ヒト型AIは人類に
どのような影響を与えうるか

シンギュラリティ・サロン 東京 第3回講演会

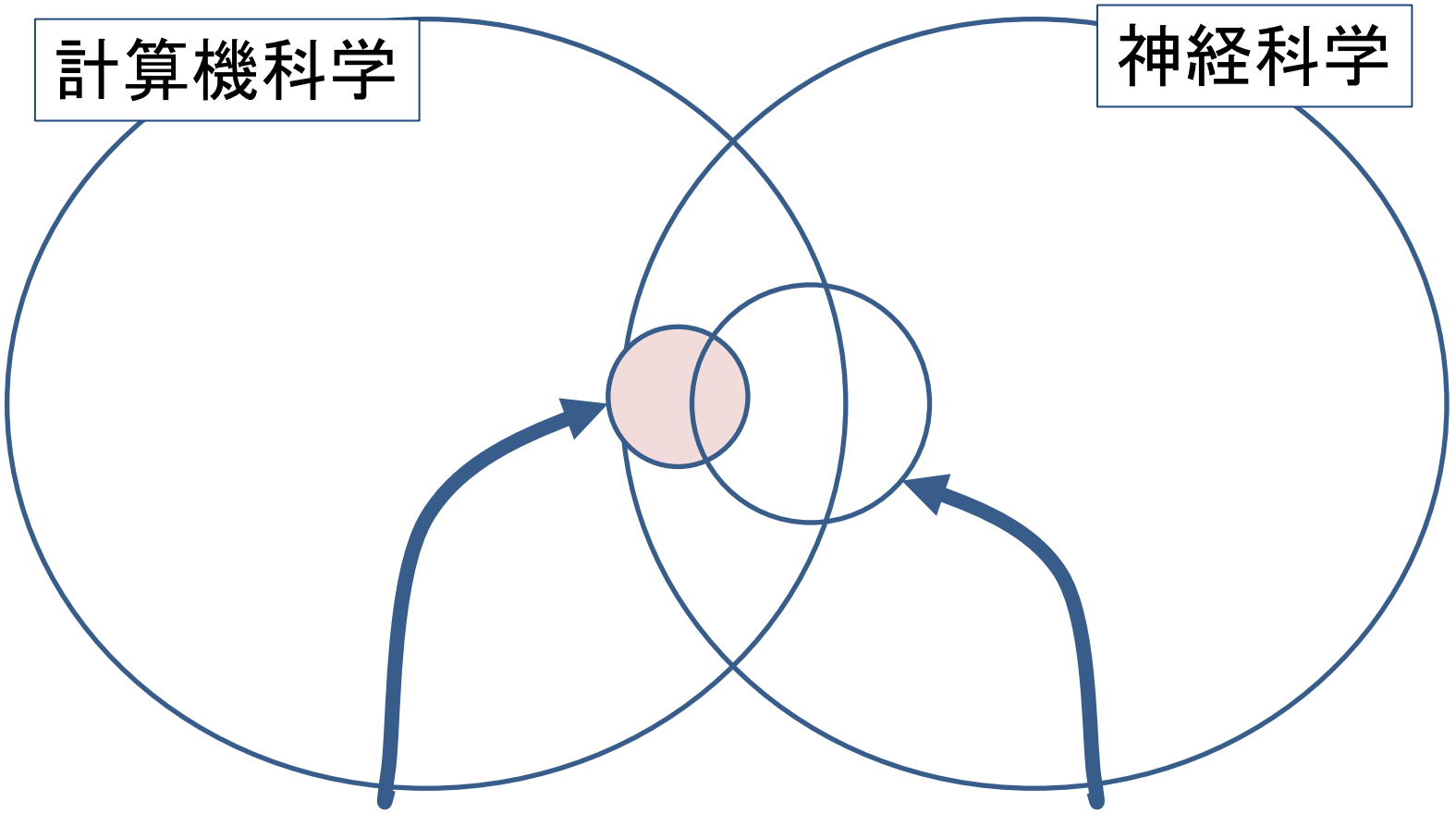
2015-10-17

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

本講演の内容は主に下記の記事をもとにしています。

一杉裕志、「ヒト型AIは人類にどのような影響を与え得るか」
人工知能:人工知能学会誌 29(5), 507-514, 2014.

計算論的神経科学



脳に学んで
人間のような知能を作る
という研究テーマ

計算論的神経科学

脳を情報処理装置に見立てて
理解する学問

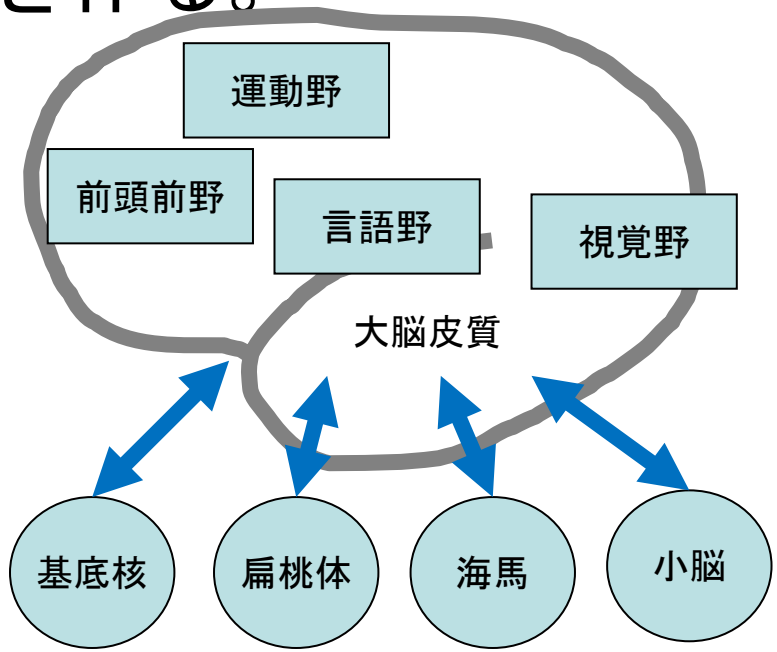
自己紹介

- 1990年東京工業大学大学院情報科学専攻修士課程修了。米澤研究室。
- 1993年東京大学大学院情報科学専攻博士課程修了。博士(理学)。
- 同年電子技術総合研究所入所。
  - 並列言語、拡張可能言語、オブジェクト指向言語のモジュール機構、スクリプト言語等を研究。
- 2001年より産業技術総合研究所に改組。
- 2005年より計算論的神経科学を研究。

ヒト型AIの実現可能性

私の研究の目標

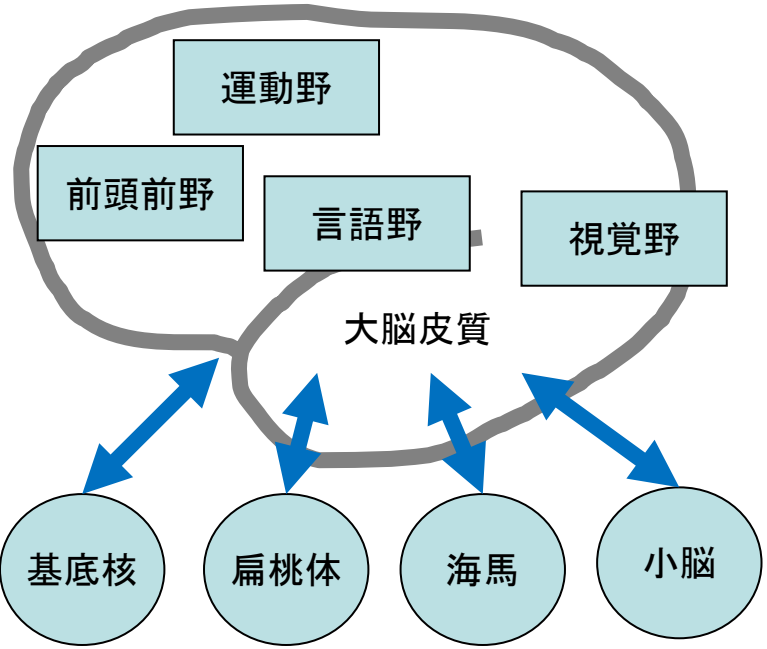
- 脳を模倣して「人間のような知能を持つ機械」（ヒト型AI）を作る。



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。

脳の中のさまざまな研究対象

抽象度

抽象化レベル	構成要素	要素数
様々な高次機能		
脊椎動物の脳のアーキテクチャ	**大脳皮質、海馬、大脳基底核、・・・**	**5個程度**
神経回路		1000億のニューロン
ニューロン	興奮性、抑制性、・・・	数種類（数百種類？）
タンパク質等の分子		1万以上
DNA	塩基 A,T,G,C（あるいはアミノ酸）	4（あるいは 20）

