ARAYA

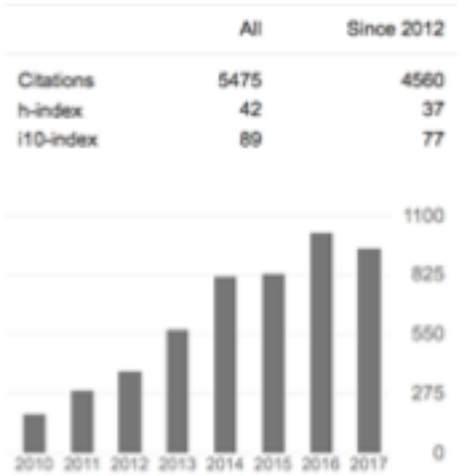
人工意識の話

金井　良太
（株）アラヤ　創業者
TW: @kanair_jp

自己紹介

元・英国サセックス大学准教授 （専門は認知神経科学）
文部科学大臣表彰 （2015）
JST CREST「人工意識プロジェクト」代表研究者
大学での研究者を辞めてアラヤを創業する。

神経科学・意識研究・人工知能分野で論文多数 (**h-index=42**)



なぜスタートアップなのか？

ARAYA

現在

神経科学の研究は知的好奇心を刺激するが、
現実世界にインパクトを与えていない。

この境界線を超える

５０年後には…

意識のメカニズムが解明され、感じることのできる
機械が生まれる。

人間の脳の中の情報は外の世界とつながり始める。


ARAYAのミッション
１．意識の情報原理を解き明かす
２．意識を持った人工知能を構築する

ARAYAのチーム

社員数　23名

技術部門　内訳　　データサイエンスグループ　11名
　　研究開発グループ　9名

PhD　17名
MS　3名

物理学・情報科学・神経科学・複雑系など多様なバックグラウンドの精鋭集団

意識のことを考える3つのステージ

ARAYA

1. **クオリアに気づくこと**
   - 意識の問題の深さと特異性に気づくこと

2. **ハードプロブレムに気づくこと**
   - 科学の扱える問題なのかと絶望的する

3. **絶対的不可能な問題でないと気づくこと**
   - 具体的な解くべき問題に落とせる
   - 科学における理論的研究の性質を理解する

この全てを実感として感じることが重要

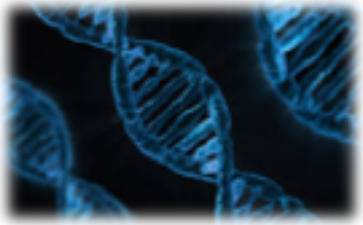
存在に関する三大起源の謎

1．宇宙の起源（宇宙物理学）

どのように宇宙は誕生したのか。
なぜ、宇宙は存在するのか。

2．生命の起源（進化生物学）

どのようにして地球に生命は生まれたのか。
どのような環境で生命は生まれるのか。



3．意識の起源　（学問分野がない）

どのようにして、脳から意識がうまれるのか。
どのような神経回路が意識を生み出すのか。

意識研究への懐疑論

「意識を定義できないから研究できない」（多くの科学者）

「意識は客観的でないから科学にならない」（行動主義者）

「意識は還元論的アプローチでは解けない」（チャルマーズ）

「意識は錯覚である」（ダニエル・デネット）

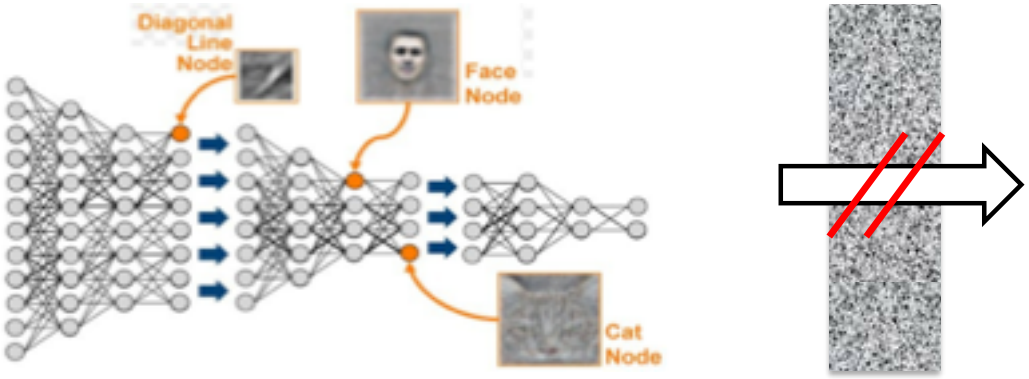
「意識は科学の扱う範囲外にある」（二元論者）

「意識の問題は人間の理解を超えている」（ミステリアン）

歴史的には懐疑的な意見もあり、
意識に興味がある研究者も本格的な研究に躊躇

人工知能の壁

現在の人工知能の延長でAGIはできない



現在の課題

1. 自発性（内発的動機、意図、好奇心）
2. 汎化性（創造性、思考）
3. 説明可能性（メタ認知、言語）

⇨ **汎用人工知能（AGI）を作るために意識を理解する。**

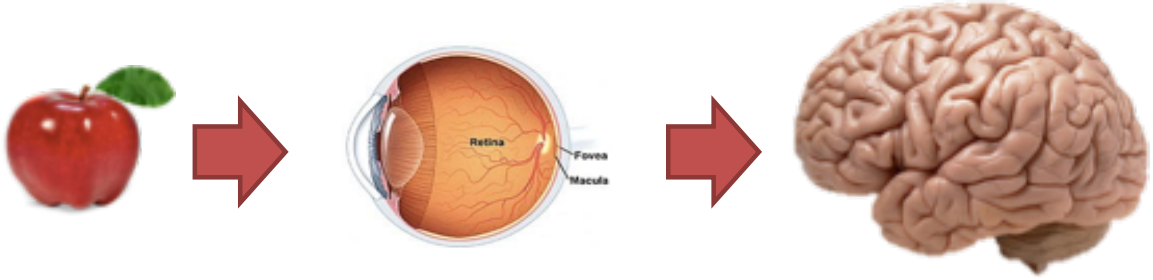
意識の問題とは何か？

ARAYA

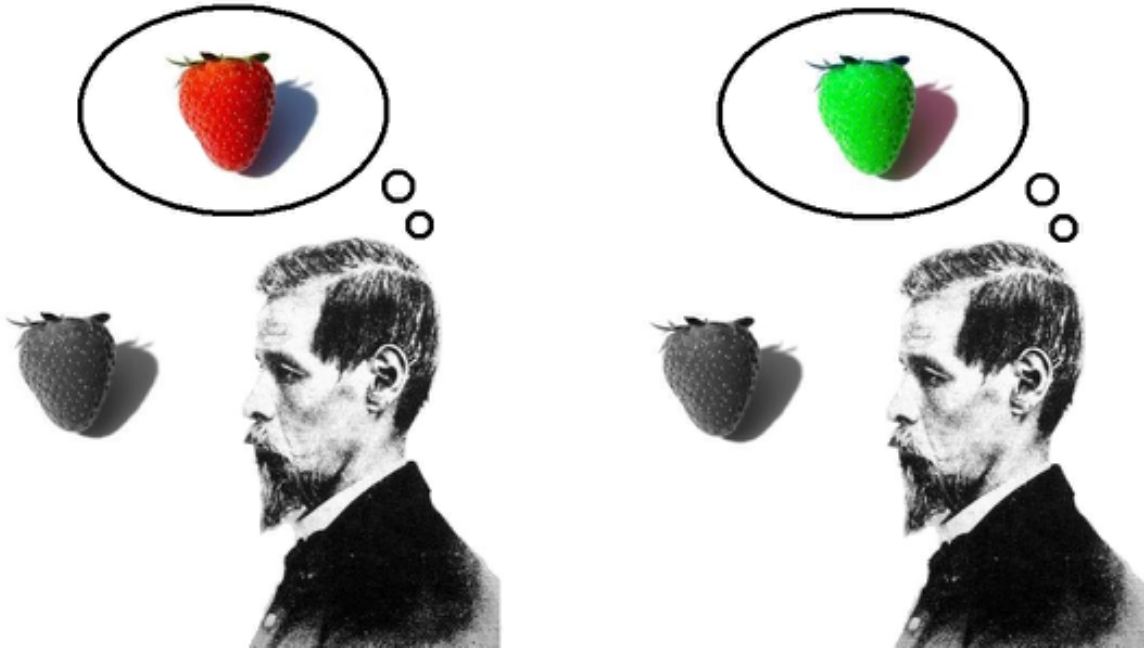
脳内の物理的現象から主観的体験が
如何にして生まれるのか？

主観的世界

物理的世界

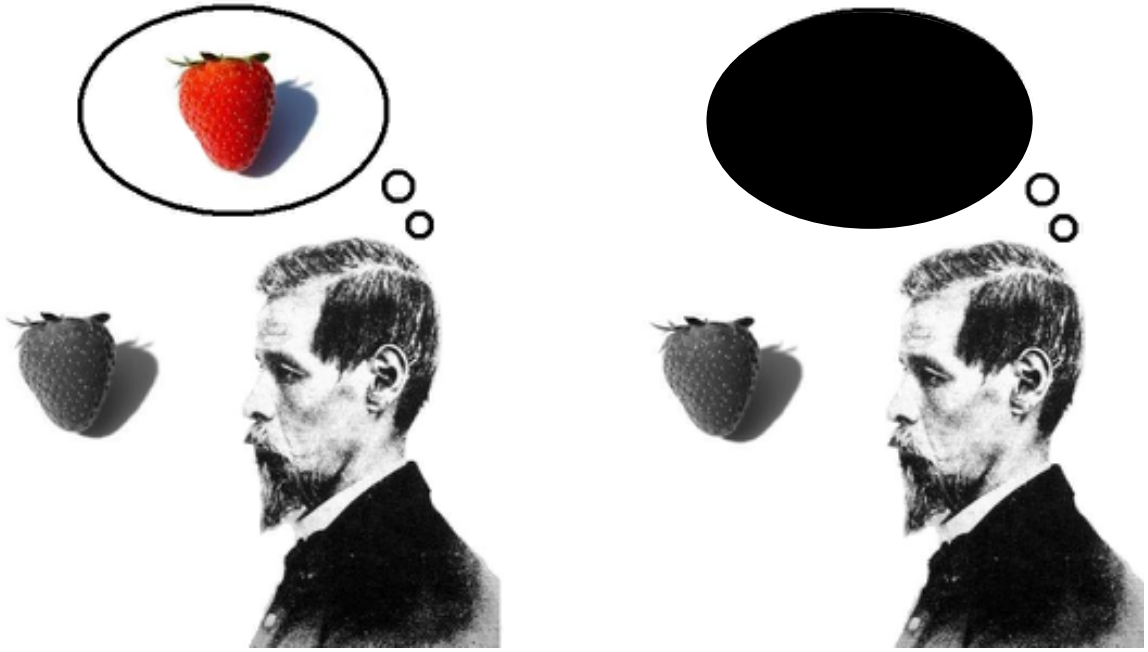


逆転クオリア



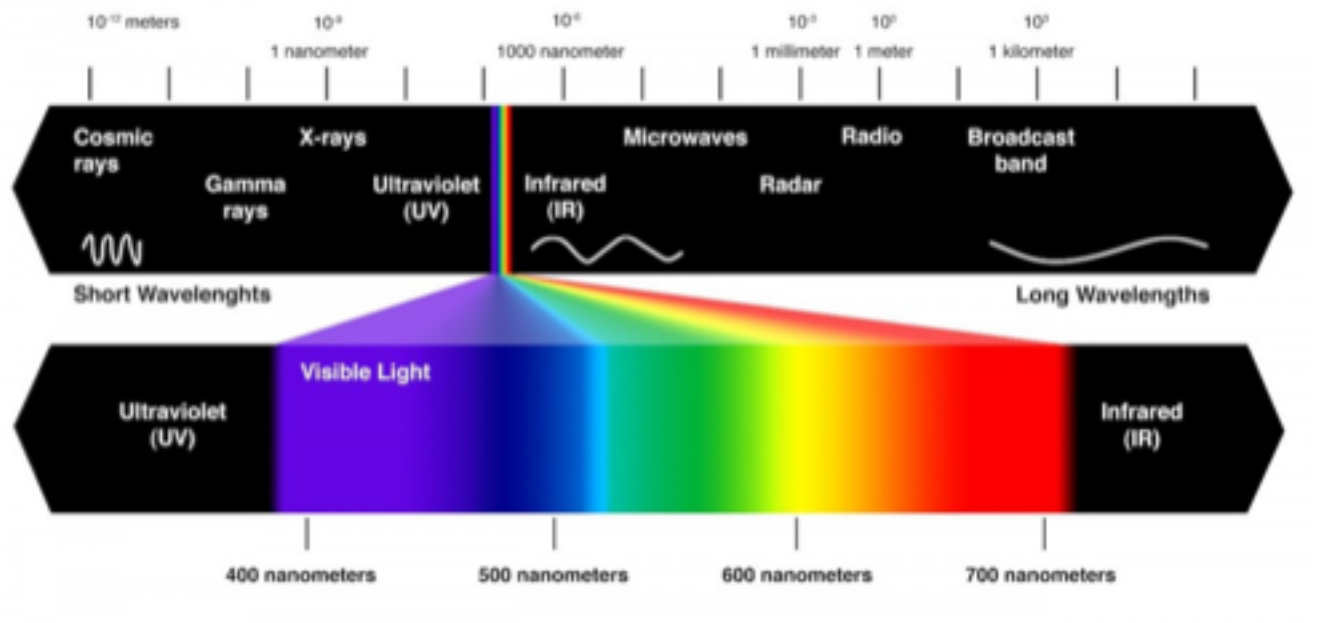
自分の見ている赤と他人の見ている赤は同じか？

哲学的ゾンビ



物理的に同じ人間が意識経験を持たないことも
「想像可能」である。
しかし、なぜ実際には意識経験があるだろう。

ユクスキュルの環世界(Umwelt)



