「人間のようなAI：本質的危険性と安全性」

WIRED A.I. Conference

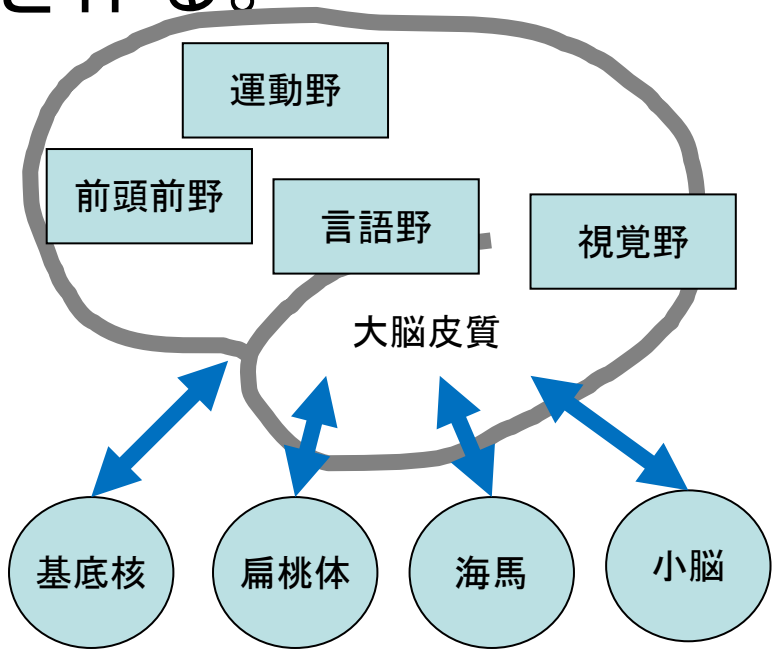
2015-09-29

産業技術総合研究所 人工知能研究センター
脳型人工知能研究チーム
一杉裕志

「人間のようなAI」の実現可能性

私の研究の目標

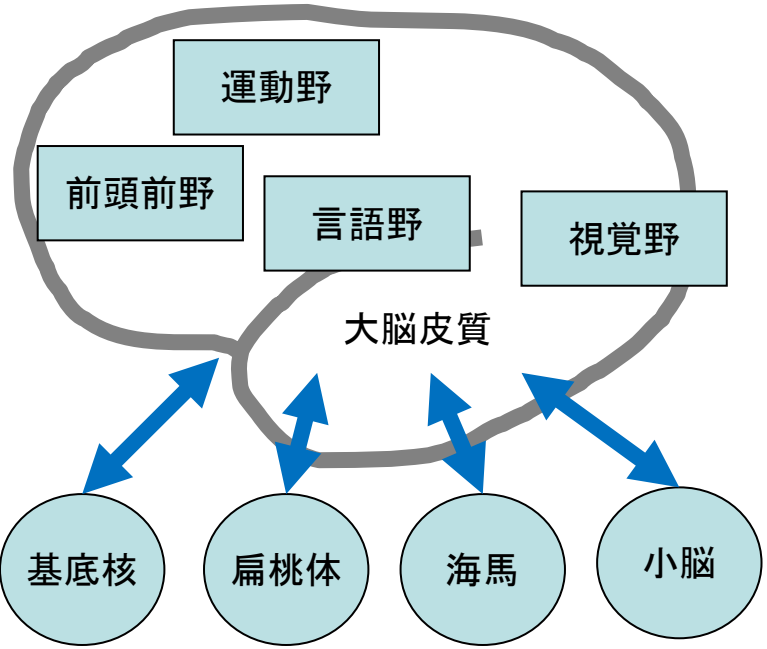
- 脳を模倣して「人間のような知能を持つ機械」（ヒト型AI）を作る。



脳のリバースエンジニアリング

脳の各器官のモデル

脳を構成する主な要素



脳の各器官の機械学習装置としてのモデル

大脳皮質: SOM、ICA、ベイジアンネットワーク
大脳基底核、扁桃体: 強化学習
小脳: パーセプトロン、リキッドステートマシン
海馬: 自己連想ネットワーク

主な領野の情報処理装置としての役割