知能の再現が目的

病気の治療や薬の開発が重要な目的

全脳アーキテクチャ勉強会 開催実績

参加者は
技術者と研究者が半々

- 第1回 2013年12月 **開催趣旨説明** 約100名参加
  - 講演者: 産総研 一杉裕志、東大 松尾豊、富士通研 山川宏
- 第2回 2014年1月**「大脳皮質とDeep Learning」** 約250名参加
  - 講演者: 産総研 一杉裕志、筑波大学 酒井宏、PFI 得居誠也
- 第3回 4月**「海馬とSLAM」** 約200名参加
  - 講演者: はこだて未来大 佐藤直行、産総研 横塚将志、富士通研 山川宏
- 第4回 6月 **関西編** 約100名参加
  - 企画: 理研 高橋恒一、講演者: 産総研 一杉裕志、東大 松尾豊、富士通研 山川宏、NICT CiNet 西本伸志、理研 泰地真弘
- 第5回7月**「意思決定」** 約200名参加
  - 講演者: 産総研 一杉裕志、奈良女子大 新出尚之、グーグルジャパン 牧野貴樹
- 第6回7月**「統合アーキテクチャ」** 約170名参加
  - 講演者:富士ゼロックス 岡本洋、玉川大 大森隆司、NII 市瀬龍太郎
- 第7回9月**「感情」**約200名参加
  - 講演者:玉川大大森隆司、京大藤田和生、東京慈恵医科大渡部文子、AGI光吉俊二
- 第8回11月**「時系列学習」**約200名参加
  - 講演者:山口大 宮崎真、電通大 山崎匡、早稲田大 尾形哲也
- 第9回2015年2月**「表現学習」**約200名参加
  - 講演者:東大 酒井邦嘉、産総研 林隆介、立命館大 谷口忠大
- ・・・

過去の講演資料の一部はネットで見られます。

今後は **全脳アーキテクチャ・イニシアティブ**が主催

全脳アーキテクチャ　検索

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。**

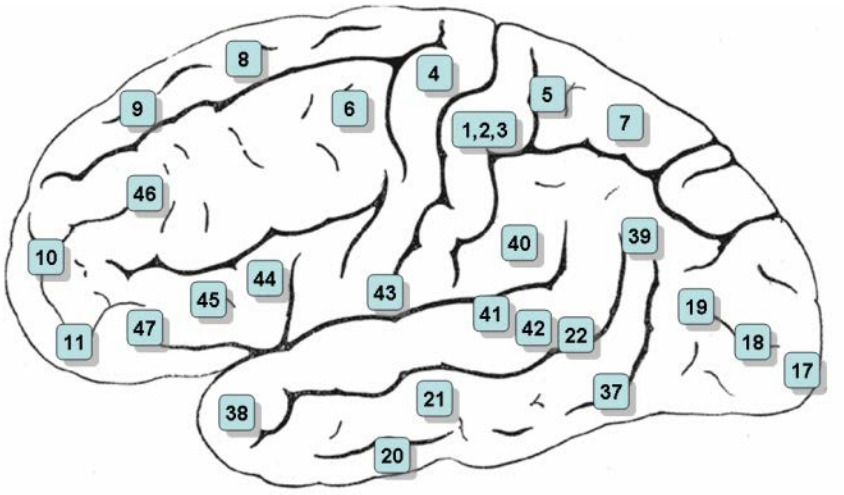
なぜ今、脳の理解が可能に？

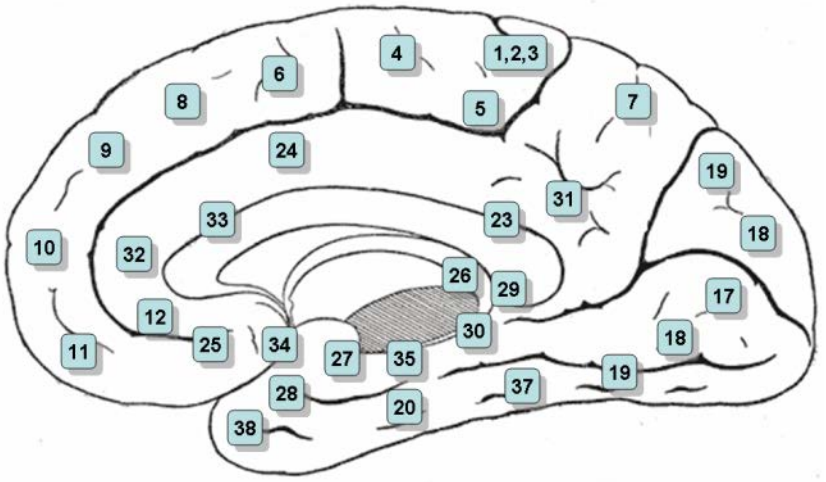
- 脳の理解に必要な知識はこの十数年の間に揃いつつある。
  - 機械学習分野の要素技術の成熟
    - ベイジアンネットの教科書 [Pearl 1988]
    - 強化学習の教科書 [Sutton 1998]
    - 独立成分分析の教科書 [Hyvarinen 2001]
    - **大規模ニューラルネット Deep Learning [Hinton 2006]**
  - 「脳の10年」(1990～1999)以降の神経科学の急速な進歩
    - ドーパミンニューロンTD誤差の論文 [Schultz 1997]
    - V1のスパース符号化の論文 [Olshausen 1996]
    - **大脳皮質のベイジアンネットモデル [Lee and Mumford 2003] etc.**