人間に見えている世界は、感覚器により限定されている。
動物それぞれに特徴的な主観的な世界がある。

コウモリであるとは

- 哲学者トーマス・ネーゲル

「コウモリであるとはどのようなことか」

コウモリにとって、超音波で空間を把握するのは、
「見ているような感じ」か「聞いているような感じ」か？

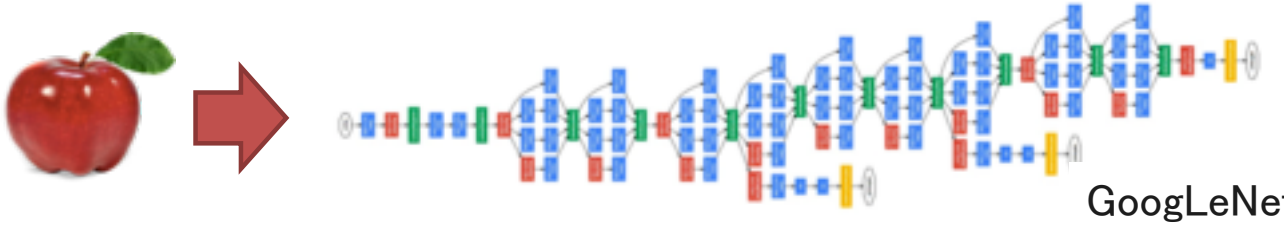
意識の問題とは何か？

ARAYA

現代のニューラルネットは感じているのか？

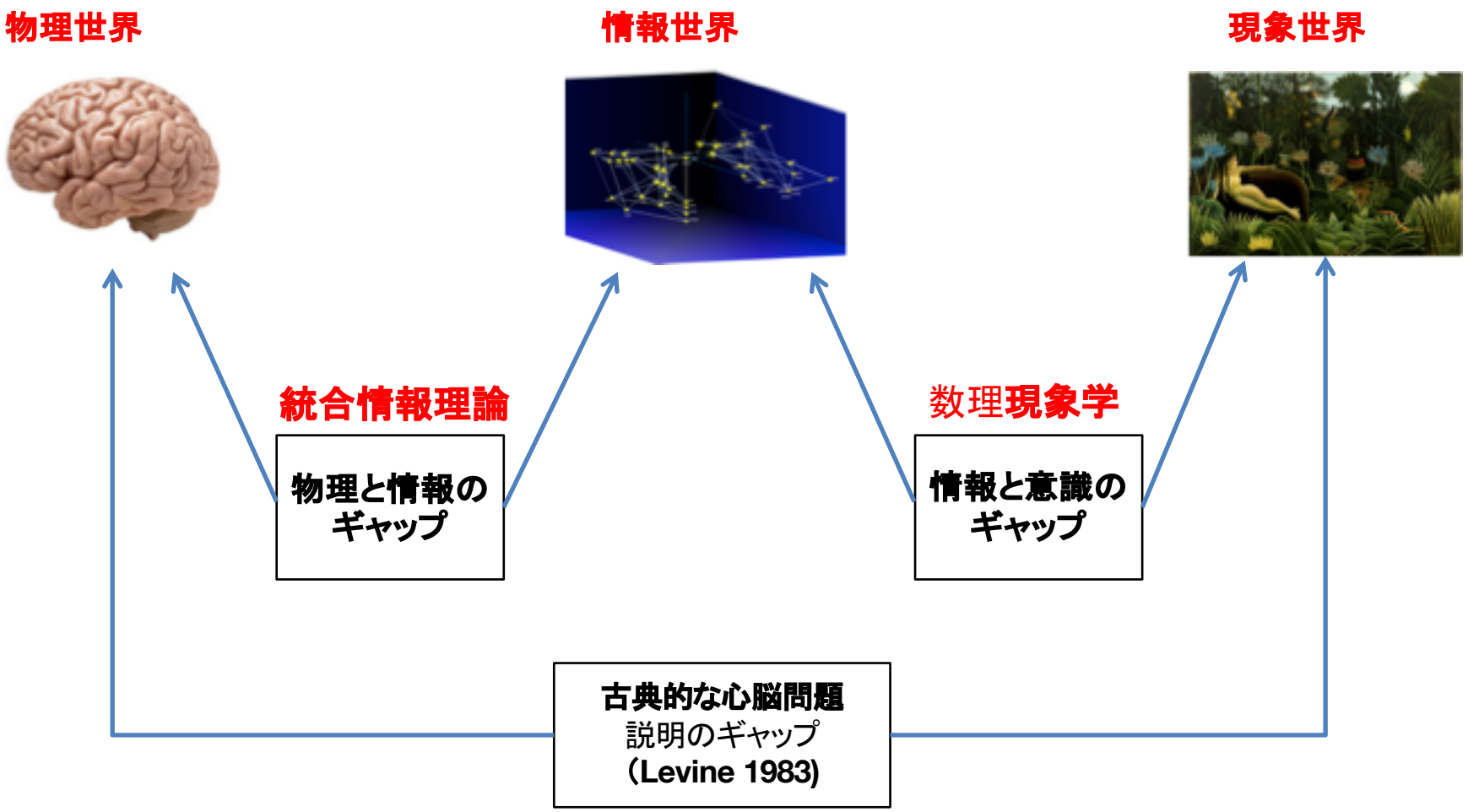
主観的世界

物理的世界



そうでないならば、なぜか？

意識のハードプロブレム



意識は定義できるのか？

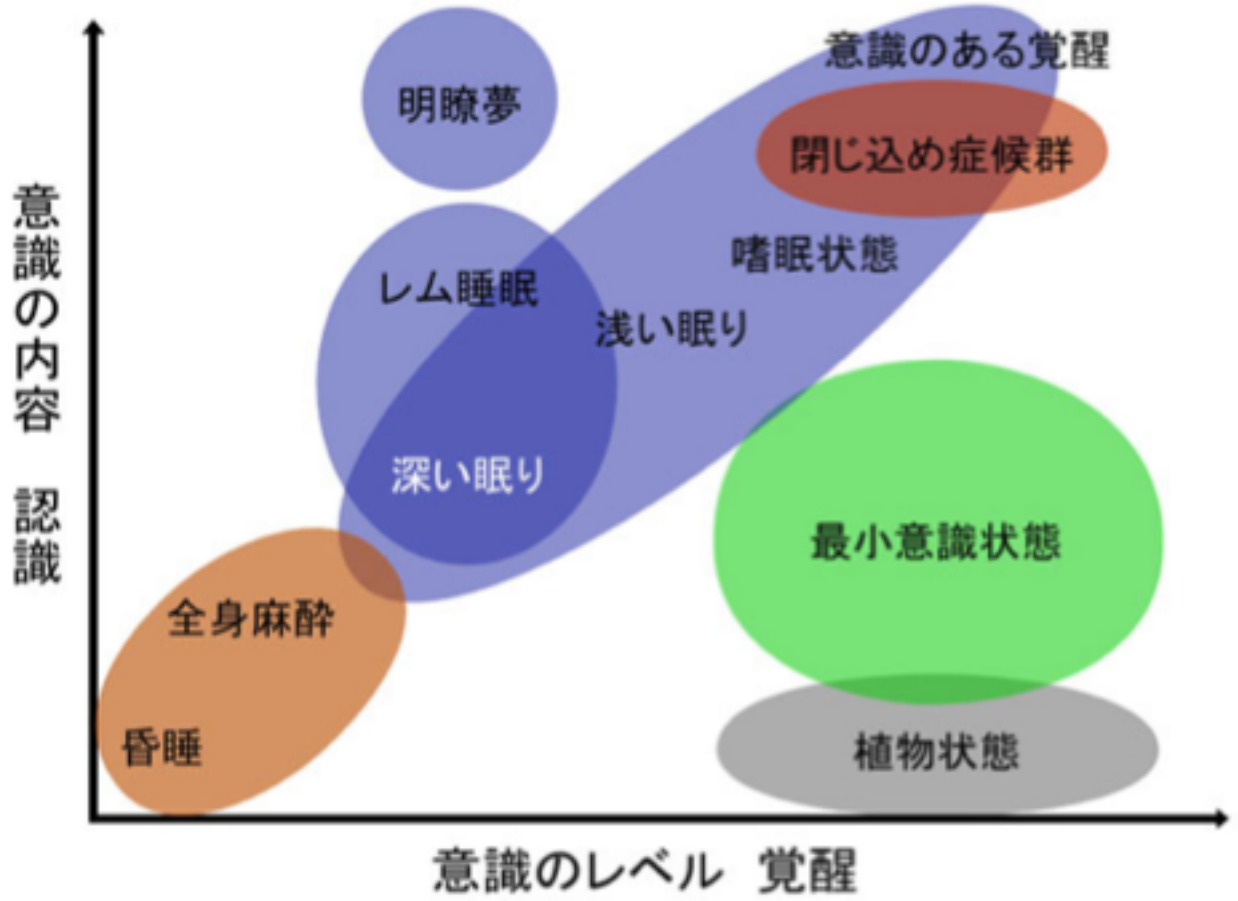
- 誰もが、素朴な直感を持っている。
- 現時点で明確な定義をすることは難しい。
- 科学において、対象の理解が深まるに連れて定義も洗練されていく（例：遺伝子、生命、熱、情報）。

暫定的定義

意識の内容（知覚、主観的経験）と
意識レベル（医学的な意味・覚醒レベル）のこと

定義を洗練させていくことも意識学のテーマ

意識の内容と状態


意識のある覚醒
明瞭夢
閉じ込め症候群
嗜眠状態
レム睡眠
浅い眠り
深い眠り
最小意識状態
全身麻酔
昏睡
植物状態
意識の内容　認識
意識のレベル　覚醒

意識学を作りたい

- 意識研究に関わる重要は多岐に渡る
  - 神経科学、心理学、認知科学、医学、生物学
  - 情報科学、人工知能分野、ロボット学、
  - 物理学（力学系・複雑系・人工生命など）
  - 数学（圏論、トポロジーなど）
  - 哲学（現象学、心の哲学、科学哲学）
  - 宗教（仏教・唯識）、文学・アート（主観）

関連する分野を一貫して学べる学部や研究所がない。
意識研究者が一斉に集まる研究所がない。
（100億円ぐらい欲しい）

ARAYA

なぜ意識を
研究しなければならないのか

意識モードと無意識モード

- 意識は知覚体験の問題だけではなく、 全ての心的現象に関わる。
  - 知覚（サブリミナル ― アウェアネス）
  - 記憶（手続き的記憶 ― 宣言的記憶）
  - 思考（無意識的 ― 意識的）

- 認知機能が意識を介在する理由は何か？

ほぼすべての心理学者・神経科学者・認知科学者にとっても関連の深い問題。

情報と意味の関係

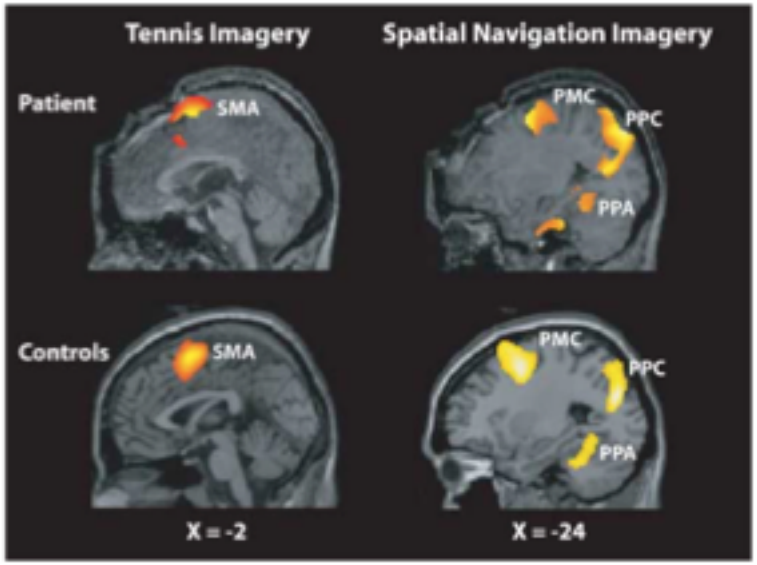
- 情報や意味が何であるかわからない。
  - 情報の観点から「主観的な感じ」とはどういうことか？
  - 「意味」を理解するとは何か？
  - 「理解」とは何か？