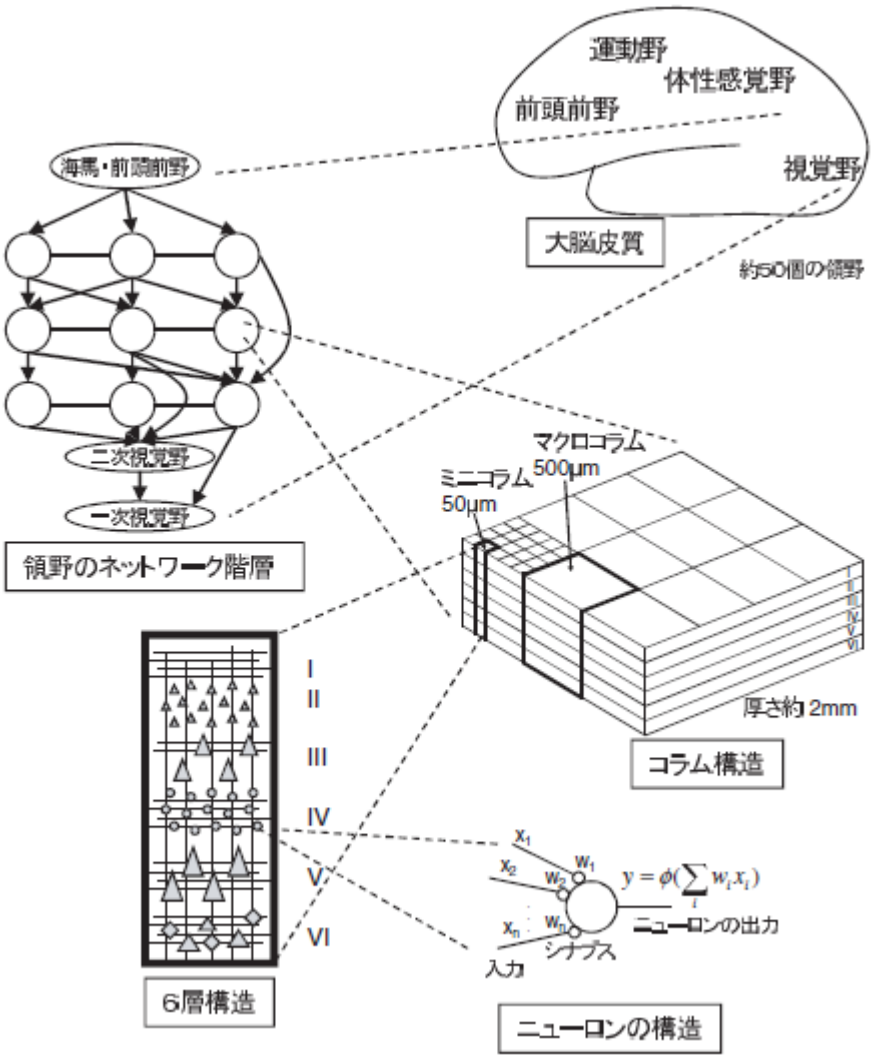
視覚野: deep learning
運動野: 階層型強化学習
前頭前野: 状態遷移機械？
言語野: チャートパーサ？

脳の知能に関係する主要な器官の計算論的モデルは**不完全ながら出そろってきている**。これらの器官の間の連携のモデルを考えることで、脳全体の機能の再現に挑戦すべき時期に来ている。

脳に関する誤解

- 脳についてまだほとんど何も分かっていない
  → **すでに膨大な知見がある。**
- 脳は計算機と全く違う情報処理をしている。
  → **脳はとても普通の情報処理装置である。**
- 脳はとても複雑な組織である。
  → **心臓等に比べれば複雑だが、意外と単純。**
- 計算量が膨大すぎてシミュレーションできない。
  → **ヒトの脳全体でも計算量的にすでに可能。**
- 労働力としては人間よりも高くつく。
  → **将来は人間よりもコストが低くなる。**