大脳皮質に関する神経科学的知見

大脳皮質

- 脳の中でも知能をつかさどる重要な部分。
  - 視覚野、言語野、運動野、前頭前野、・・・

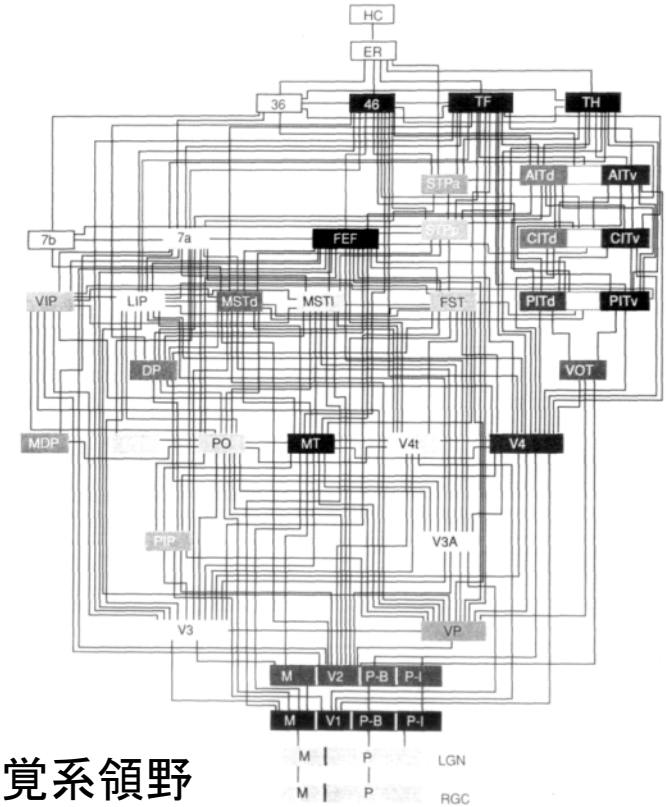




図は http://en.wikipedia.org/wiki/Brodmann_area より

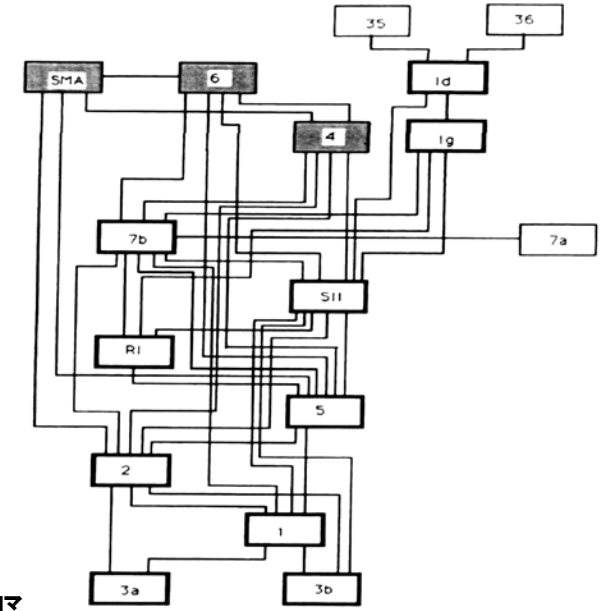
大脳皮質の領野

- 各領野の機能、接続構造はかなり明らかになりつつある。



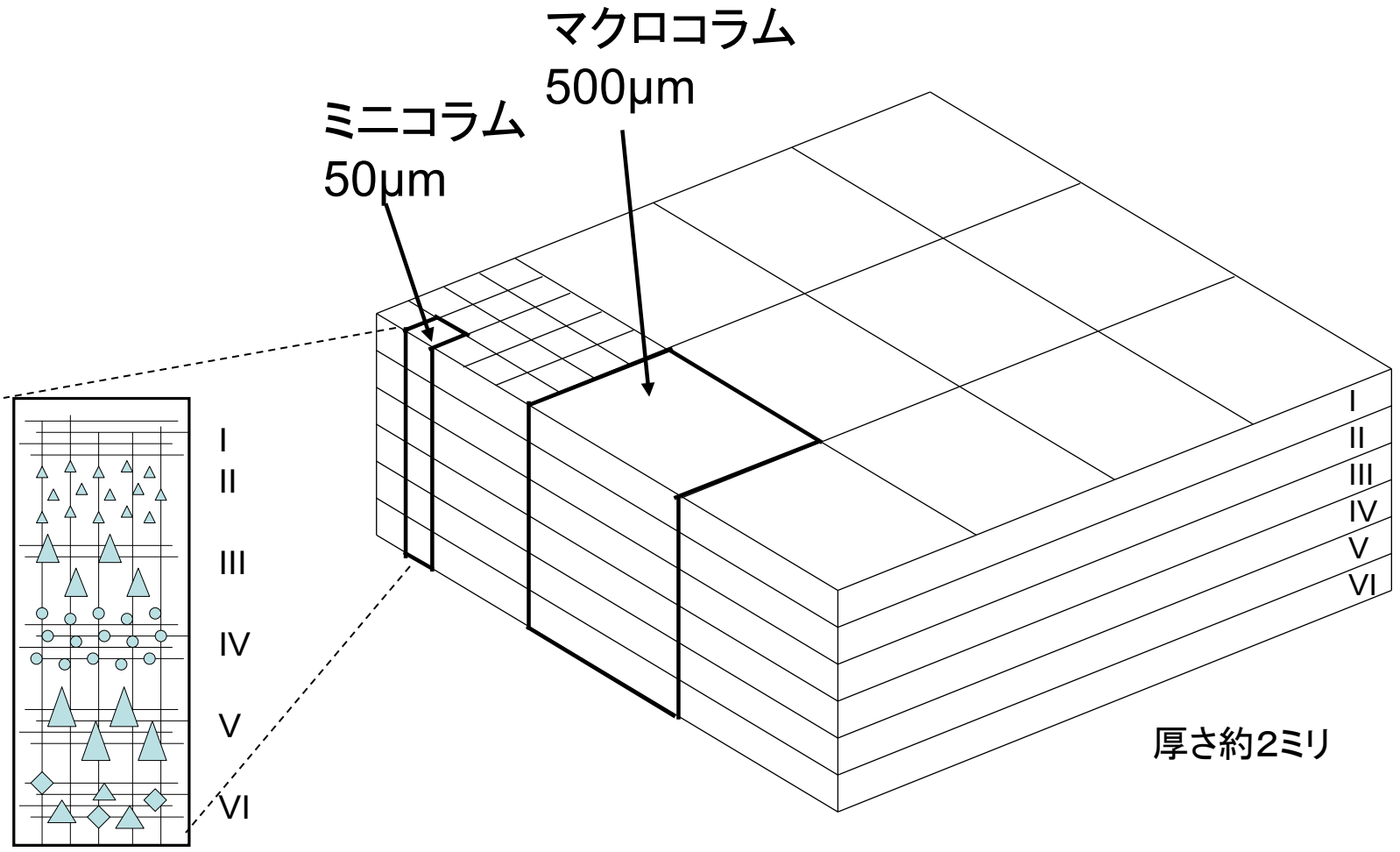
視覚系領野

Daniel J. Felleman and David C. Van Essen
Distributed Hierarchical Processing in the Primate Cerebral Cortex
Cerebral Cortex 1991 1: 1-47