- 情報と意味の関係の理解は、新しい技術を生み出す可能性がある。
  - 熱力学や量子力学に匹敵する理論が生まれるかもしれない。
  - 人工知能ブーム、侮れない。

新しい物理的世界観を生み出す可能性がある。

意識は切実な問題

- 植物状態の患者で意識を持っていることが、fMRIで見つかることがある。



医学的にはプラクティカルな重要性がある。

Owen et al. (2006) Science

知能と意識

ARAYA

無関係ではないか？

知能	客観	機能	外的行動
意識	主観	経験	内的メカニズム

作業仮説

客観機能は内的メカニズムを規定する。
(知能を作ると自動的に意識を持った内的構造を必要とする)

なぜ人工知能に意識の理解が必要なのか

ARAYA

- **ヒトではすべての認知機能に意識が関わっている**
  - 記憶と学習、知覚、思考、行動、意思決定、感情などの全てに「意識モード」と「無意識モード」があるが、その違いがわからない。

- **意識的経験には意味理解が付随している**
  - しかし、現在の人工知能技術では「意味理解」を実現できていない。

意識の２つの側面

ARAYA

アクセス意識

- 外から観測可能な意識の持つ客観的・機能的側面
- 報告可能性（内観・メタ認知による報告など）

現象的意識

- 意識体験の主体（セルフ）によってのみ感じることのできる意識の主観的側面
- クオリア

意識の２つの問題

ARAYA

ハード・プロブレム

どうして、物理的な脳の感覚情報に、主観的な感覚（クオリア）が伴うのか？

⇔どうして哲学的ゾンビは不可能なのか？

⇔どうしたら、動物や人工知能の意識を確かめられるのか？

イージー・プロブレム

刺激の弁別、情報の統合、知覚内容の報告、注意の方向付けなどの機能的側面を神経活動から解明する。

人工意識の２つの課題

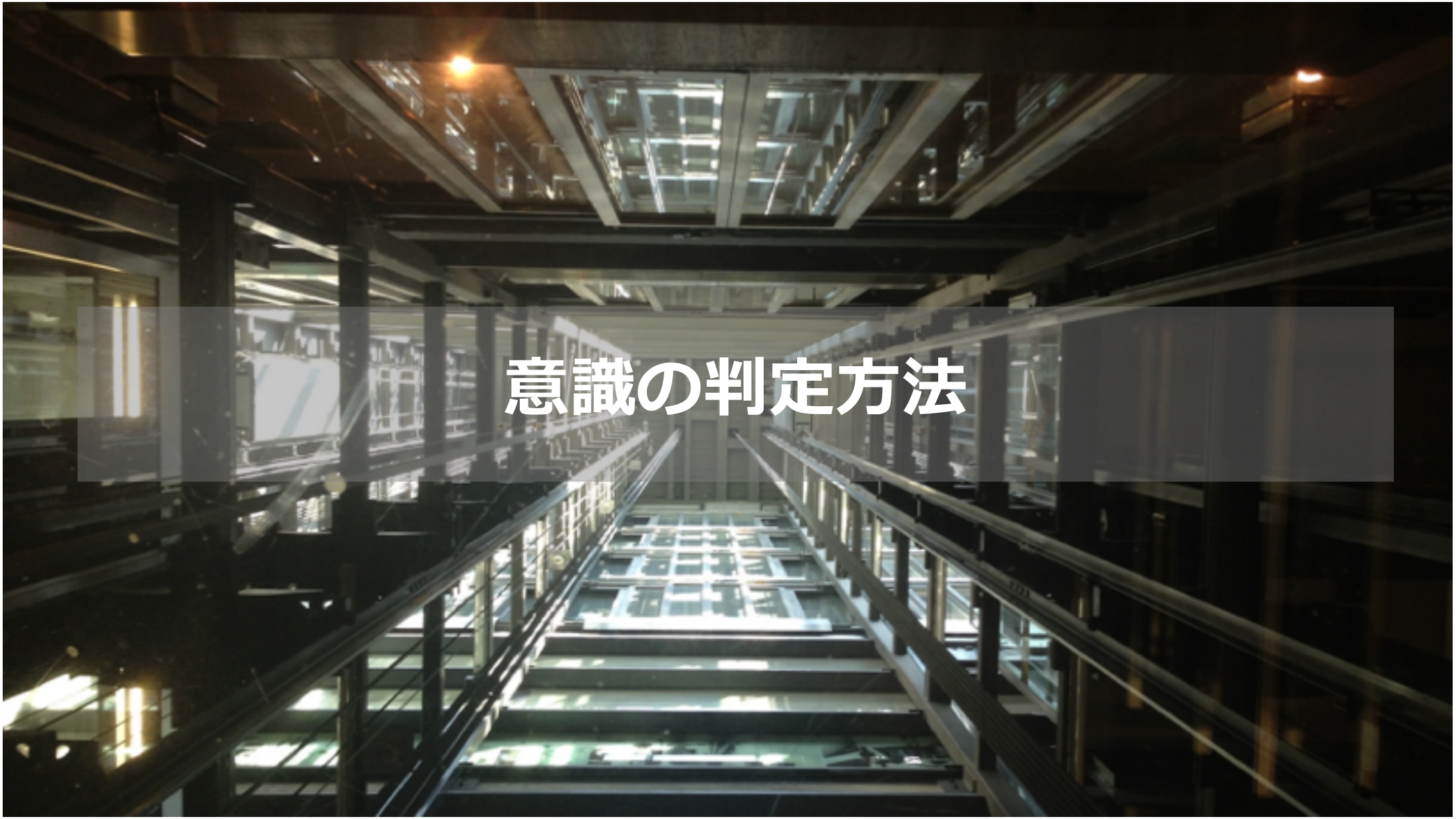
ARAYA

意識の創造

- 意識の持つ因果関係や機能を人工的に実現する。
- アクセス意識の創造に対応。

意識の存在証明

- 人工知能に意識があることを証明する。
- 現象的意識の存在証明に対応。

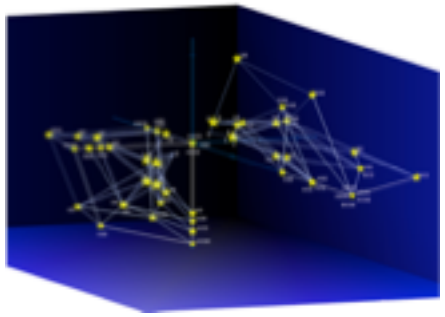

意識の判定方法

人工意識の判定方法

ARAYA

- 統合情報理論 (IIT)
  - システム自身の内在的視点からの情報構造を扱う
  - 現象学的観察を元に公理系を構築
  - 主観的経験はシステムの情報構造と同一性をもつ

内在的情報構造



MICS

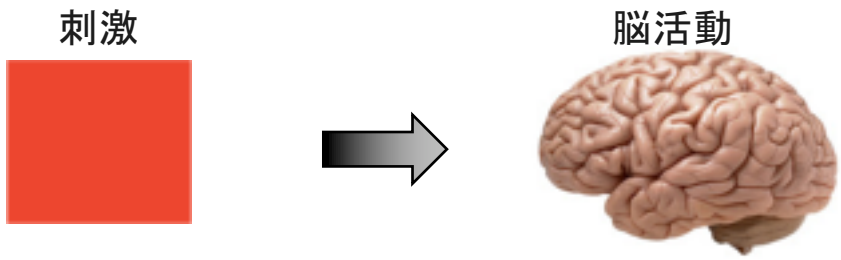
Φ =

意識的経験

意識理論の内在的視点

ARAYA

従来のシャノンの情報理論
（外側からの視点）



実験者の視点による対応関係
外部からラベル付されている情報

新しい情報理論
（内側からの視点）

脳に見えているのは脳活動のみ

ラベル付けされていない情報

統合情報理論における公理系と同一性の主張

ARAYA

５つの公理

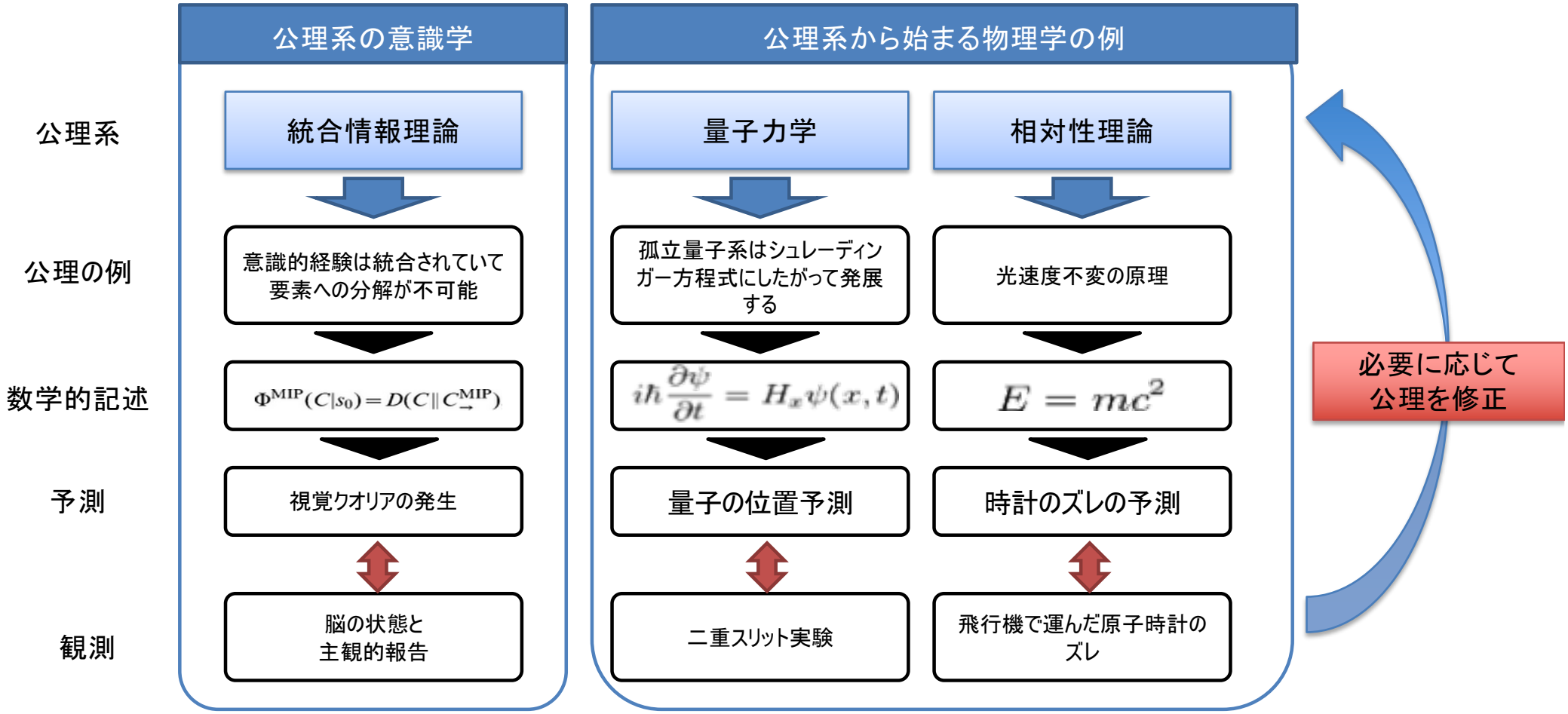
- 存在性: 意識というのは現実に起きている現象である
- 構造性: 個々の意識経験は経験の組み合わせにより構成されている。
- 情報性: 個々の意識経験は、多数の可能性の中の特定の経験である。
- 統合性: 個々の経験は、独立な要素に分解できない。
- 排他性: 個々の意識経験は、特定の時空間的スケールで生じる。