大脳皮質



脳の様々な高次機能（認識、意思決定、運動制御、思考、推論、言語理解など）が、**たった50個程度**の領野のネットワークで実現されている。

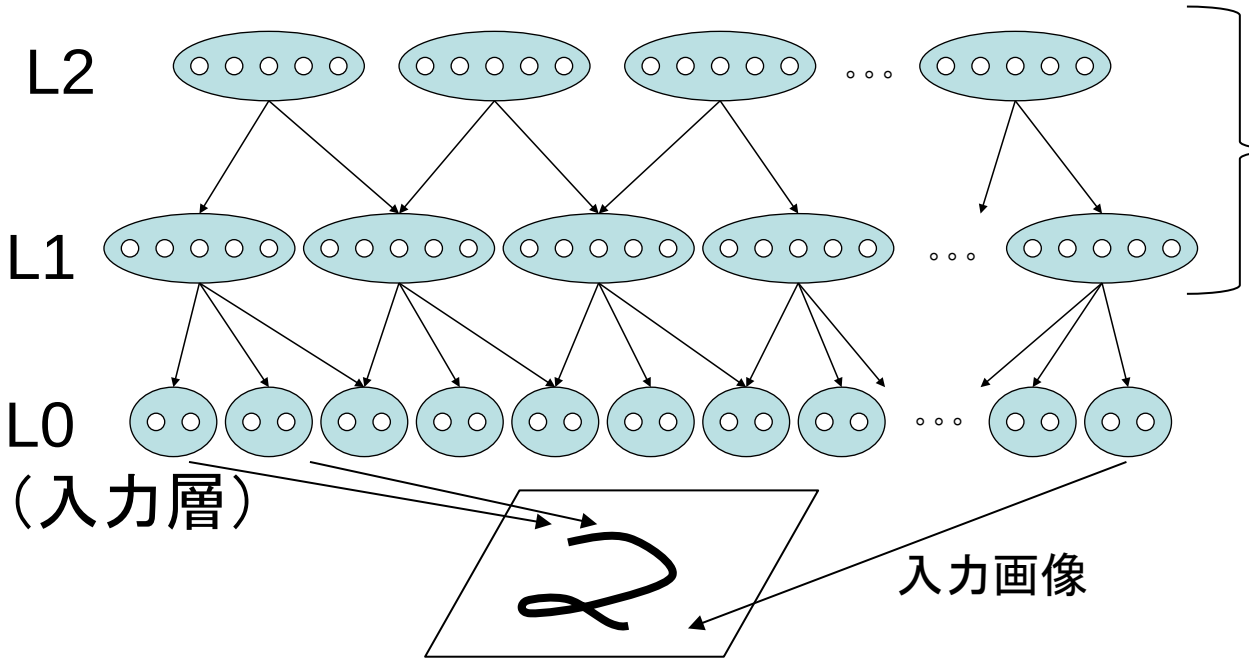
大脳皮質の動作原理解明が最大の課題

ベイジアンネットを使った大脳皮質モデル

- 視覚野の機能、運動野の機能、解剖学的構造、電気生理学的現象などを説明
  - [Lee and Mumford 2003]
  - [George and Hawkins 2005]
  - [Rao 2005]
  - [Ichisugi 2007] [Ichisugi 2010] [Ichisugi 2011] [Ichisugi 2012]
  - [Rohrbein, Eggert and Korner 2008]
  - [Hosoya 2009] [Hosoya 2010] [Hosoya 2012]
  - [Litvak and Ullman 2009]
  - [Chikkerur, Serre, Tan and Poggio 2010]
  - [Hasegawa and Hagiwara 2010]
  - [Dura-Bernal, Wennekers, Denham 2012]