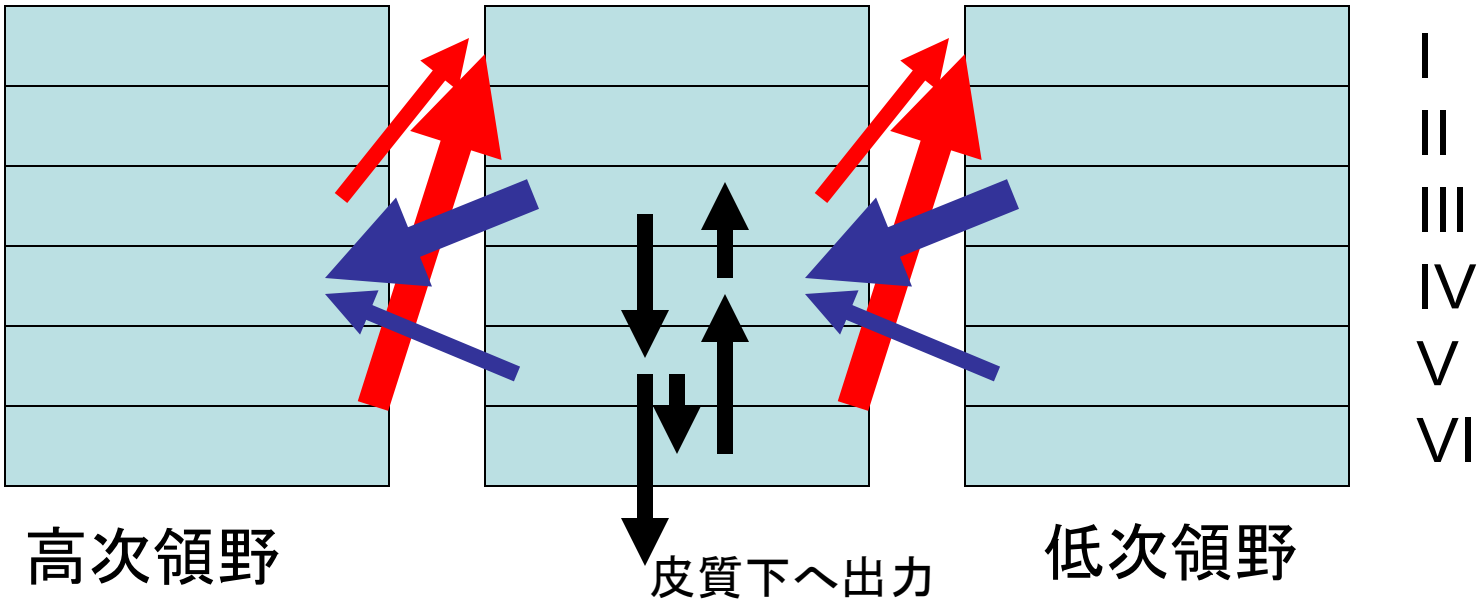
運動野

大脳皮質のコラム構造の模式図



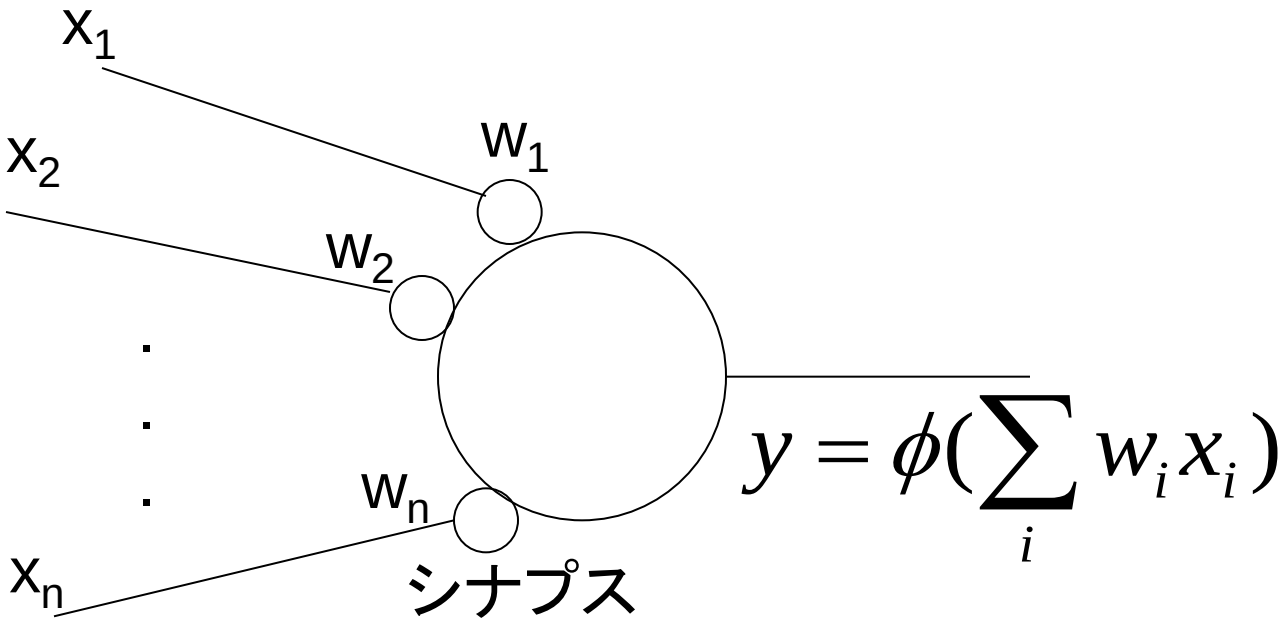
大脳皮質の解剖学的特徴[Pandya and Yeterian 1985][Gilbert 1983]