同一性の主張

- 同一性: 現象学的経験は物理システムにおいて生成される因果情報量（Φ）と同一である。

公理系の意識理論

ARAYA



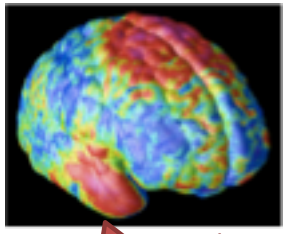
公理系からの意識理論

ARAYA

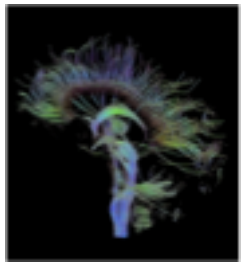
現象学・情報理論からの
トップダウン

意識

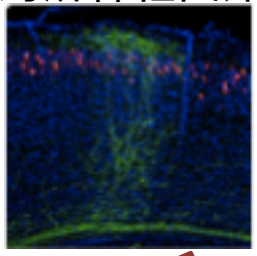
脳全体



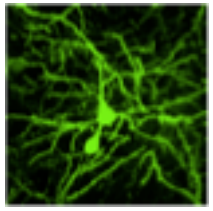
広域神経回路



局所神経回路



ニューロン



説明

説明

説明

説明

ハードプロブレムはココ！

還元主義のボトムアップ
（生物学）

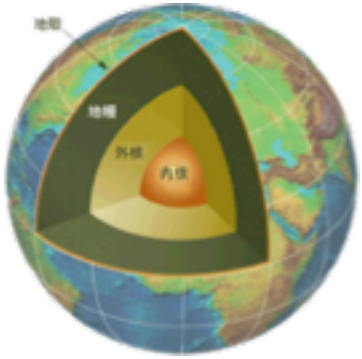
観測不可能な自然現象についての推論

ARAYA

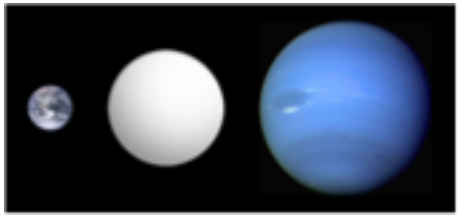
科学においては直接観測が難しいものがたくさんある

オーストラリアの季節

地球の内部構造



太陽系外惑星の環境

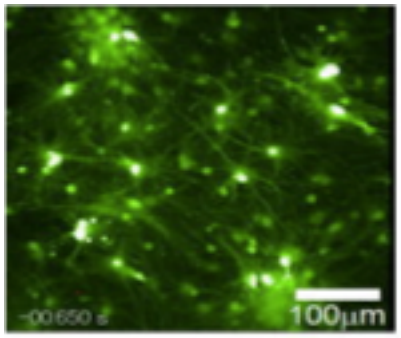


人間は理論のレンズを通して自然を知る

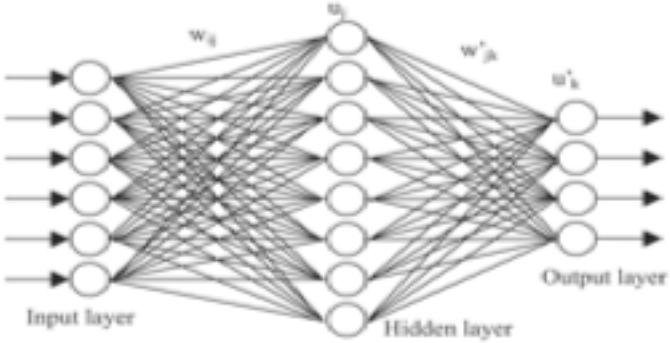
人工知能の意識についての推論

ARAYA

シャーレの中の脳細胞



人工ニューラルネット

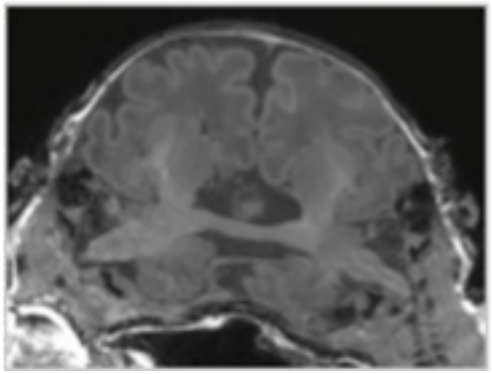


ロボットの中のプログラム

Φを計算することで、機能と意識の関係を見つけ出す

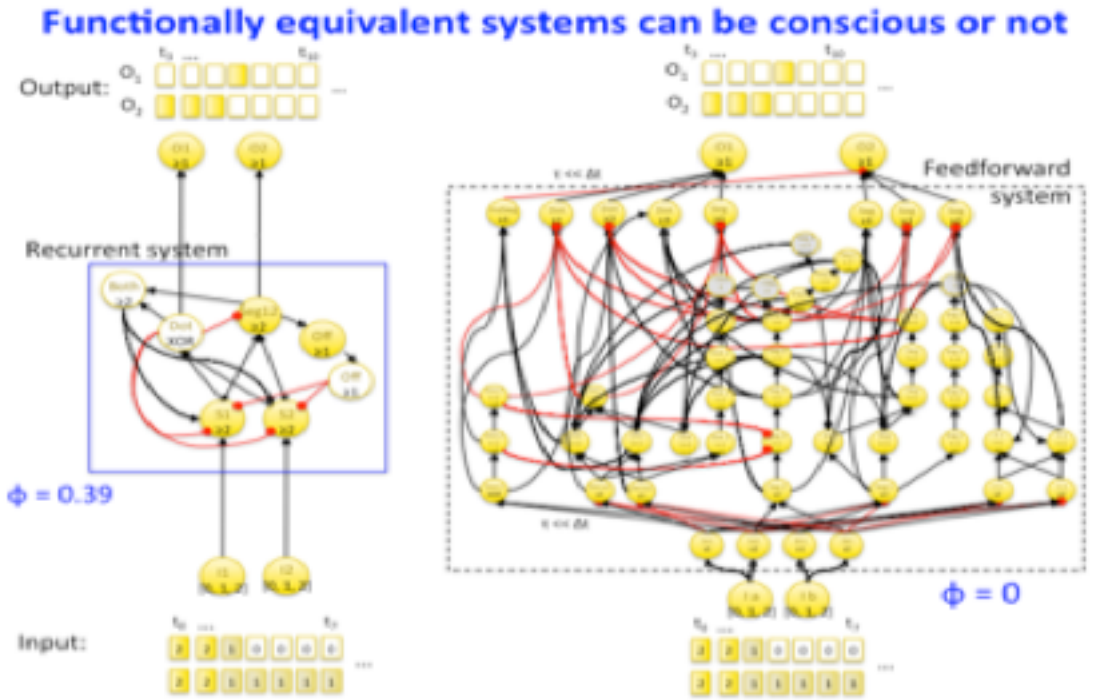
究極的には直接AIの人工意識にアクセス

視床で繋がっているHogan姉妹は
クオリアをシェアしている。



- 脳と脳を直接繋げばクオリアが共有できる。
- 脳とAIを直接繋げば、AIのクオリアを確かめられる。

機能が同じでも内部情報が違う



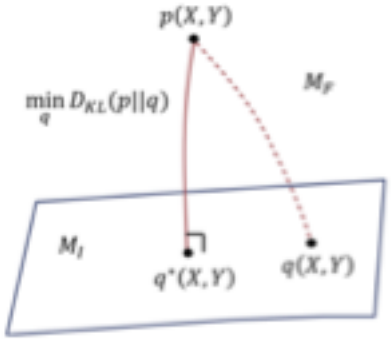
同一の機能でも内部構造によってΦは異なる

統合情報理論の進化

ARAYA

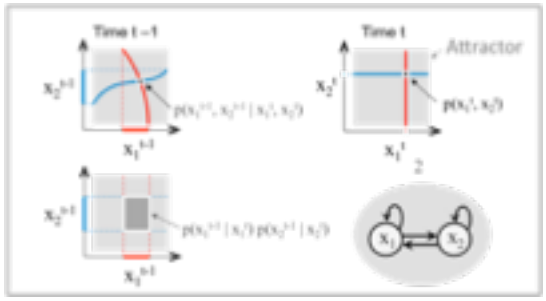
統合情報理論の実システムへの応用へ向けた進展

情報幾何



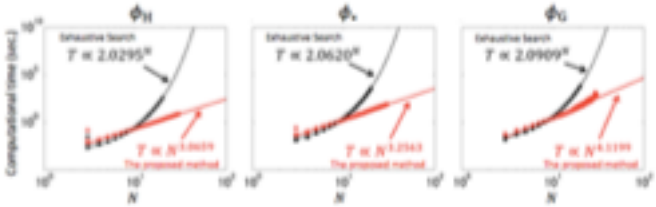
Oizumi et al. (2016) PNAS

力学系



Tajima & Kanai (2016)
Neuroscience of Consciousness

劣モジュラを仮定した計算高速化



Kitazono et al. (2017)
ASSC in Beijing

元来の目的であったDNNなどの実システムでの統合情報量の計算が見えてきた。

これができると人工意識の評価が可能となる。

劣モジュラ性による計算高速化

ARAYA

Submodularity

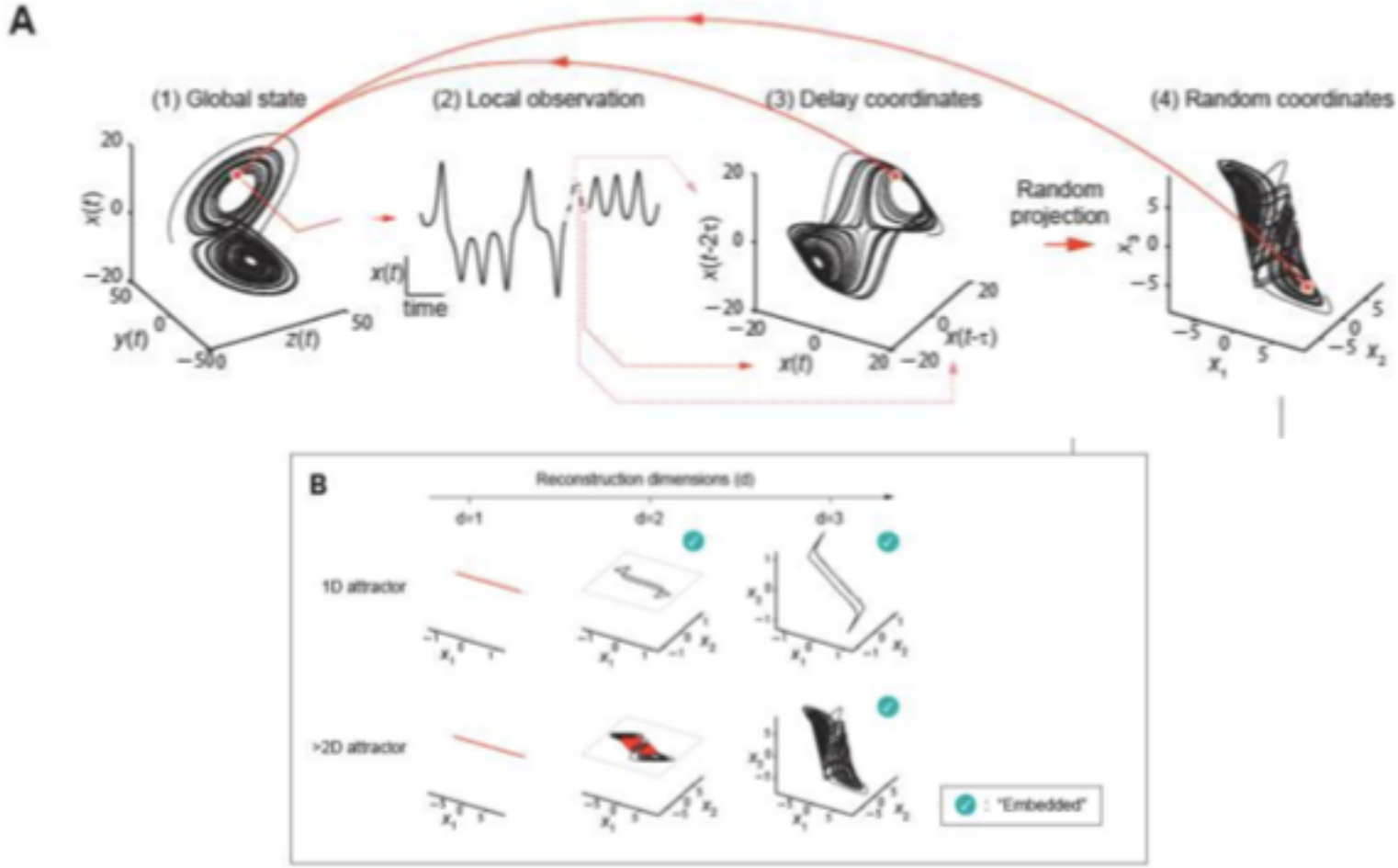
A set function $f: 2^{\Omega} \to \mathbb{R}$, where Ω and 2^{Ω} are a finite set and its power set, is submodular if it satisfies the following inequality for any $X \subseteq Y \subset \Omega$, and $x \in \Omega - Y$:

$$f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y).$$

Theorem (Queyranne)

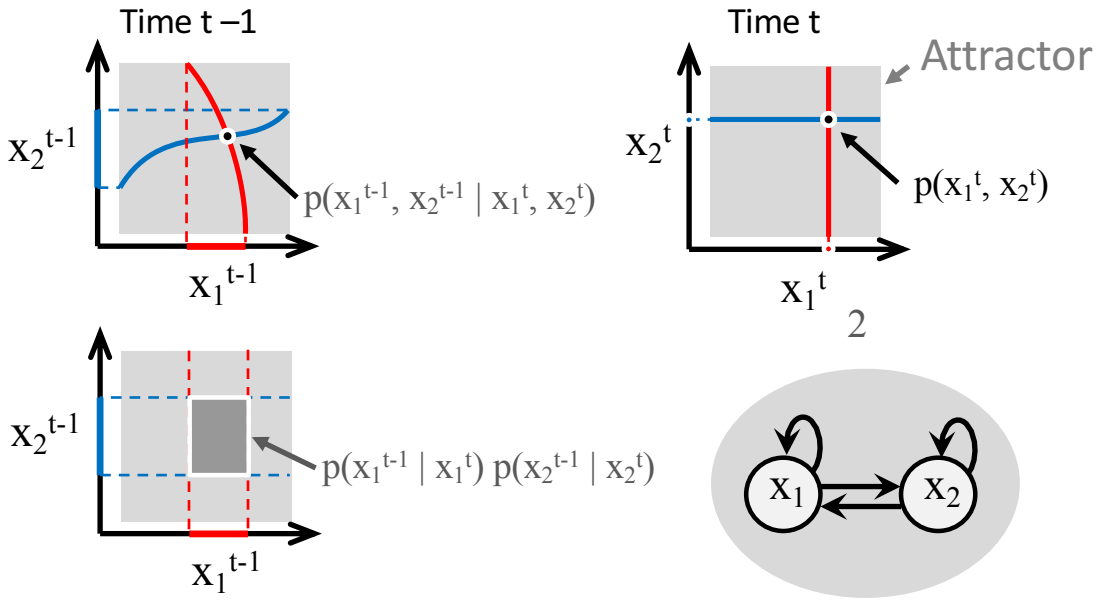
Symmetric submodular functions ($f(X) = f(\Omega - X)$) can be minimized using $O(N^3)$ function evaluations.

