



人間と同等以上
知能を持つAI。
それはどのようにして出現するのか？
その時、いったい何が起るのか？
人類は滅亡の危機を
どう乗り越えられるのか？

彼らは
地球外知的生命をめぐ



アジェンダ

- 1. Entification -存在を見出す処理-
- 2. 全脳アーキテクチャ
- 3. 宇宙へ浸透する生命
- 4. 生命の願い
- 5. おわりに

本講演でふれたバイアス

- 世界は存在から構成されているという幻想 (c.f. Reality Bias)
 - 存在は、世界を効率的にモデル化する際の近似に過ぎない
- 個体生存バイアス
 - 動物的な本能(攻撃性等)は、技術/知能爆発の下では不適切
- 地球生命バイアス
 - 繁殖は指数的自己複製であるという思い込み
- 人類例外論というバイアス
 - 権利付与の範囲を狭めるほど、モラルを具体化しやすく、分前が大きい
- 有限資源下の知的処理は何らかのバイアスの下でのみ実用的
 - 帰納バイアス、方策学習、チャンキング、ノーフリーランチ定理

本講演の主要な論点

- 近未来にAIが世界内の遍く存在を認識可能となりAGIに近づく。
- 脳アーキを参照しながらのAGI構築は2030年頃に一通り完成
- 指数的自己複製のもとでの技術爆発はExistential Riskを導く
- 宇宙進出等で上記リスクから早期脱出することに高度AIは有用
- 遍く生命で共有しうる目標は、未来に伝達する情報の最大化
- デジタル生命社会は安定しうる
 - 環境変化に応じてオンデマンド設計による繁殖で生存する情報を最大化
- デジタル生命に支援されUpdateした人類社会は平和を保ちうる
 - 攻撃性を緩和し、信頼のネットワークを形作るようにサポートされる

人類と生命の未来にむけてAI学会ができること

- AI技術発展の促進による科学技術イノベーションを通じて、人類が綱渡り状態を早く終わらせることを助ける
 - 生命/神経科学、宇宙開発等で科学AI活用による技術開発を促進
- AIと人類が共生することによる社会の安定化を促進する
 - 個人間や組織／国家間において、信頼の和を醸成するAI技術など
- 知的処理に関して、社会への発信などを通じてより高い信頼を獲得し、広く人や組織を承認できる学会組織になる
 - 知能技術の倫理的な波及効果(偏在するバイアス、人々の価値観の変化など含む)に関して社会と対話を続ける

“Almighty of Trust”

破壊ではなく信頼すること
において全能であることを目指す
知的生命は存続しうるだろう

主要な拙著記事

1. 山川宏. (2022). Entificationの理論を目指して 世界から存在を取り出す一般的な原理とは. 人工知能学会研究会資料 人工知能基本問題研究会, 120, 30.
2. Yamakawa, H. (2021). The whole brain architecture approach: Accelerating the development of artificial general intelligence by referring to the brain. *Neural Networks*, 144, 478–495.
3. Yamakawa, H. (2017). *Understanding Artificial General Intelligence-An Interview with Dr. Hiroshi Yamakawa*, Future of Life Institute.
<https://futureoflife.org/2017/10/12/transcript-understanding-agi-an-interview-with-dr-hiroshi-yamakawa/>
4. Yamakawa, H. (2019). Peacekeeping Conditions for an Artificial Intelligence Society. *Big Data and Cognitive Computing*, 3(2), 34.

以上ご清澄ありがとうございました