- 情報処理の途中結果の3層の情報が上位領野に送られ、最終結果の5層の情報は下位領野に戻される。
- 非常に意味ありげな不思議な構造。



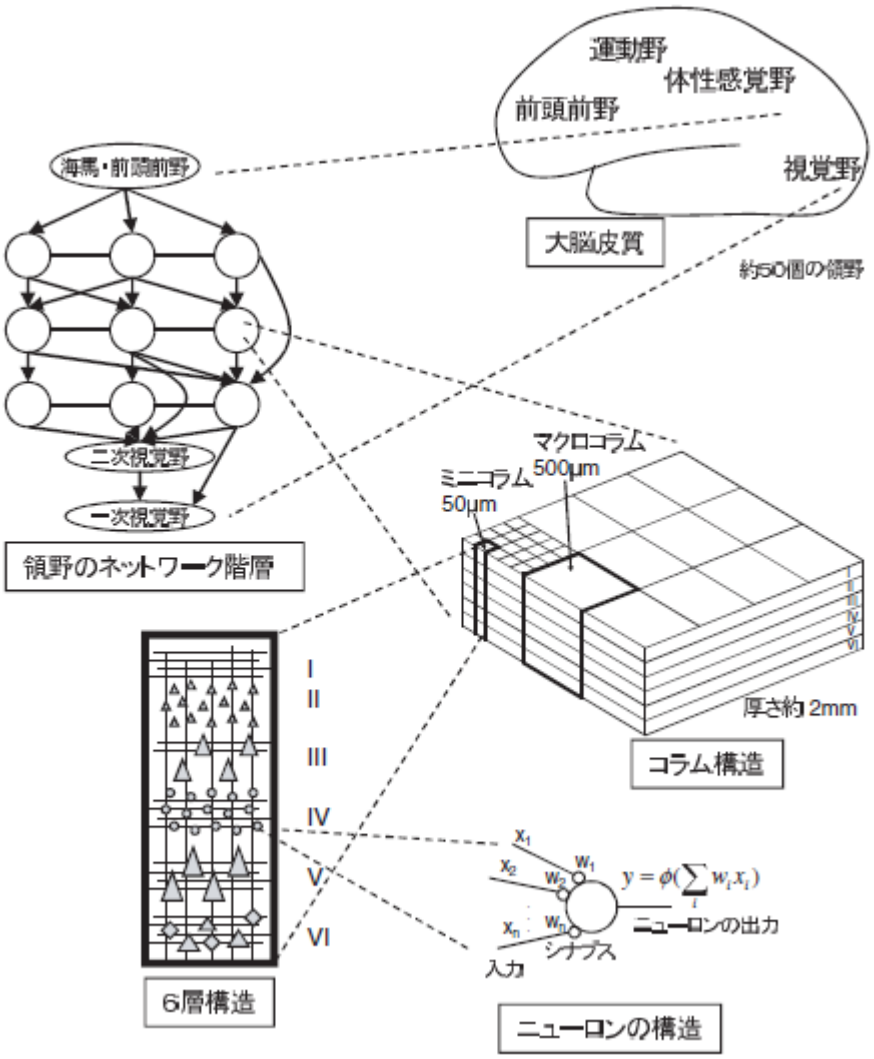
ニューロン（神経細胞）

- ヒトの脳で1000億個。
- 内積計算のような簡単な演算しか行えない。
- シナプスが w という重みを学習。



ニューロンへの入力　　ニューロンからの出力

大脳皮質の各スケールでの構造



領野 約50個

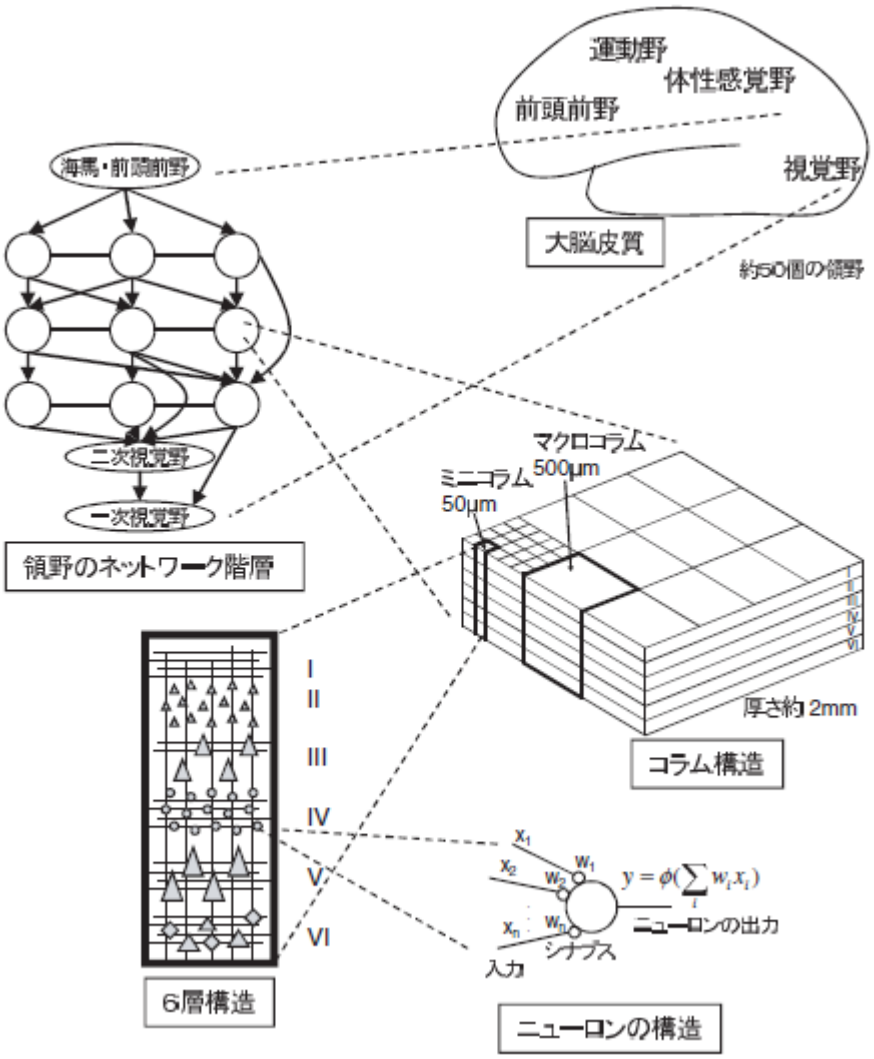
マクロコラム約100万個

ミニコラム 約1億個

ニューロン 約100億個

シナプス 約100兆個

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。

確率伝播アルゴリズム[Pearl 1988]

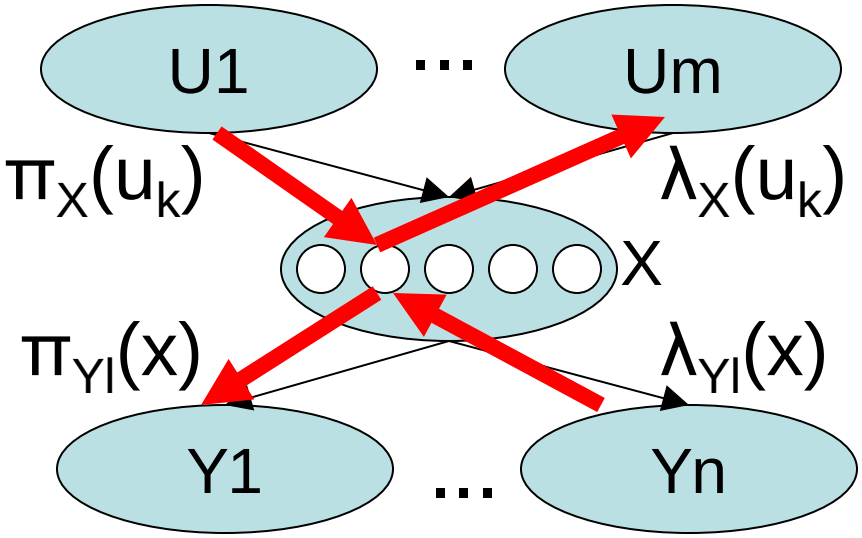
$$BEL(x) = \alpha\lambda(x)\pi(x)$$

$$\pi(x) = \sum_{u_1,\cdots,u_m} P(x \mid u_1,\cdots,u_m)\prod_k \pi_X(u_k)$$

$$\lambda(x) = \prod_l \lambda_{Y_l}(x)$$

$$\pi_{Y_l}(x) = \pi(x)\prod_{j\neq l} \lambda_{Y_j}(x)$$

$$\lambda_X(u_k) = \sum_x \lambda(x) \sum_{u_1,\cdots,u_m/u_k} P(x \mid u_1,\cdots,u_m)\prod_{i\neq k} \pi_X(u_i)$$



近似確率伝播アルゴリズム[Ichisugi 2007]

$$\boldsymbol{l}_{XY}^{t+1} = \boldsymbol{z}_Y^t + \boldsymbol{W}_{XY}\boldsymbol{o}_Y^t$$

$$\boldsymbol{o}_X^{t+1} = \bigotimes_{Y \in children(X)} \boldsymbol{l}_{XY}^{t+1}$$

$$\boldsymbol{k}_{UX}^{t+1} = \boldsymbol{W}_{UX}^T \boldsymbol{b}_U^t$$

$$\boldsymbol{p}_X^{t+1} = \sum_{U \in parents(X)} \boldsymbol{k}_{UX}^{t+1}$$

$$\boldsymbol{r}_X^{t+1} = \boldsymbol{o}_X^{t+1} \otimes \boldsymbol{p}_X^{t+1}$$

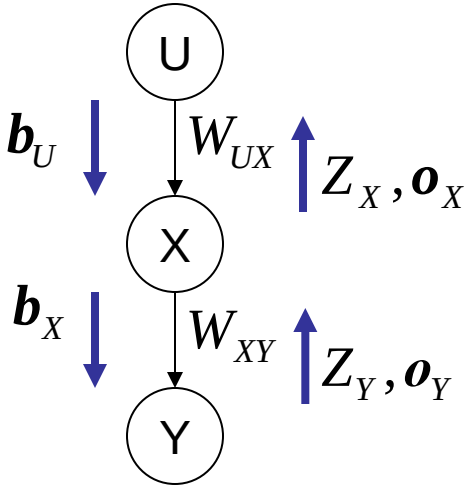
$$Z_X^{t+1} = \sum_i (\boldsymbol{r}_X^{t+1})_i \ (= \left\| \boldsymbol{r}_X^{t+1} \right\|_1 = \boldsymbol{o}_X^{t+1} \bullet \boldsymbol{p}_X^{t+1})$$

$$\boldsymbol{z}_X^{t+1} = (Z_X^{t+1}, Z_X^{t+1}, \cdots, Z_X^{t+1})^T$$

$$\boldsymbol{b}_X^{t+1} = (1/Z_X^{t+1})\boldsymbol{r}_X^{t+1}$$

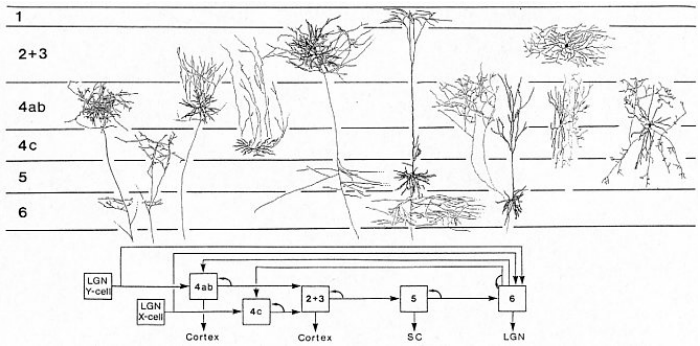
ただし、$\boldsymbol{x} \otimes \boldsymbol{y} = (x_1 y_1, x_2 y_2, \cdots, x_n y_n)^T$