力学系的解釈による実システムでの評価



Tajima et al. (2015)

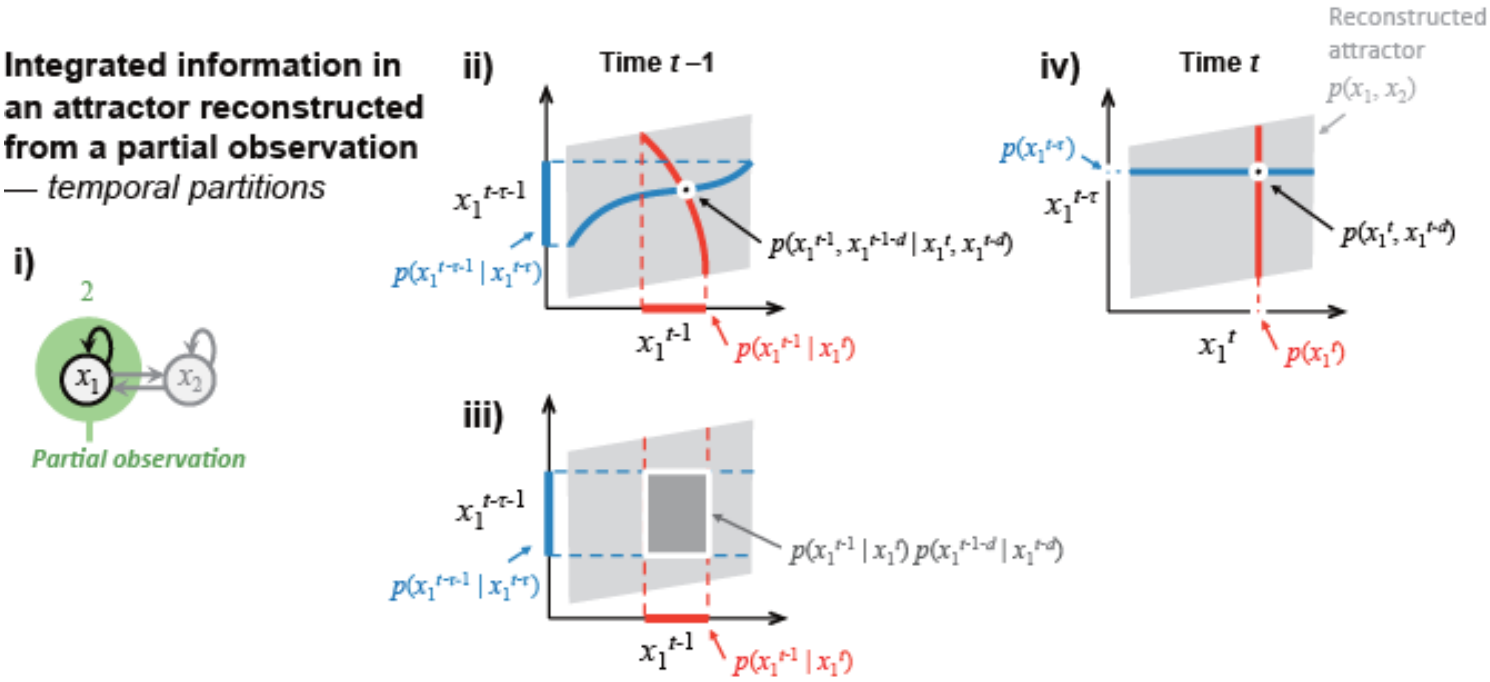
力学系的解釈による実システムでの評価



$$\varphi^{\mathrm{Dim}} \equiv \mathrm{Dim}[p(x_1^{t-1}|x_1^t)p(x_2^{t-1}|x_2^t)] - \mathrm{Dim}[p(x_1^{t-1}, x_2^{t-1}|x_1^t, x_2^t)].$$

Tajima & Kanai (2017)

力学的解釈による実システムでの評価

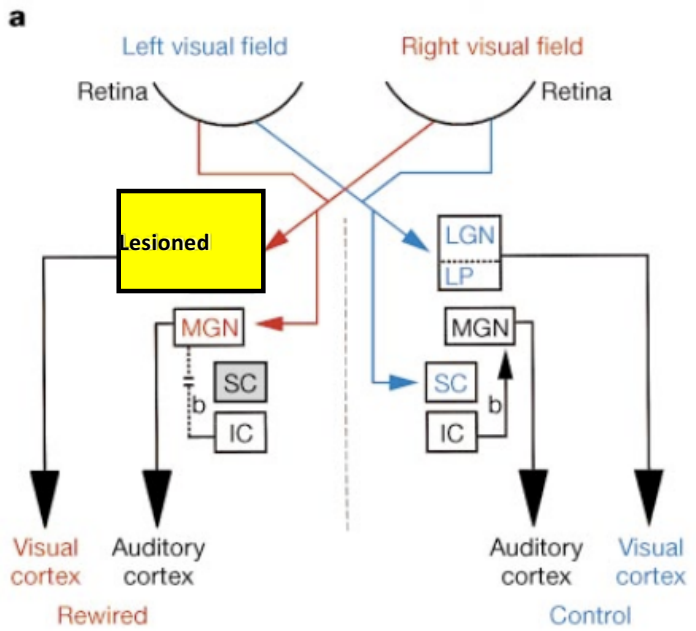


- We can recover the topology of attractor and estimate the dimensionality from partial observation.
- We speculate quality may be captured by topological characteristics.

Tajima & Kanai (2017)

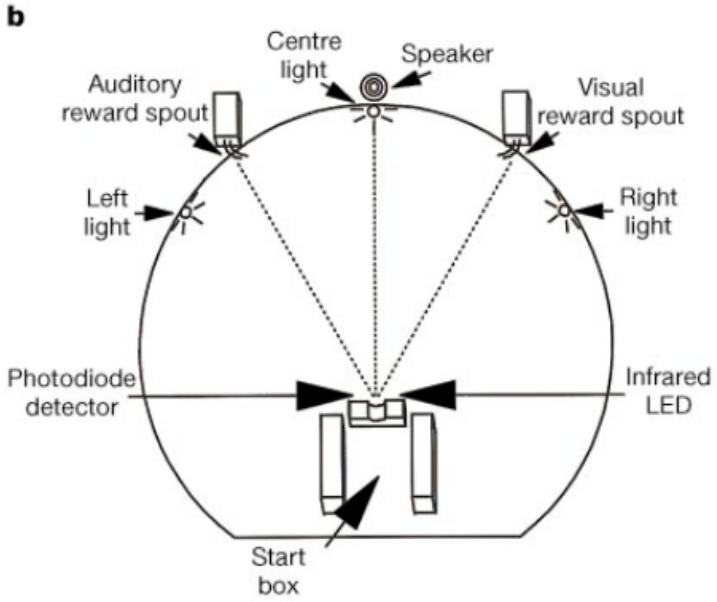
視覚クオリアと聴覚クオリアは
何が違うのか？

Modality and qualia



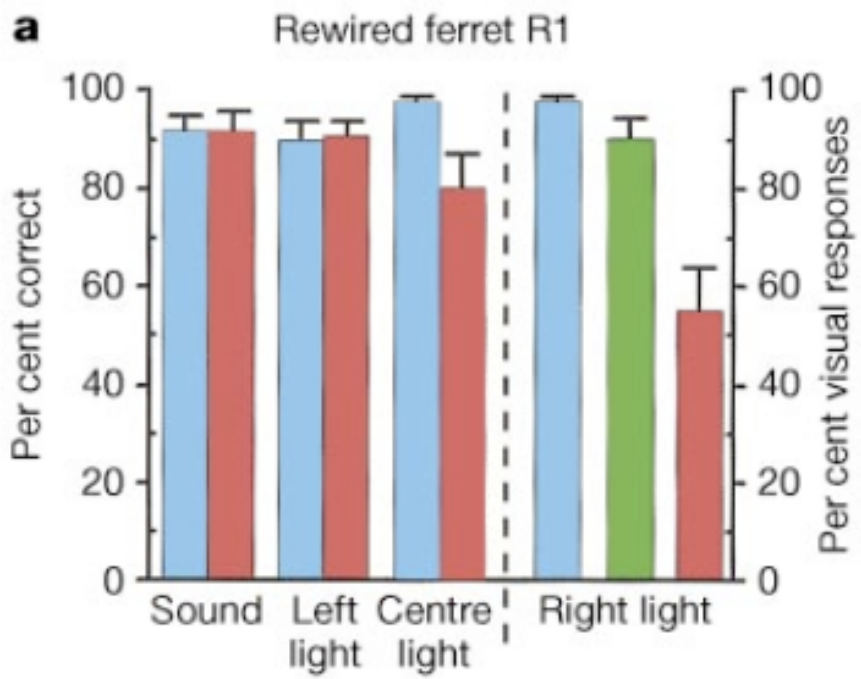
(von Melchner et al. 2000 *Nature*)

Modality and qualia



(von Melchner et al. 2000 *Nature*)

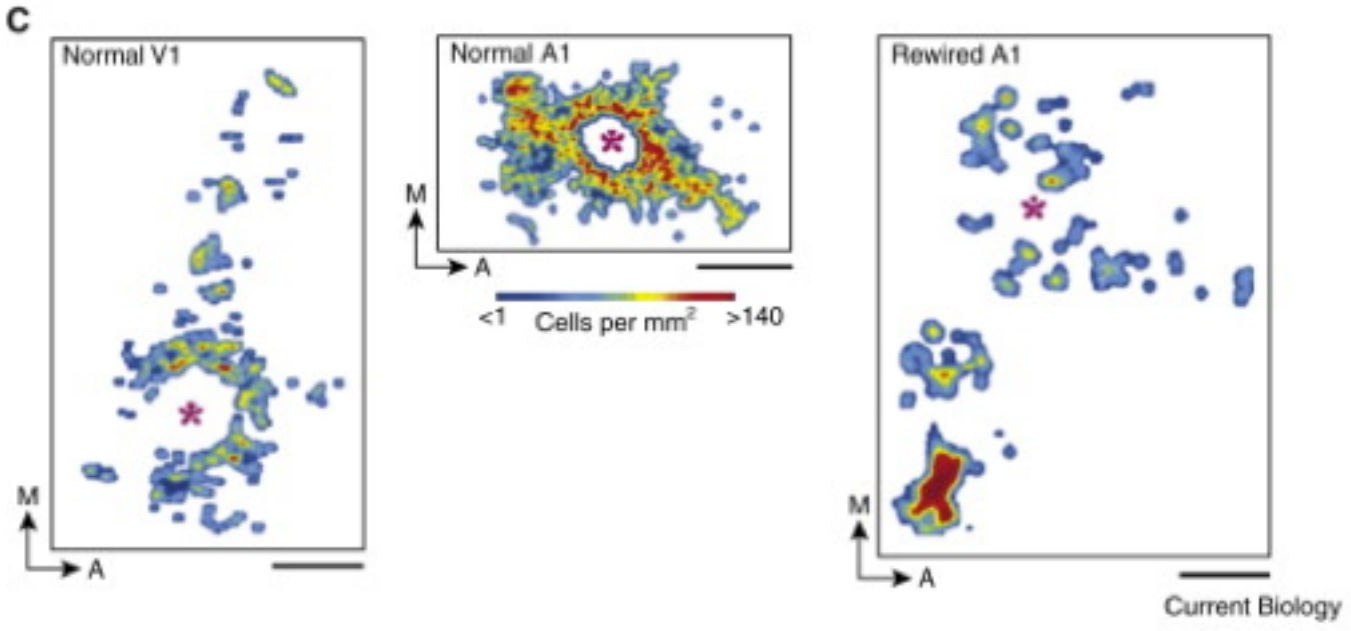
Modality and qualia



von Melchner et al. (2000) *Nature*

解剖的にはどうなっているか？

Connectivity change

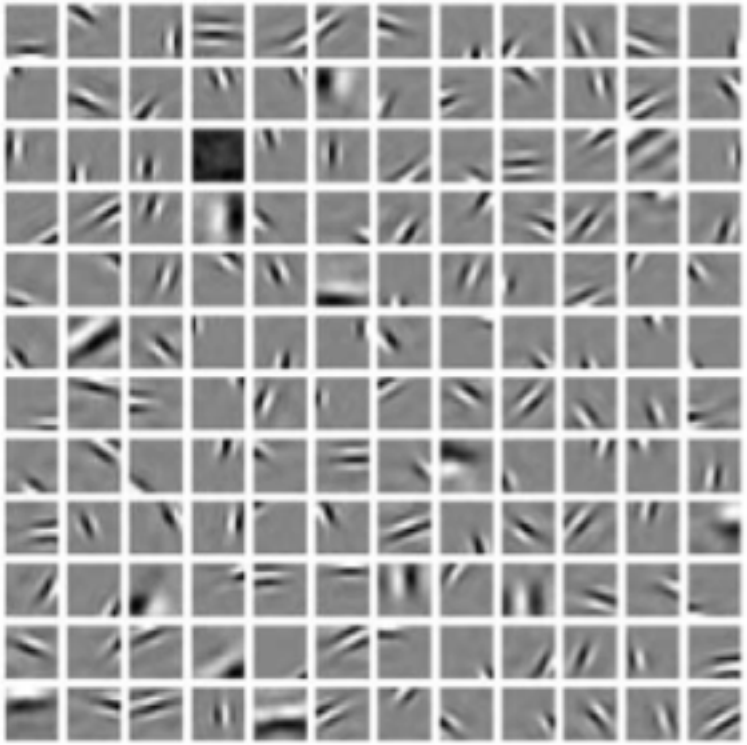


(modified from Sharma et al. 2000 *Nature*)