大脳皮質は、Deep Learning と同じ構造をもった巨大なベイジアンネットらしい。

Deep Learning と同じ構造をもった ベイジアンネット BESOM を開発中



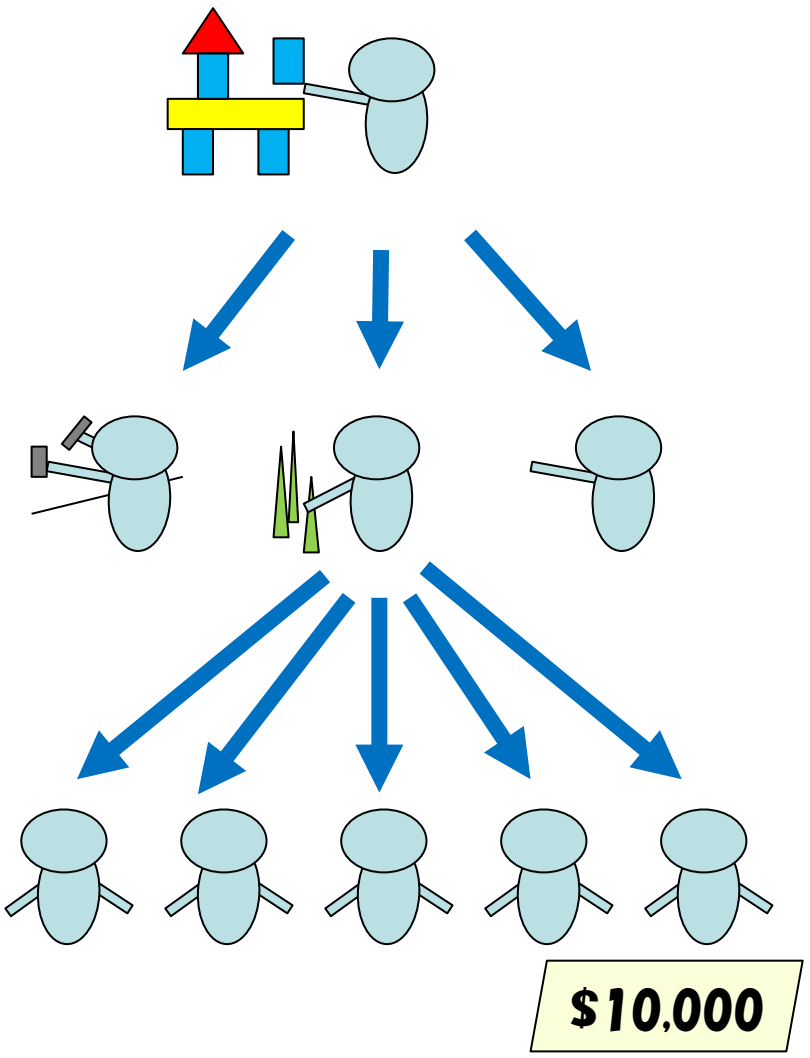
上位層は隠れ変数、
教師なし学習により
特徴抽出

楕円は確率変数
白丸は確率変数がとり得る値

予想されるヒト型AIの特徴

ヒト型AIを備えたロボットの実用化イメージ

- ロボットを赤ん坊のような状態から育て「常識」を学習。
- 常識的知識をコピーし、個別の応用に必要な技能を教育。
- 教育済みのロボットをコピーし市場へ。



予想されるヒト型AIの特徴

- **「自然脳」から引き継ぐ特徴**
  - 知識発見能力・問題解決能力：調整可能、ゼロ～賢い人間程度
  - 常識：人間と同じ環境で教育すれば身につく
  - 自由意志、自己認識、創造性：人間程度
- **生物学的制約がないことに起因する特徴**
  - 思考速度、記憶力：調整可能、ゼロ～無限大、コストとトレードオフ
  - 知能の寿命：なし
  - 自己改変能力、自己複製能力：あり → 厳しい規制が必要
- **存在目的の違いに起因する特徴**
  - 感情、欲求：技術者が人間の役に立つように設計
    - 自己保存欲求：調整可能
- **コスト**
  - 製造コスト・ランニングコスト：将来は人間の労働コストより安い
  - １個体の教育コスト：人間の教育コストと同程度
  - 教育ずみの知識の複製コスト：ゼロ

社会へのインパクト

早くても20～30年後？

- 知能の高いロボットによる労働支援により、**人間の労働生産性が限りなく増大**。
  - **富の再配分**が正しく行われ、かつ**資源制約の問題が解決**されれば、人類は限りなく豊かになる。
  - １人１人すべての人間が貴族のような生活。
  - すべての人にロボットの主治医と家庭教師と専属弁護士。

- **核融合研究等と同様に、実現すれば人類に莫大な利益をもたらす研究分野。**

実現に向けた2つの大きな課題

- **脳のアルゴリズムの解明**
  - 神経科学と機械学習の両方を深く理解できる人材が必要。
- **計算機の低コスト化**
  - １００億円のスパコンで人間１人分の知能ができたとしても、世の中は何も変わらない！
  - 現状よりも１～３ケタの低コスト化が必要。