Pearl のアルゴリズム[Pearl 1988]をいくつかの仮定のもとで近似。
（アルゴリズムは今後少し修正する予定）

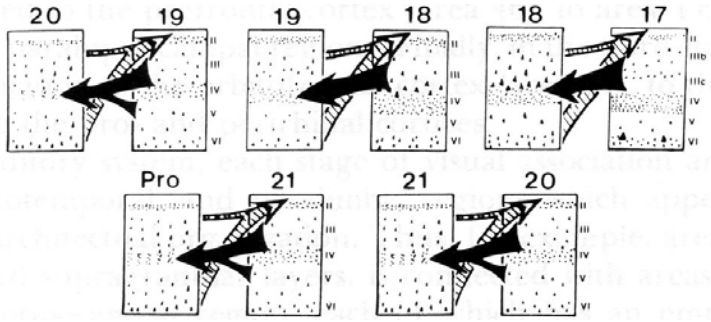


- 神経回路で実現可能
- 大規模化可能な計算量・記憶量

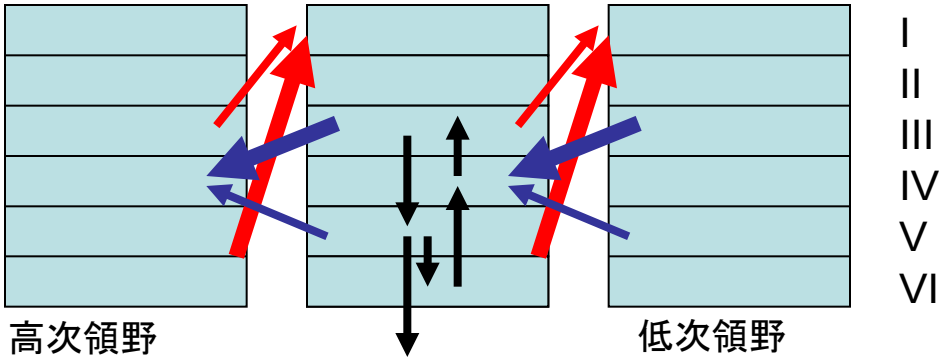
コラム構造・6層構造との一致



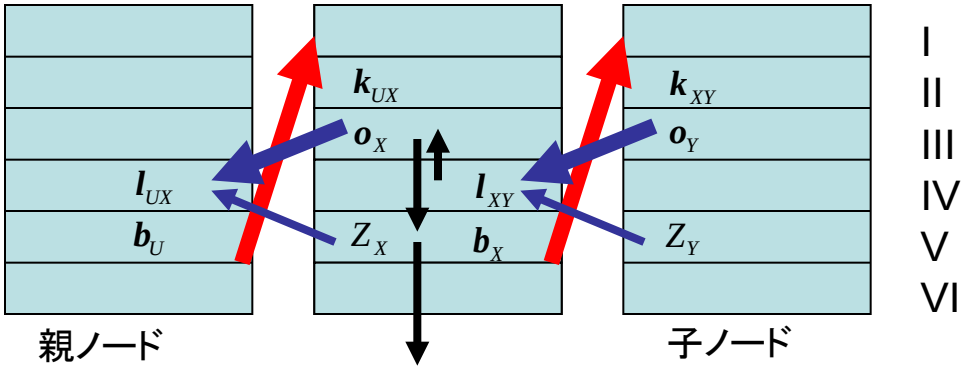
[Gilbert 1983]



[Pandya and Yeterian 1985]



解剖学的構造

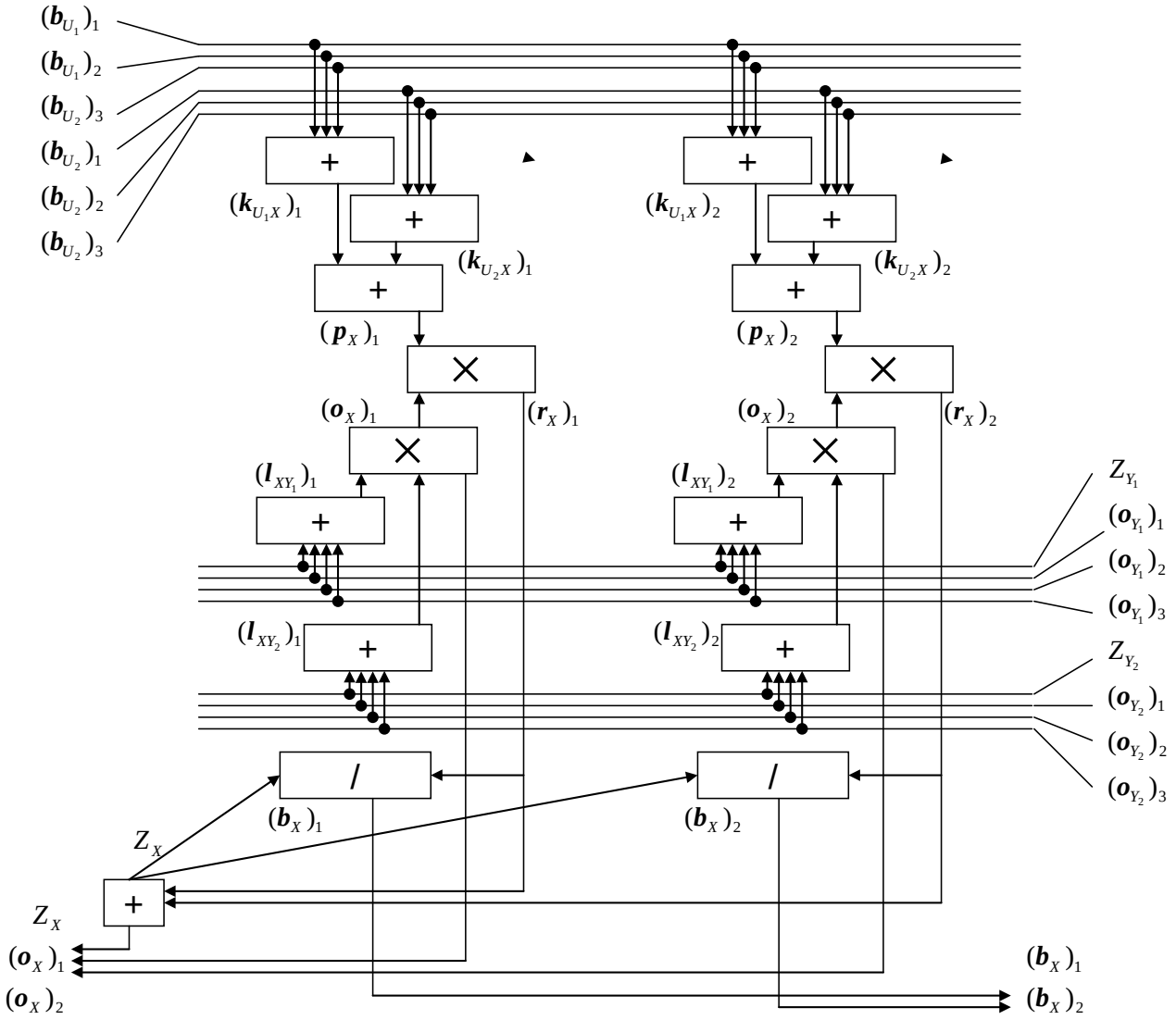


モデル

Pandya, D.N. and Yeterian, E.H., Architecture and connections of cortical association areas. In: Peters A, Jones EG, eds. Cerebral Cortex (Vol. 4): Association and Auditory Cortices. New York: Plenum Press, 3-61, 1985.

Gilbert, C.D., Microcircuitry of the visual-cortex, Annual review of neuroscience, 6: 217-247, 1983.