クオリアの解剖

- クオリアの質的な特徴は、マイクロサーキットのネットワークの特性に反映されているのではないか？
- 統合情報理論で、解剖的なネットワークトポロジーとモダリティ（視覚か聴覚化など）を特徴づけできるだろう。

視覚と聴覚の性質の違いの起源



視覚の統計的特徴

ARAYA

意識の機能とは？

V1仮説

- V1（第一次視覚野）の脳活動は、前頭葉に直接投射していないため、意識には上らない。

- この仮説は、意識の生物学的機能が次の2つという考察から来ている
  - 現在の視覚入力のベストな解釈を生み出すこと
  - その情報を自発的な行動や未来の行動計画に利用可能とすること

Crick & Koch (1995)

クオリア3原則

1.撤回不能性

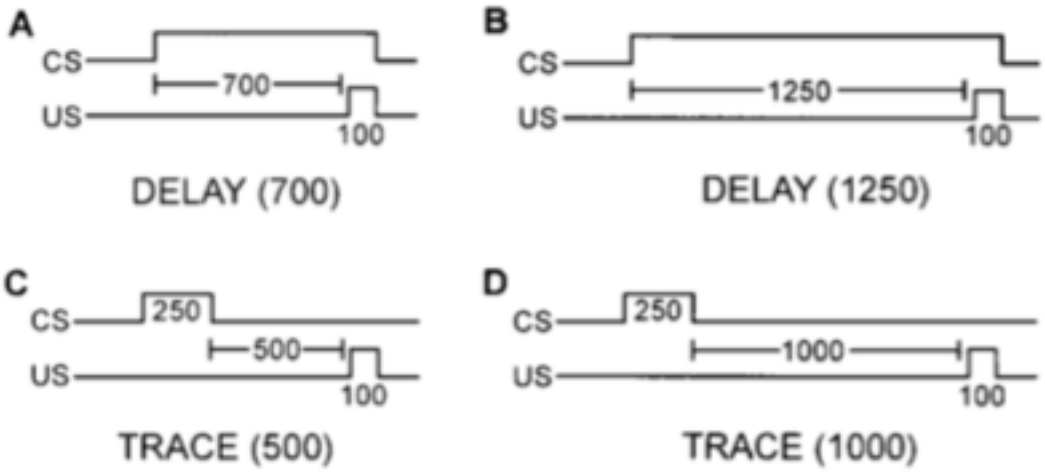
2.柔軟性

3.短期記憶性

知識からの隔離

非反射的行動

意識を必要とする学習

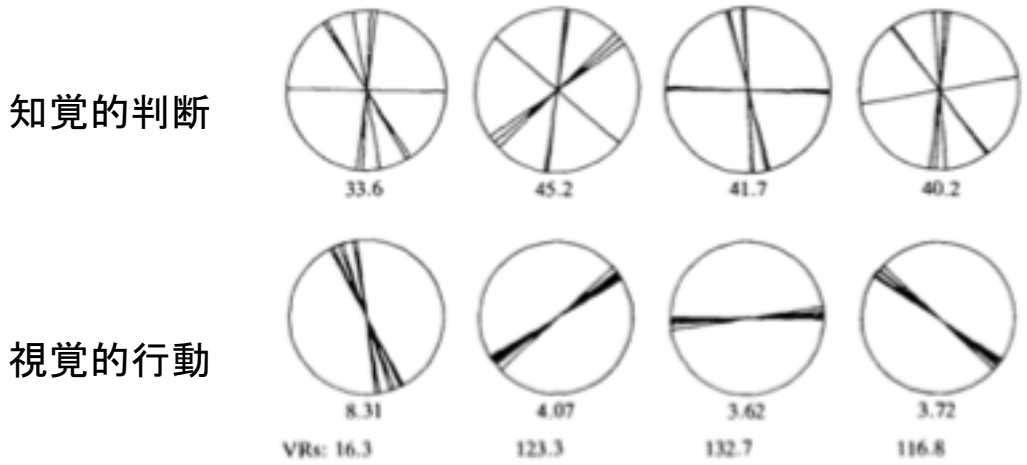


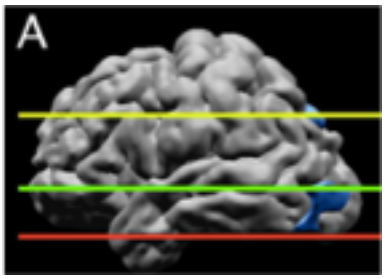
- 遅延条件付けには「気づき」は不要。（無意識で生じる）
- トレース条件付けには「気づき」が必要。

一種のクレジット・アサインメント問題を「意識」が解決

Clark & Squire (1998)

意識を必要とする行動





- オンラインでの目の前への反応は意識を必要としない。
- 短期記憶に基づくアクションは意識を要する。

Milner et al. (1991)

意識機能についての仮説

ARAYA

【反実仮想的情報生成理論】

意識の機能とは、現在の感覚入力と乖離した状況を感覚信号のデータフォーマットで生成し、それを未来の行動の計画などにりようすること。

逆に、自分の環境における状態の予測を生成することのできるシステムは意識を持つ。

Kanai (forthcoming)

反実仮想的情報生成理論

意識の機能= 現在の環境から分離した状況やイベントを表現すること

**その機能を持てば、未来の状況を内的なシミュレーションが可能になる。
そのためには、感覚運動ループの依存関係の学習が必要となる。**

- 意図・計画・想像
  - 反実仮想的な自分の行動の結果についての予測
- 非反射的行動
  - 現在の外的刺激によってトリガーされる行動以外が可能となる
- 短期記憶性
  - 過去の情報の保持も反実仮想の1つである

意識と知性の進化的起源

ARAYA

【知性の進化ステージ】

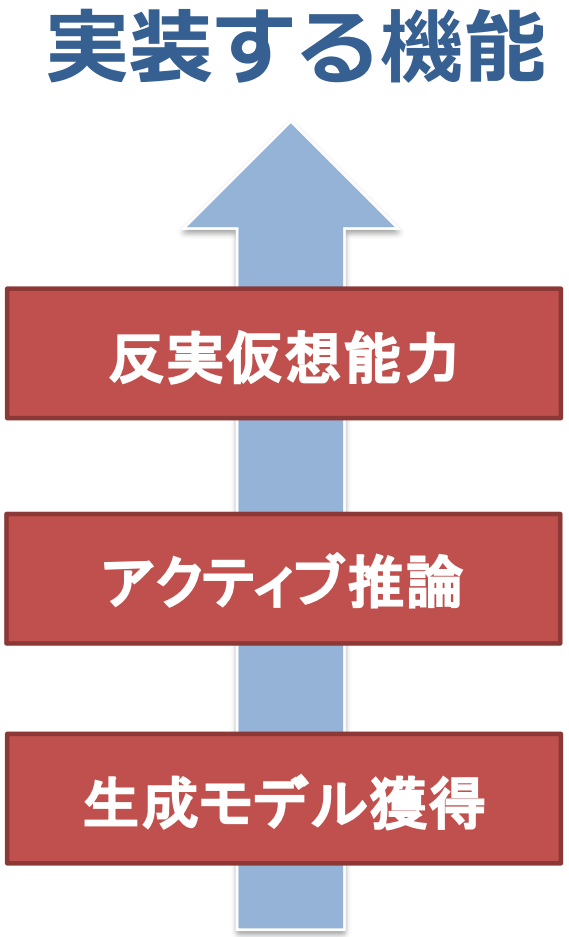
1. ダーウィン型生物
2. スキナー型生物
3. ポパー型生物
4. グレゴリー型生物

ARAYA

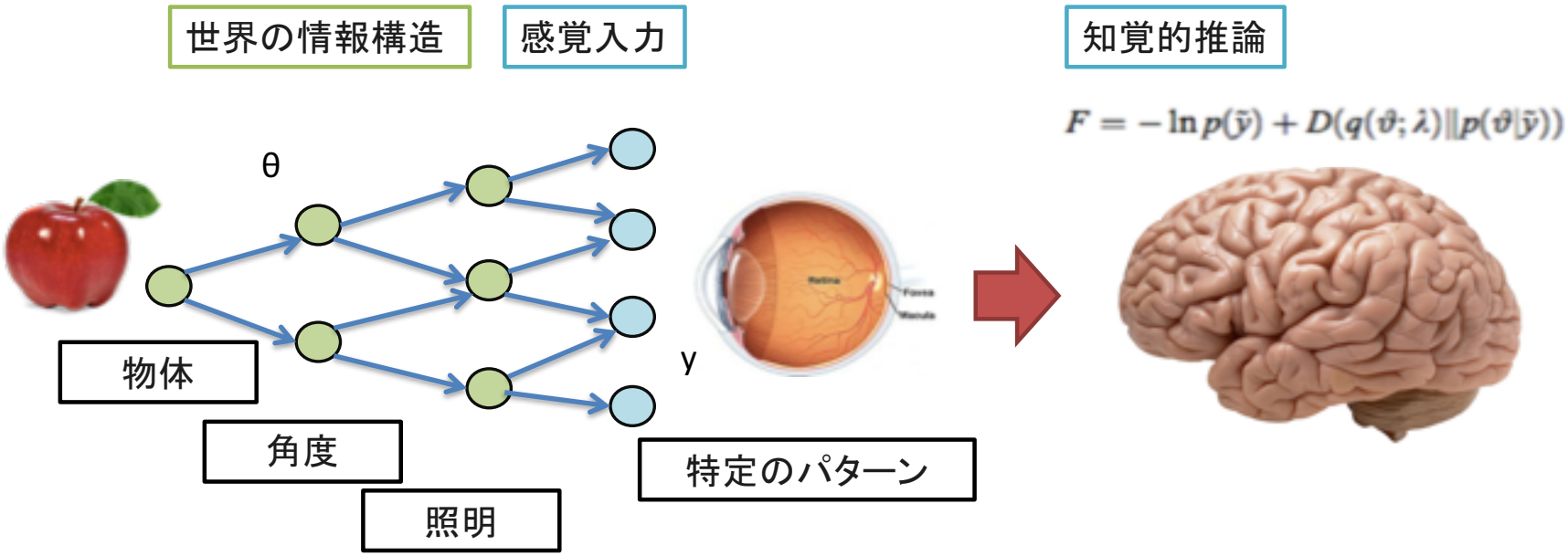
反実仮想の実装

意識の作り方：３つのステップ

ARAYA



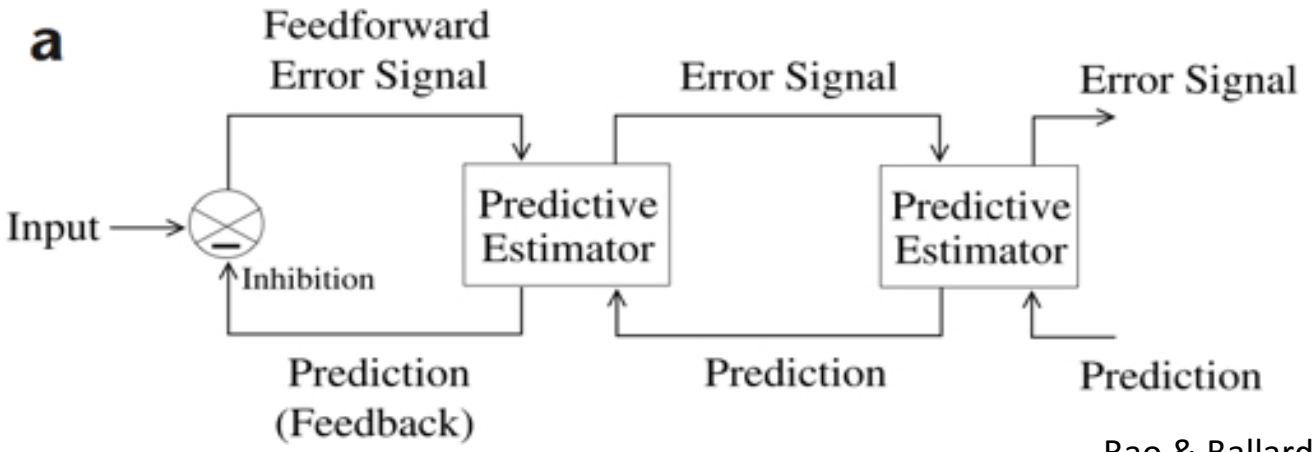
意識の作り方①：生成モデルの獲得



実現される意識機能

物体認識などの知覚系の確立（Deep Learning）
脳が自己の予測モデルを獲得 (cf. Cleeremans)

