資料３－２

AI開発ガイドラインについて
透明性と制御可能性を中心とした緒論

高橋恒一

理化学研究所
慶應義塾大学
全脳アーキテクチャ・イニシアティブ

Whole Brain Architecture Initiative

「AI開発ガイドライン」(仮称)に掲げる原則の項目及び内容(たたき台)

① **透明性の原則**

ＡＩネットワークシステムの動作の検証可能性及び説明可能性を確保すること。

② **利用者支援の原則**

ＡＩネットワークシステムが利用者を支援し、利用者に選択の機会を適切に提供するよう配慮すること。

③ **制御可能性の原則**

人間によるＡＩネットワークシステムの制御可能性を確保すること。

④ **セキュリティ確保の原則**

ＡＩネットワークシステムの頑健性及び信頼性を確保すること。

⑤ **安全保護の原則**

ＡＩネットワークシステムが利用者及び第三者の生命・身体の安全に危害を及ぼさないよう配慮すること。

⑥ **プライバシー保護の原則**

ＡＩネットワークシステムが利用者及び第三者のプライバシーを侵害しないように配慮すること。

⑦ **倫理の原則**

ＡＩネットワークシステムの研究開発において、人間の尊厳と個人の自律を尊重すること。

⑧ **アカウンタビリティの原則**

ＡＩネットワークシステムの研究開発者が利用者など関係するステークホルダーに対しアカウンタビリティを果たすこと。

AIシステムの設計、実装、配置展開の各段階における透明性

ソフトウエア工学に基づくフロー

1. 要求分析	規格化、オープン設計、アセスメント
2. 設計	
3. 詳細設計	オープンソース、文書化
4. 実装	
5. 配置前学習	学習データセットの公開、学習結果出力
6. 配置・展開	動作ログの作成、保全
7. 配置後学習	学習結果の出力
8. 事後検証	動作ログ、学習結果の検証可能性確保

従来のソフトウエア工学を超えた問題

- フローの各段階が互いに侵食する （複合学習問題： Sculley *et al,* NIPS14）
- 詳細設計、実装以後も学習により動作が変化する

二つのオープン化：開放性 vs. 透明性

- **開放性**
  - ⇨ API、プロトコル、記述形式公開によるエコシステム構築
  - ⇨ 情報の流通性、相互運用性、プライバシー
  - 開発コストの分散、プラットフォームへのデータの誘い込みなどで経済的な合理性があるため、行政の役割は主導よりもむしろ舵取り、監視、独占防止等。
  - オープンソース

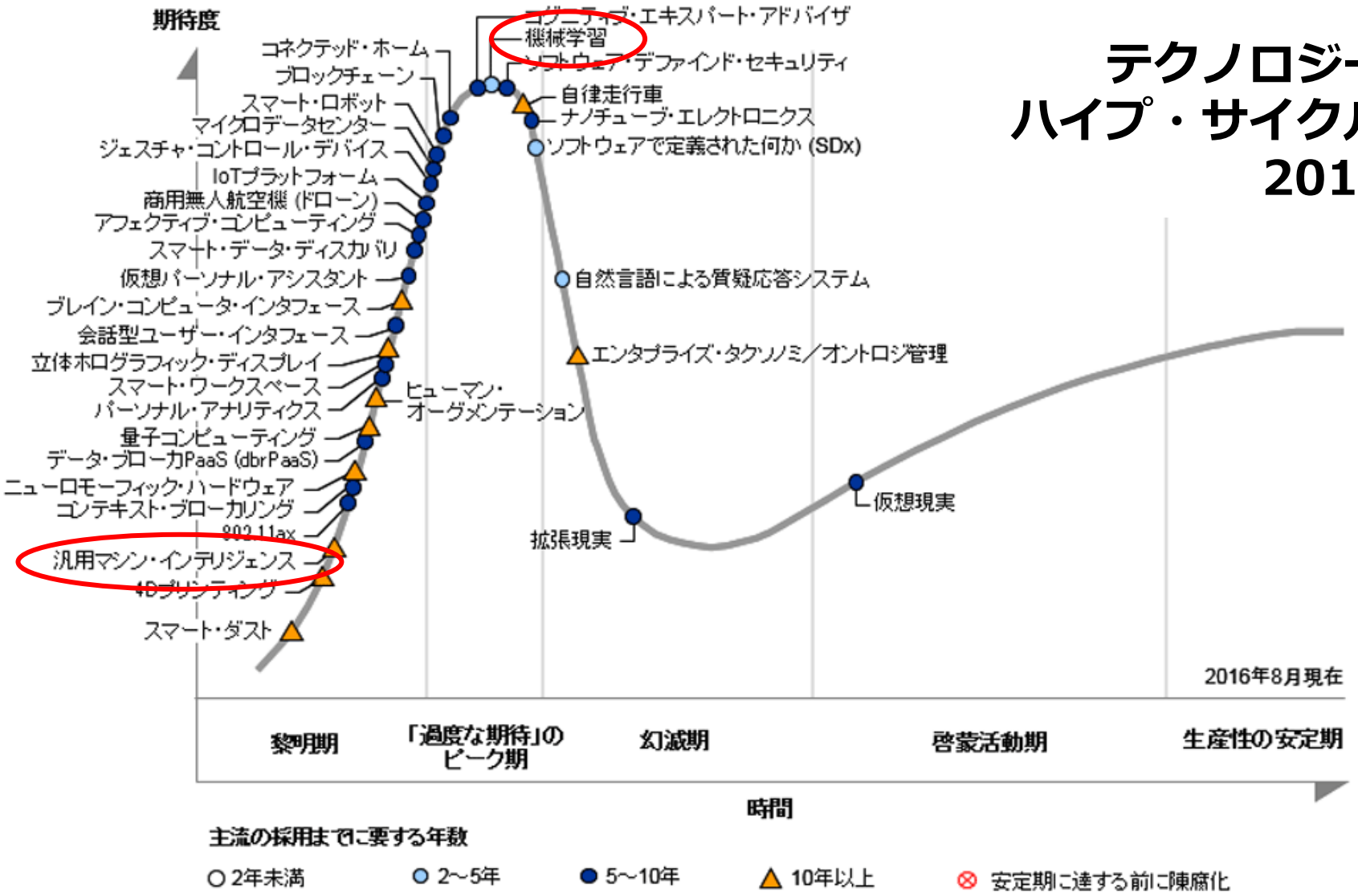
- **透明性**
  - ⇨ 実装、価値関数、学習内容のブラックボックス化の回避
  - ⇨ 解釈可能性 （自己説明能力の付与）
  - ⇨ ロギング、チェックポインティング等
  - 必ずしも経済的合理性は自明ではない。
    - ログ保持、自己説明的アーキテクチャの開発などはコストセクター。（訴訟、PLリスク、ブランドリスクなどは存在。）
    - 共通化により産業振興にメリット。
  - 各国の組織（OpenAI, Partnership on AI、FLI等）や行政との連携でガイドライン作成、規制の共通化。