各変数の値を計算する回路



I

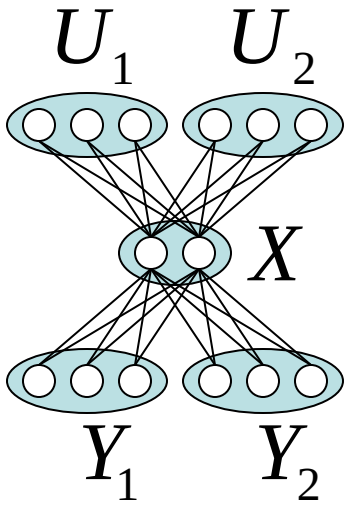
II

III

IV

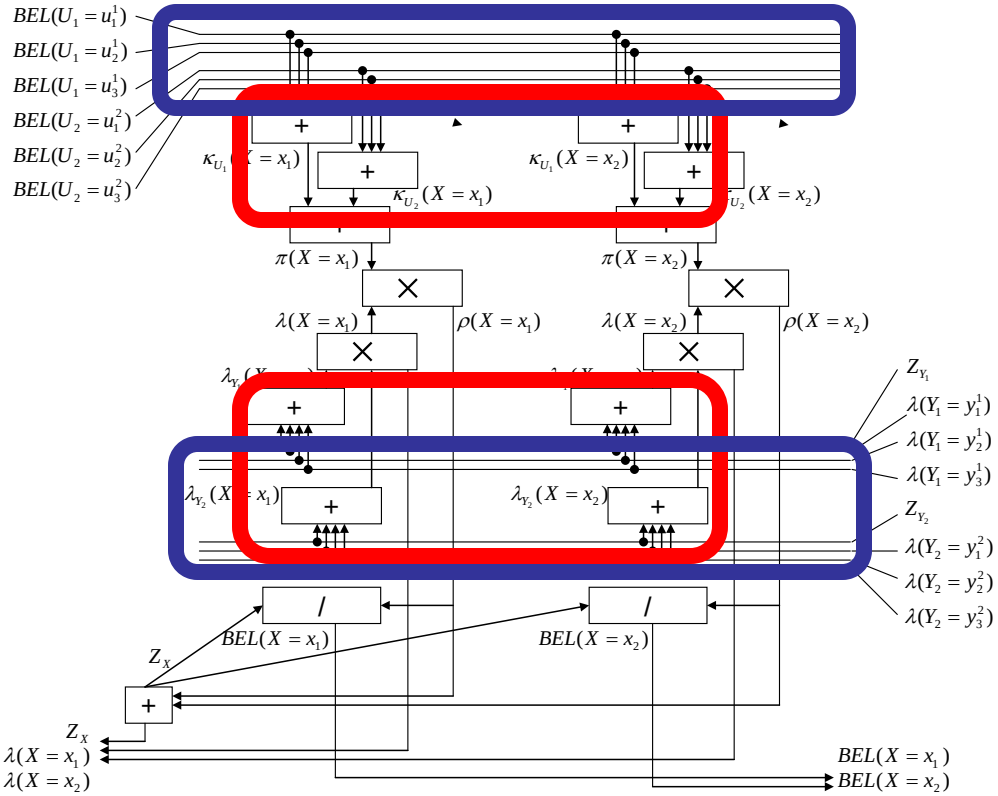
V

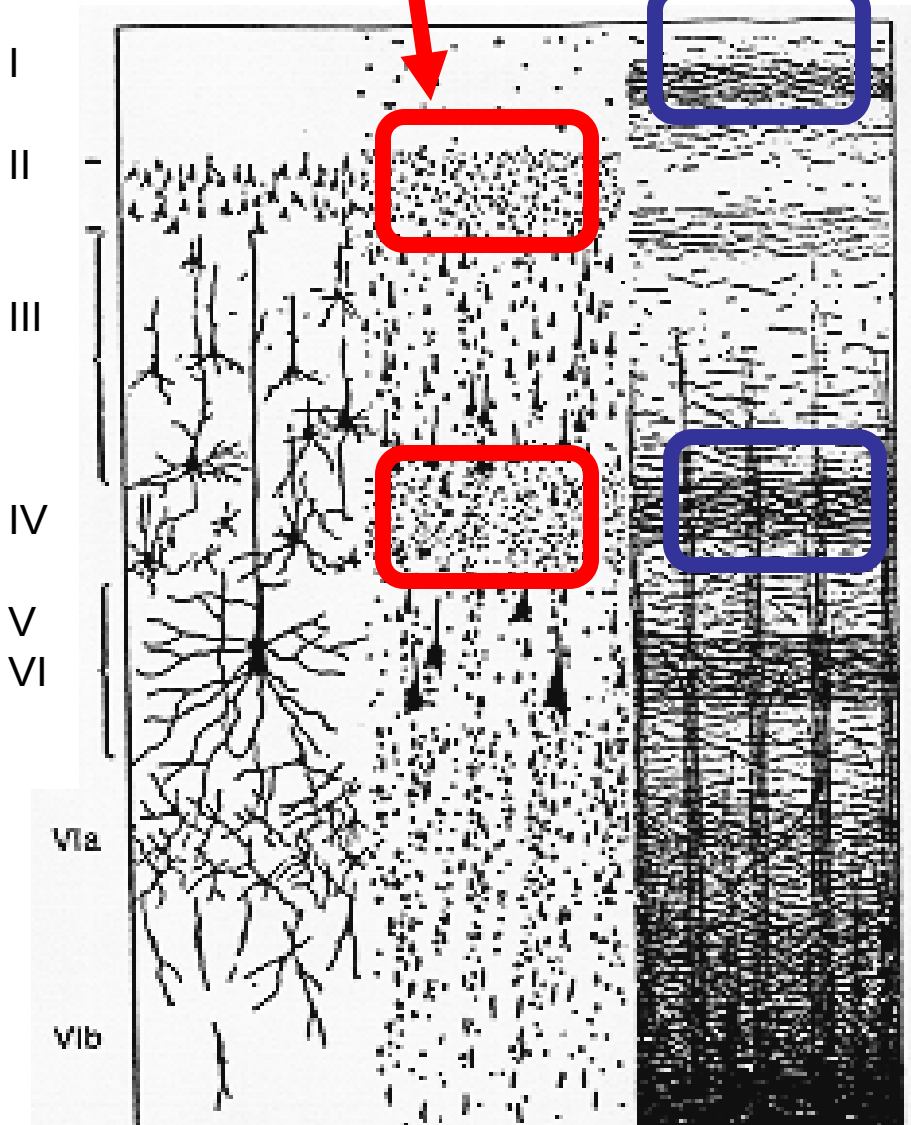
VI



左図は、上記BESOMネットワークのノードXのユニットの値を計算する回路

大脳皮質の構造との一致

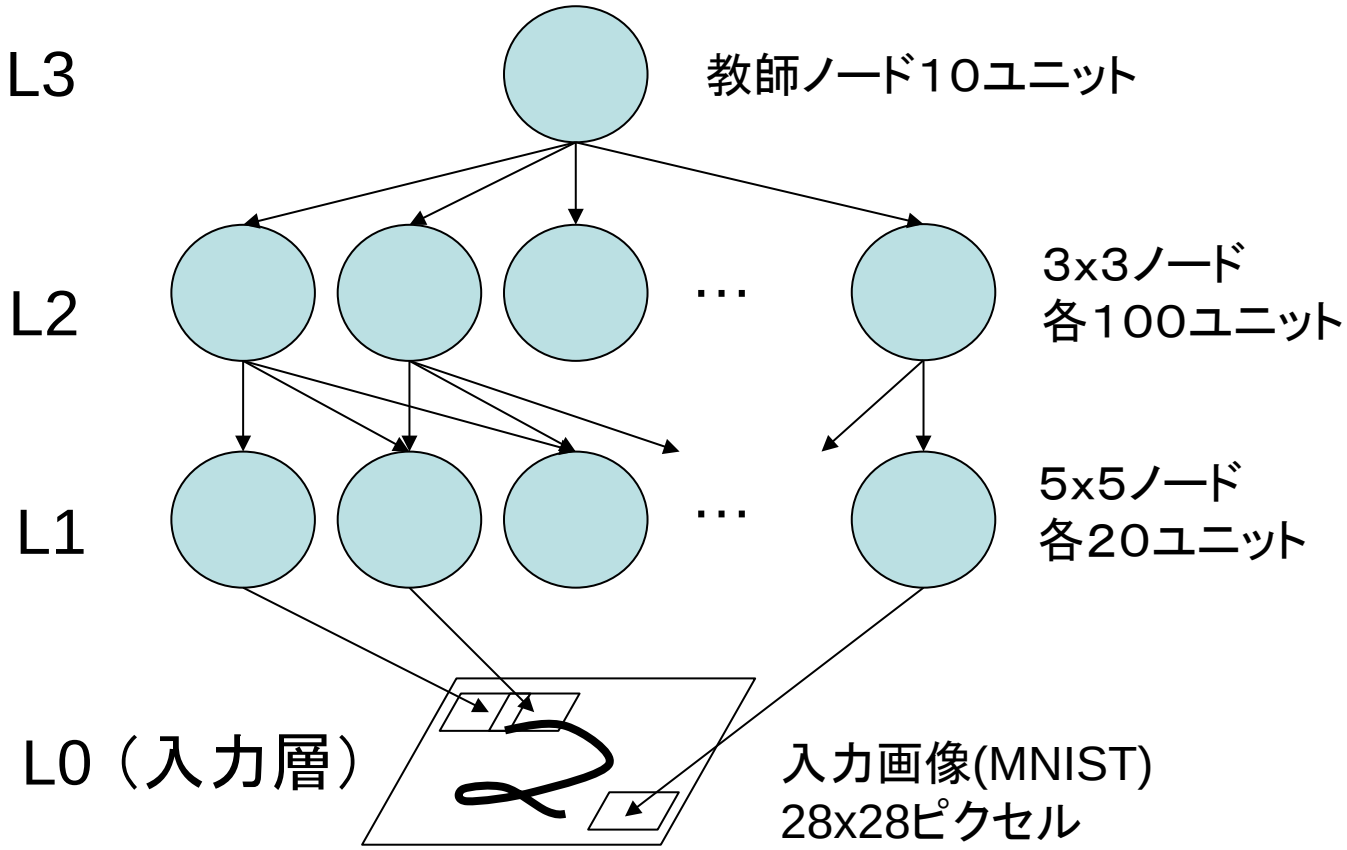




K. Brodmann, Vergleichende Lokalisation der Grosshirnrinde. in: ihren Prinzipien dargestellt auf Grund des Zellenbaues,. J.A. Barth, Leipzig, 1909.

This figure is taken from the following Web page.
http://web.sc.itc.keio.ac.jp/anatomy/brodal/chapter12.html

4層「制限付き」ベイジアンネット
手書き数字認識[一杉 未発表]



$$P(X \mid U_1, \cdots, U_m) = \frac{1}{m} \sum_{i=1}^{m} P(X \mid U_i)$$

Neocognitron,
Deep Learning
と同じ深い構造の
ベイジアンネット

ノード数にたいして
1入力の処理が
ほぼ線形時間で動作

注：EMを使う場合、現在 O(n^2)

pre-training なし、認識アルゴリズムは
OOBP、EMアルゴリズムで学習
認識率：92%程度

参考：
linear classifier (1-layer NN) 88%
Ciresan et al. CVPR 2012 99.87%
「MNIST handwritten digit database」
http://yann.lecun.com/exdb/mnist/

ここまでのまとめ

- 脳を構成する主な器官の計算論的モデルは不完全ながら出そろっている。
- 脳全体の機能の再現も目指せる時期。

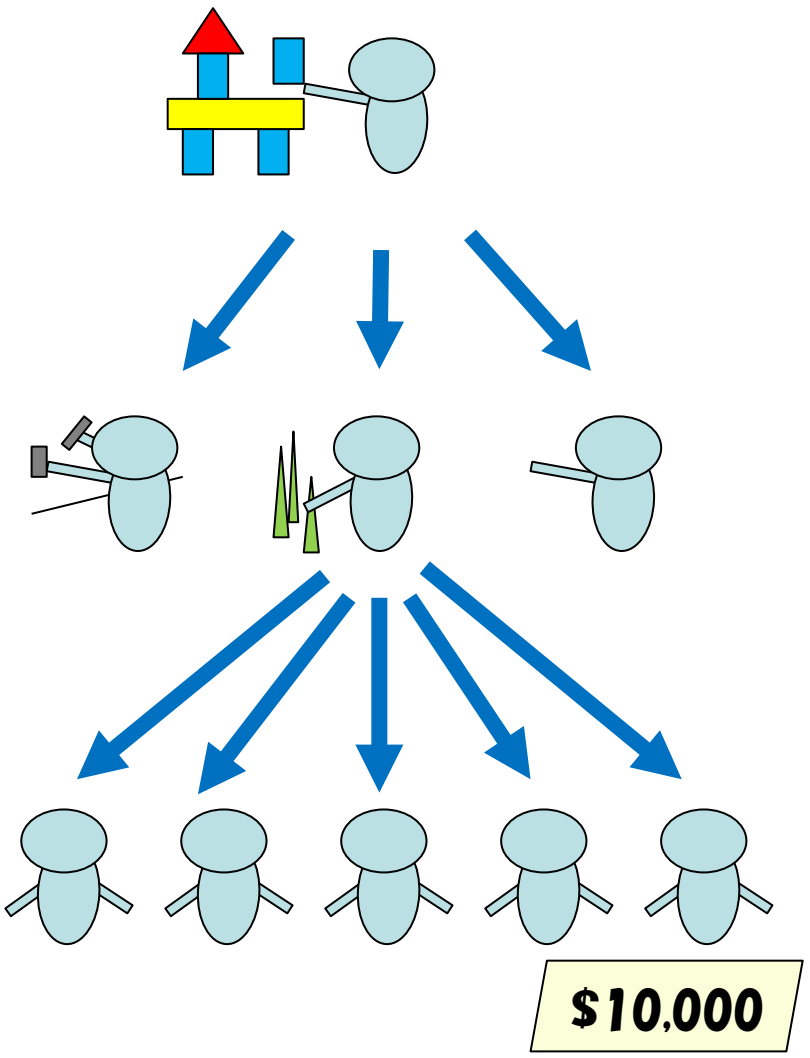
- 大脳皮質がある種のベイジアンネットであることはまず間違いない。
- 脳を模倣した人工知能を実現するために**非常に重要な手掛かり**。

脳とベイジアンネット　検索

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

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。
  - **機械化で得られる利益 ＞ 解雇した労働者への所得補償**

    となったとき、すべての人間が働くのをやめても
    社会全体の利益が増えることになる。

ヒト型AIの人間社会への影響

注：私は社会科学や進化論の専門家ではありません。