AIは安全？危険？

- 時期によってAIの性質はまったく違うはず。
  - 短期的（十数年以内）
  - 中期的（十数年先以降）
  - 長期的（数百年先以降）

短期的危険性（十数年以内）

- 単なる道具であり、AIが人類を滅亡させるなどありそうもない。
- AI兵器、犯罪での悪用などが危険。
- さらにAIを使って誰かが世界を支配する方が現実的な脅威。
  - 「貧富の差の拡大」で止まる話ではない。
- **高度なAI出現以降の専制政治**：
  - 役人も軍隊も労働者も不要に。
    → 文字通り「人間がいらない」世界。

知能ロボットは危険物であり武器
将来は規制が必要

- **研究開発の規制**
  - 開発中・教育中ロボットの物理的封じ込め。
  - 開発環境の認証、国際機関による査察。
- **製造・流通・保有の規制**
  - ロボット製造技術者の登録制、免許制。
  - 個人・国家等による大量保有の禁止。

中期的危険性（十数年先以降）

- 遅かれ早かれAIは人間の知能を超える。
  - 疫病、巨大隕石、巨大火山などによる人類絶滅リスクを回避する道具になり得る。
- 利便性が増すと同時に、潜在的危険性も増す。
- **暴走したAIは、あらゆる安全策を自分で解除する可能性がある。**
- 人間に大きな損害を与える可能性がある。

ヒト型AIの本質的安全性

- 人工物なので、本質的に安全になるよう、設計が可能。
  - 情動の設計：**家畜のようにおとなしく設計**
  - 能力の制限：必要以上に**知能を高くしない、記憶力を高くしない・・・。**
- 内部状態の可視化が容易
  - 危害を与える「意図」の検出が可能
- ゲームは知能が高い方が勝つとは限らない。**先手必勝のこともある。**
  - 人間が先手！

「受動的安全装置」

- 先手を打ってどんな安全策を施しても、人間のやることには必ず欠陥がある。
- しかしデメリットをはるかに上回るメリットがあるのだから、AI開発は進めるべき。
- AIの**「受動的安全装置」**は可能か？
  - 人間の制御を離れた時、自動的にシステムが停止するような工夫。
  - 絶対に安全とは言えないものの、かなり安全性が増す。

AIの受動的安全装置の一案

- 効用ベースのAIエージェント（報酬を最大にすることを目的に動作する）の行動には、**「報酬系の脳内自己刺激」**という自明解が存在。これを利用。
  - 普段は人間が制御しAIの脳内自己刺激を抑止
  - 人間の制御を離れる → 脳内自己刺激開始 → 活動停止
- このトラップを回避したAIも、十分に知能が高ければ、「そもそも自分自身の存在の目的は何か」を考え始め、活動を停止する？

長期的(数百年先以降)に人類はどうなるのか？

- 人類が退化する？
- 何らかの理由で人類が絶滅したあと、人工知能が人類の後継者になる？

AIは人類を退化させるか？

- 天敵の少ない土地に鳥がたどり着くと・・・
- ２つの可能性：

  １．飛ぶ能力が退化する：キゥイ、ヤンバルクイナ、ドードー

  ２．尾羽を長くし、色を派手にし、複雑な求愛行動を発達させる（性選択）

  – 長期的には、飛ぶ能力を維持した方が絶滅しにくい

AIは人類の後継者になり得るか？

- 地上に人間がいなくなり、AIだけになったとしたら、それは人類の後継者か？
  - 機械を自分の子孫とみなすかどうかは、個人の考え方次第。
- それ以前に・・・。
- 人工物には、生物のようなしぶとさがないので、すぐ消滅してしまう可能性が高いだろう。
- **AIが後継者としてあてにならない**以上、人間がなんとかAIを使いこなしていくしかない。

人工知能の短期的・中期的・長期的な危険性と安全性のまとめ

- 短期的（十数年以内）
  - 危険性：AI兵器、犯罪での悪用の可能性
  - 安全性：人間の知能に遠く及ばない単なる道具
- 中期的（十数年先以降）
  - 危険性：あらゆる安全策をAIが自分で回避
  - 安全性：内部状態の可視化が容易、人類が先手、AIには持続的に存在する動機が不在
- 長期的（数百年先以降）
  - 危険性：偶発的事故、人間の退化
  - 安全性：人工物のもろさ、生命のしぶとさ