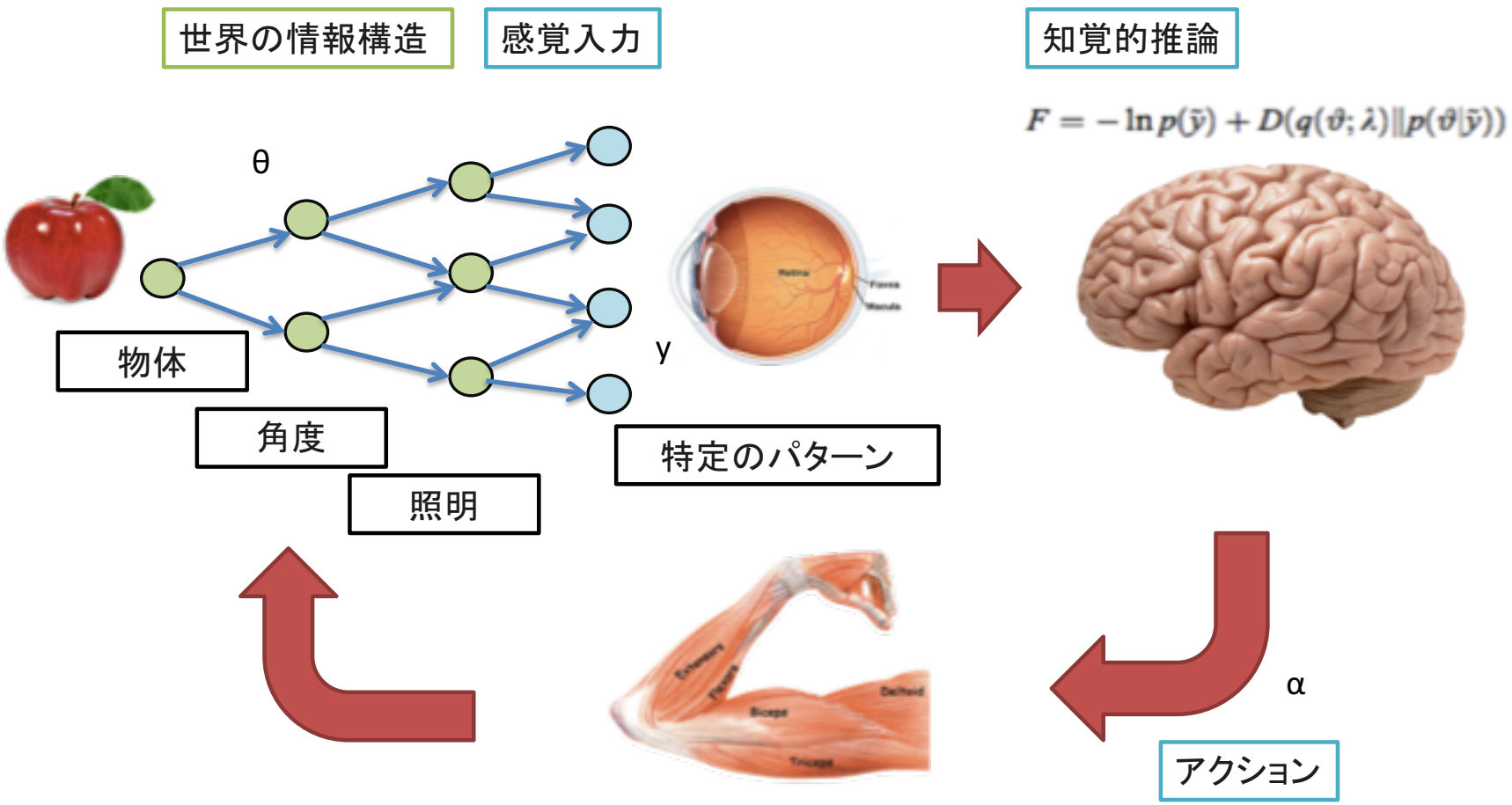
脳内での予測符号化理論



Rao & Ballard (1999)

$$E_1 = \frac{1}{\sigma^2}(\mathbf{I} - f(\mathbf{U}\mathbf{r}))^{\mathrm{T}}(\mathbf{I} - f(\mathbf{U}\mathbf{r})) + \frac{1}{\sigma_{td}^2}(\mathbf{r} - \mathbf{r}^{td})^{\mathrm{T}}(\mathbf{r} - \mathbf{r}^{td})$$

意識の作り方②：アクティブ推論

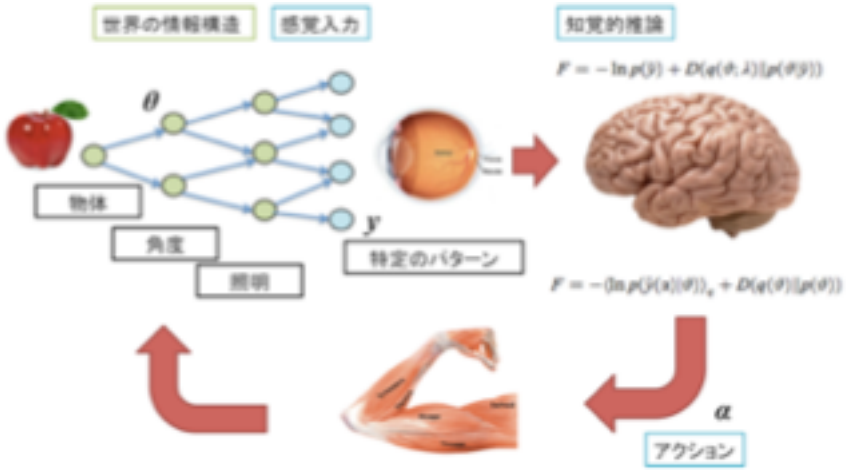


意識の作り方②：アクティブ推論

ARAYA

知覚での予測を確認するように
行動が引き起こされる

知覚と行動を統一したモデルが構築される



アクティブ推論によって実現される意識機能

世界と自己との関係についてのモデルが獲得される

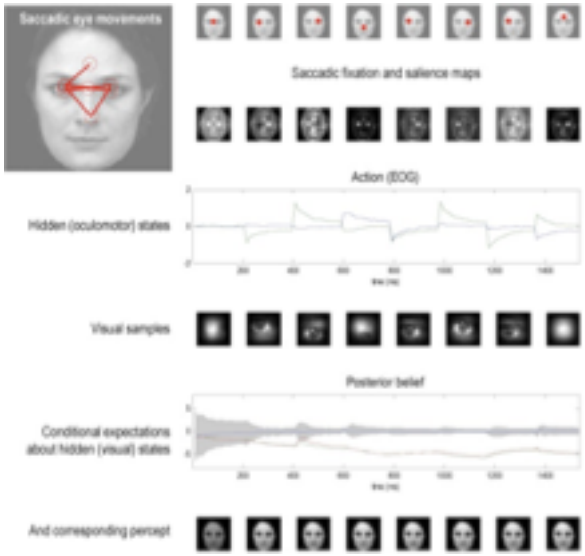
The Dark Room Problem

反実仮想への拡張：Saccade as experiments

ORIGINAL RESEARCH ARTICLE
Front. Psychol., 28 May 2012 | http://dx.doi.org/10.3389/fpsyg.2012.00151

Perceptions as hypotheses: saccades as experiments

Karl Friston[1]*, Rick A. Adams[1], Laurent Perrinet[1,2] and Michael Breakspear[3]



Information Gain

次の視線の方向を、潜在変数のエントロピーが下がるように選択する。

$$\tilde{\eta}_u(t) = \arg\max_{\tilde{\eta}_j} S(\tilde{\eta}_j)$$

$$S(\tilde{\eta}_j) = -H[q(\tilde{\psi} \mid \tilde{\mu}_x(t+\tau), \tilde{\mu}_v(t+\tau), \tilde{\eta}_j)]$$

$$= \tfrac{1}{2} \ln \left| \partial_{\tilde{\psi}} \tilde{\varepsilon}_j^T \Pi_\omega \partial_{\tilde{\psi}} \tilde{\varepsilon}_j + \Pi_\psi \right|$$

意識の作り方③：反実仮想能力

ARAYA

仮説：意識の機能＝現実と切り離した世界を想像する力

世界と自己との関係についてのモデルが獲得されるとそれを用いた
シミュレーションが可能となる。

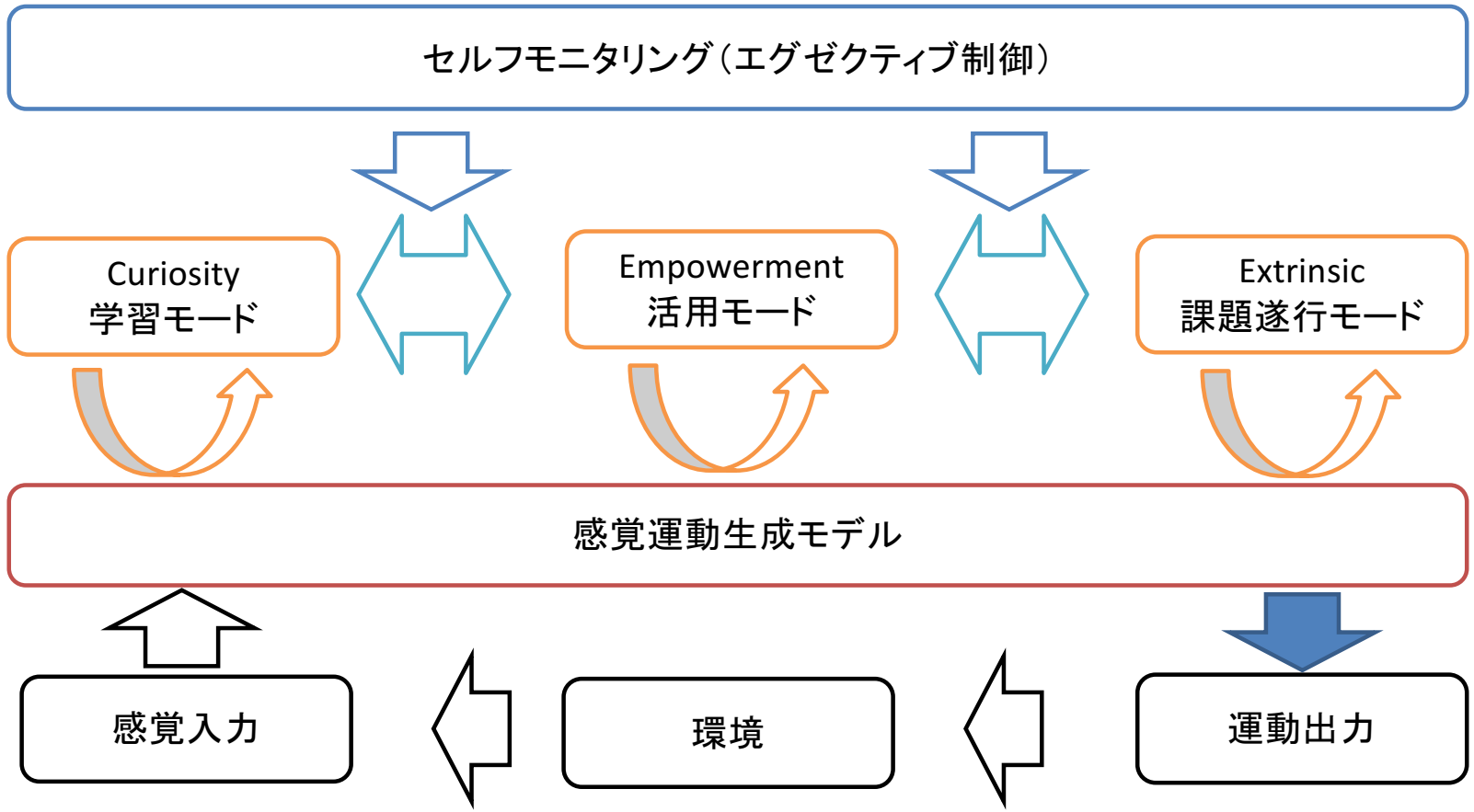
反実仮想能力によって生まれる機能

- 意図
- 思考
- 想像
- 夢

意識アーキテクチャの構想

ARAYA

意識アーキテクチャの実装



Computer Science > Learning

Counterfactual Control for Free from Generative Models

Nicholas Guttenberg, Yen Yu, Ryota Kanai

(Submitted on 22 Feb 2017 (v1), last revised 9 Mar 2017 (this version, v2))

We introduce a method by which a generative model learning the joint distribution between actions and future states can be used to automatically infer a control scheme for any desired reward function, which may be altered on the fly without retraining the model. In this method, the problem of action selection is reduced to one of gradient descent on the latent space of the generative model, with the model itself providing the means of evaluating outcomes and finding the gradient, much like how the reward network in Deep Q-Networks (DQN) provides gradient information for the action generator. Unlike DQN or Actor-Critic, which are conditional models for a specific reward, using a generative model of the full joint distribution permits the reward to be changed on the fly. In addition, the generated futures can be inspected to gain insight in to what the network 'thinks' will happen, and to what went wrong when the outcomes deviate from prediction.