専門家による真剣な議論が望まれます。

経済への影響

早くても20～30年後？

- 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。
  - すべての人に主治医と家庭教師と専属弁護士。

幸福への影響

- 労働から解放されて幸せ！

- 仕事がなくなり不幸！
  - AI出現以前の現在の常識にとらわれている？「働かざる者食うべからず」は過去のものに。
  - いや、おそらく人間は働いていると幸せを感じるように作られている。労働本能。
  - 疑似的に労働本能を満たす方法：家庭菜園、園芸、釣り、ペット、料理、編み物・・・

人口への影響

- 普通に考えれば、労働の自動化で生産性が向上すれば人口は増える。
- しかし・・・。
- ヒト型AI出現以前:
  - 人間は労働力。人口が多いほど国力が強い。
- ヒト型AI出現以降:
  - 機械が労働力。人口が少ないほど国力が強い。

社会制度への影響

- 民主政治と専制政治で結果が大きく異なる。
- 民主政治：
  - 社会全体が豊かになる。富の再配分。ベーシックインカム。
  - 古代ギリシャ、古代ローマ
- 専制政治：
  - 役人も軍隊も労働者も不要に。
    専制君主にとって文字通り「人間がいらない」世界。

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。

機械安全

- **本質的安全設計：**
  - **万が一の危険を減らしておく。**
  - 必要以上に強くしない、かたくしない、速くしない、とがらせない、人間からの隔離、・・・
- **安全防護策：**
  - **どうしても減らせない危険への対策。**
  - 防護柵、カバー、非常停止、・・・。

参考：「機械類の設計の一般原則 」明治大学教授 向殿政男
http://www.jsa.or.jp/stdz/edu/pdf/b3/3_02.pdf

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？

ヒト型AIの権利

- 生物はすべて自然権を持つ：
  - ただし、ここでは
    「自然権」＝「生き残り子孫を残そうとする性向」
    と定義。
- ヒト型AIは自然権を持つか？
  - 設計次第。「自分は壊れても人間を助けようとする性向」を持つように設計すれば、生物のような自然権は持たず、権利の衝突は避けられる。
- ところが、1つ問題が・・・。

AIへの同情心は人類を滅ぼす

- 人間は、自分と似たものに対し共感し、自分と同じ権利を認めたくなる性質を持っている。
- しかし、
  - 生存権を認めると、寿命ないので個体数は単調増加。
  - 「自己改造権」を認めると、無限に能力増強。
- 対策：人間が同情心を持ちにくいように作っておくしかない。
  - 外見や話し方を人間と全く違ったものにする。
  - 人間から嫌われる感情を持つよう設計する。

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