想定すべき技術体系

- 第三次AIブーム
  - 機械学習中心
  - 認識、意思決定、行動のそれぞれが分断
    - Deep learning、IBM Watson、IoT

- 第三次ブーム以降
  - 認知＝行動サイクルの再統合でより自律性が高まる
  - 自動化装置の性能＝自律性の高さ＝経済的合理性
  - 技術的な鍵としては？
    - 認知アーキテクチャ、汎用性（汎化能力、転移学習）
  - ある意味主流への回帰
    - サブサンプションアーキテクチャ （掃除ロボット）
    - 自動運転 （乗り物、ロボット、建物等）
    - 自動運営（経営、店舗、工場、金融、ICTシステム等）
- ただし、当然のことながら技術トレンドの予測は難しい。

テクノロジー ハイプ・サイクル 2016



出典：　ガートナー社　プレスリリース　2016年8月25日

認知アーキテクチャの形式

（いずれも環境との相互作用により動作する）

1　現状の主流　（非認知アーキテクチャ）

センシング　　推論　　行動
認知　　アクチュエーション

2　古典的認知アーキテクチャ

センシング　→　認知　→　意思決定　→　行動　→　アクチュエーション

3　包摂アーキテクチャ

計画
行動
センシング　→　反射　→　アクチュエーション

参考：　脳型認知アーキテクチャ（一例）

思考（新皮質）
情動（辺縁系）
センシング　→　生存（脳幹）　→　アクチュエーション

透明性、制御可能性を担保するにあたって留意すべき対象は推論・学習形式･アーキテクチャにより異なる

共通

- ソースコード　→　アーキテクチャによっては相対的に重要度低下
- 行動ログ、通信ログ
- 判断ログ、根拠　→　アーキテクチャによっては困難

機械学習：　教師あり・なし学習

- →　獲得表象
- 訓練データセットに依存
- 根拠の説明：例えばNNによる分類モデルではアテンション技術などが対応するが、複雑なケースにおいてクリアカットな解釈は困難。

記号推論：　（エキスパートシステム）

- →　推論ルールセット

強化学習　（DQN、AlphaGo）、**脳型アーキテクチャ**：

- →　報酬系（目的関数）
- 「行動のプログラム」から**「行動原理」のプログラム**（強化学習）**「行動原理の学び方」のプログラム**（脳型アーキテクチャ）へ
- 「報酬ハック」の危険性は各国組織が指摘(Partnership on AI, OpenAI, FLI)

技術の高度化に従って、実際には複合的アーキテクチャとなりつつある。

まとめ

- AI開発ガイドラインにおいて、透明性、制御可能性、セキュリティー、倫理、アカウンタビリティーなどは相互に結びついた複合性がある。
- 情報システム一般に、開発プロセスの各段階におけるアセスメントは可能だが困難。AIシステムにおいてはさらに各段階の相互侵食が進む。
  - AIシステムではシステム構築後、配置前、配置後において学習により動作が変更。
- 今後の中長期的技術トレンドとして、認知＝意思決定＝行動を統合した認知アーキテクチャの台頭により自律性が高まる可能性。
- 自律的アーキテクチャにおいては、行動の直接的なプログラムから、行動原理のプログラム、あるいは行動原理の学習方法のプログラムへと比重が移ると考えられ、報酬ハックなどの新たなセキュリティーリスクが発生する可能性。

参考：脳型認知アーキテクチャにおける価値システム

- AIの効用、経済的価値は自律性に比例する。
- 神経科学では１次欲求中枢と身体性から階層的に情動、欲求、価値観、行動原理などが構築されているという見方がある。
- AIにどう価値システムを実装するかは**技術的課題**であるとともに、artificial moral agent構築の**倫理学的、社会学的課題**であり、対攻撃性規制など**政策、行政的課題**でもある。
  （ヒトと「似ているが非なる」価値システムを持ったエージェントを社会がどう受容するか。）
- ⇛　人文社会学との共創によるAGI研究が必要
  - 慶應SFCAI社会共創ラボ、AIRなど