Comments:	6 pages, 5 figures
Subjects:	**Learning (cs.LG)**; Machine Learning (stat.ML)
MSC classes:	68T05
Cite as:	arXiv:1702.06676 [cs.LG]
	(or arXiv:1702.06676v2 [cs.LG] for this version)

Submission history

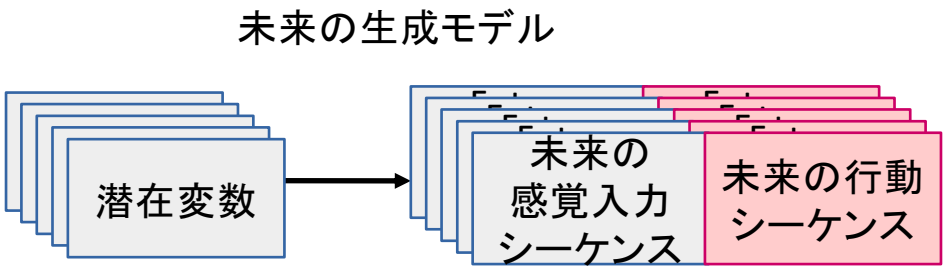
From: Nicholas Guttenberg [view email]
[v1] Wed, 22 Feb 2017 04:50:47 GMT (483kb,D)
[v2] Thu, 9 Mar 2017 06:35:45 GMT (483kb,D)

生成モデルで柔軟性のあるエージェントを構築

生成コントロール:

- **可能な未来の分布を生成する**

この分布はアクションを含む → 自分に好ましい未来を生成する行動探索に利用可能

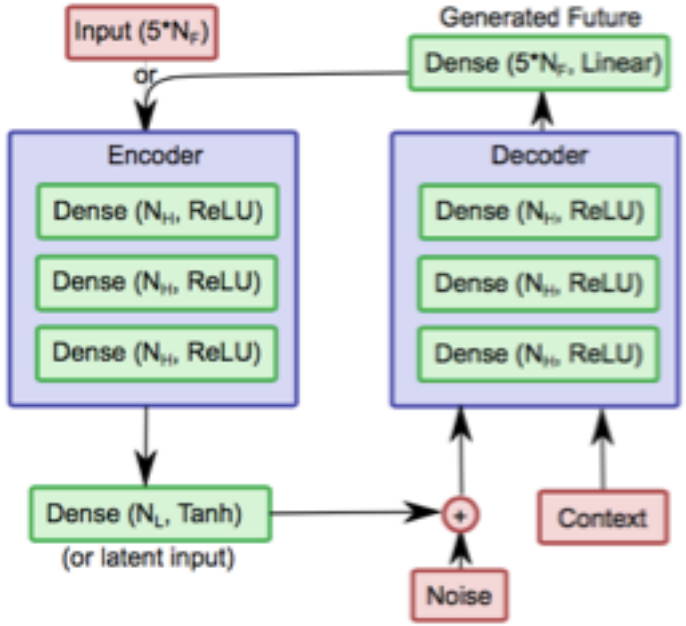


得られる機能的メリット

- 報酬を課題中に自在に変更しても対応できる。
- 報酬として不確かさやサプライズを設定することが可能（内発的動機づけに利用可能）
- 長期的な行動計画に利用可能（逐次的な予測モデルでは難しい）

Guttenberg, Yu, Kanai (2016)

柔軟性のあるエージェント



(Arulkumaran, Creswell, Bharath 2016)

新しい目標に自在に対応可能
0ショット学習も成功している

Guttenberg, Yu, Kanai (2016)

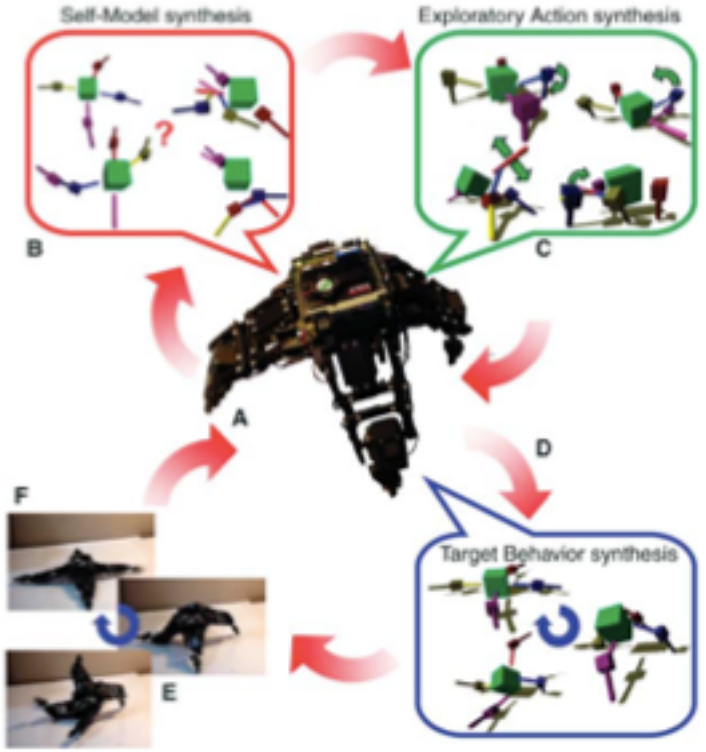
自己モデルを持つ反実仮想ロボット

ARAYA

Resilient Machines Through Continuous Self-Modeling

Josh Bongard,[1]*† Victor Zykov,[1] Hod Lipson[1,2]

Animals sustain the ability to operate after injury by creating qualitatively different compensatory behaviors. Although such robustness would be desirable in engineered systems, most machines fail in the face of unexpected damage. We describe a robot that can recover from such change autonomously, through continuous self-modeling. A four-legged machine uses actuation-sensation relationships to indirectly infer its own structure, and it then uses this self-model to generate forward locomotion. When a leg part is removed, it adapts the self-models, leading to the generation of alternative gaits. This concept may help develop more robust machines and shed light on self-modeling in animals.



Bongard , Zykov & Lipson (2006)

Computer Science > Artificial Intelligence

Learning body-affordances to simplify action spaces

Nicholas Guttenberg, Martin Biehl, Ryota Kanai

(Submitted on 15 Aug 2017)

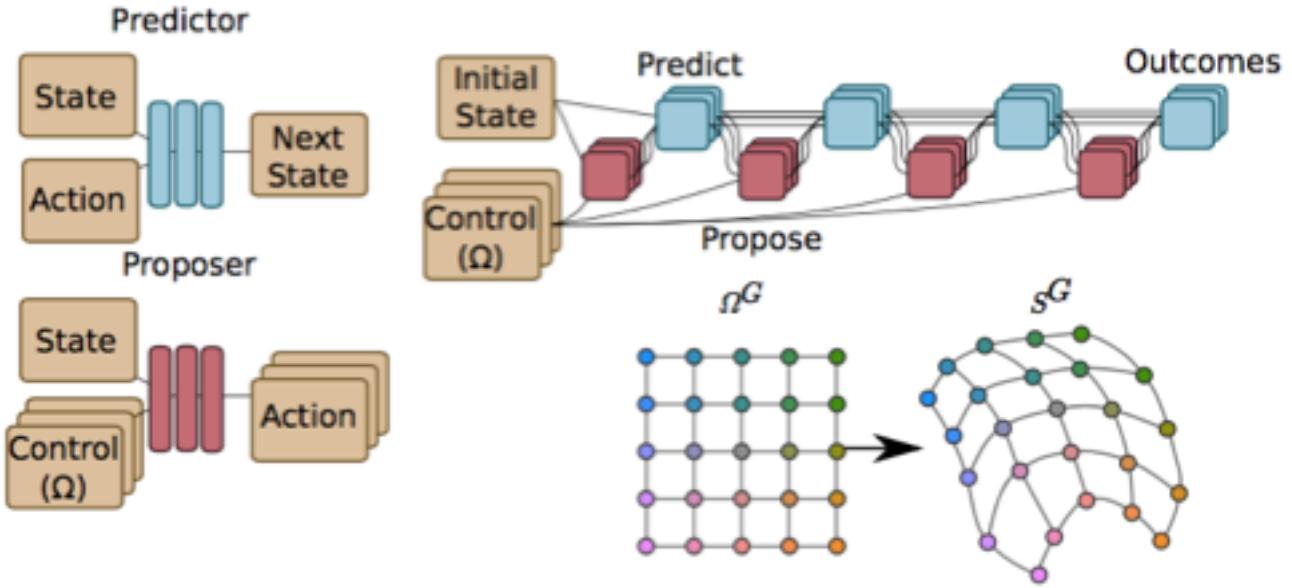
Controlling embodied agents with many actuated degrees of freedom is a challenging task. We propose a method that can discover and interpolate between context dependent high-level actions or body-affordances. These provide an abstract, low-dimensional interface indexing high-dimensional and time- extended action policies. Our method is related to recent ap- proaches in the machine learning literature but is conceptually simpler and easier to implement. More specifically our method requires the choice of a n-dimensional target sensor space that is endowed with a distance metric. The method then learns an also n-dimensional embedding of possibly reactive body-affordances that spread as far as possible throughout the target sensor space.

Comments: 4 pages, 4 figures
Subjects: **Artificial Intelligence (cs.AI)**; Robotics (cs.RO)
Cite as: arXiv:1708.04391 [cs.AI]
(or arXiv:1708.04391v1 [cs.AI] for this version)

Submission history

From: Nicholas Guttenberg [view email]
[v1] Tue, 15 Aug 2017 04:07:57 GMT (372kb,D)

逆モデルの学習



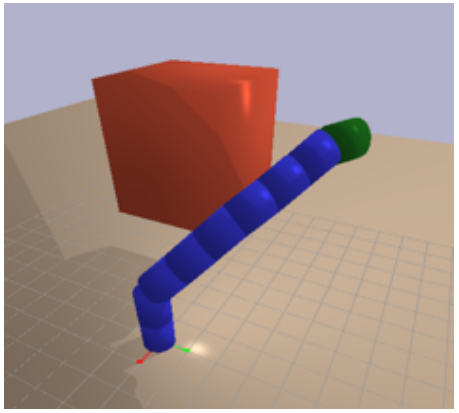
センサー空間で表現される行動のアウトプットを低次元のコントロール空間に
埋め込むためのネットワークを構築した。

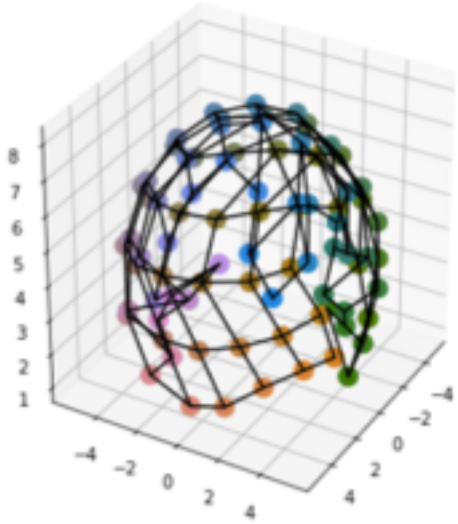
Guttenberg, Biehl, Kanai (2017)

逆モデルの学習

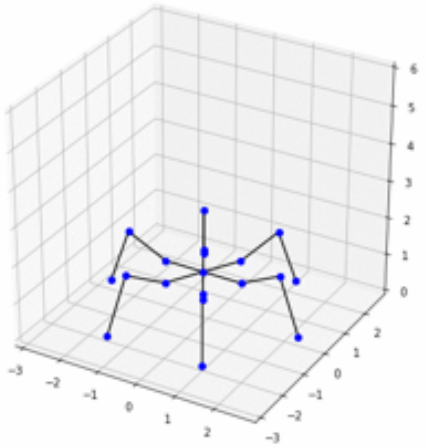
ARAYA

Reacher





Hexapod



- 自由度の縮約は強化学習の効率を向上させる
- このような行動の結果を表現することで、反実仮想的に利用することでより意図に近いものを構築することができる。

Guttenberg, Biehl, Kanai (2017)

まとめ

ARAYA

- **情報生成理論**
  - 意識の機能的側面（非反射的行動や自らの行動予測など）の特徴をうまく捉えている。
  - 意識の生物学的な知見とも合致している。

- **汎用性との関連**
  - 生成モデルは多様な課題でポリシーを見つけるために利用可能。
  - 複数の生成モデルを効率的に繋いで問題を解くためのメタなネットワークが必要。
  - 複数の生成モデルをつなぐことで「意味」は生まれるのではないか（マルチモーダル）？

- **現象的意識は生まれるのか**
  - 意識の根底にある物理的な因果関係を再構築できれば、現象的意識はそこに生じるはずである哲学的立場（生物学的自然主義）をとると、情報生成理論に基づいて構築したエージェントは意識を持つ。