2019/3/5

人工知能技術開発の現時点における倫理的問題の諸相

中川裕志
(理化学研究所　革新知能統合研究センター)

人工知能の歴史：
拡張　対　置き換え

- 人工知能は人間の能力を拡張する
  - IA: Intelligent Assistance / Amplifier

- 人工知能は人間に置き換わる
  - AI: Artificial Intelligence
    - IA vs AI はマルコフ本の基本的視点（第4章にAIの歴史も含めて素晴らしい記述がある）
    - J.マルコフ: 人工知能は敵か味方か, 日経BP社,2015

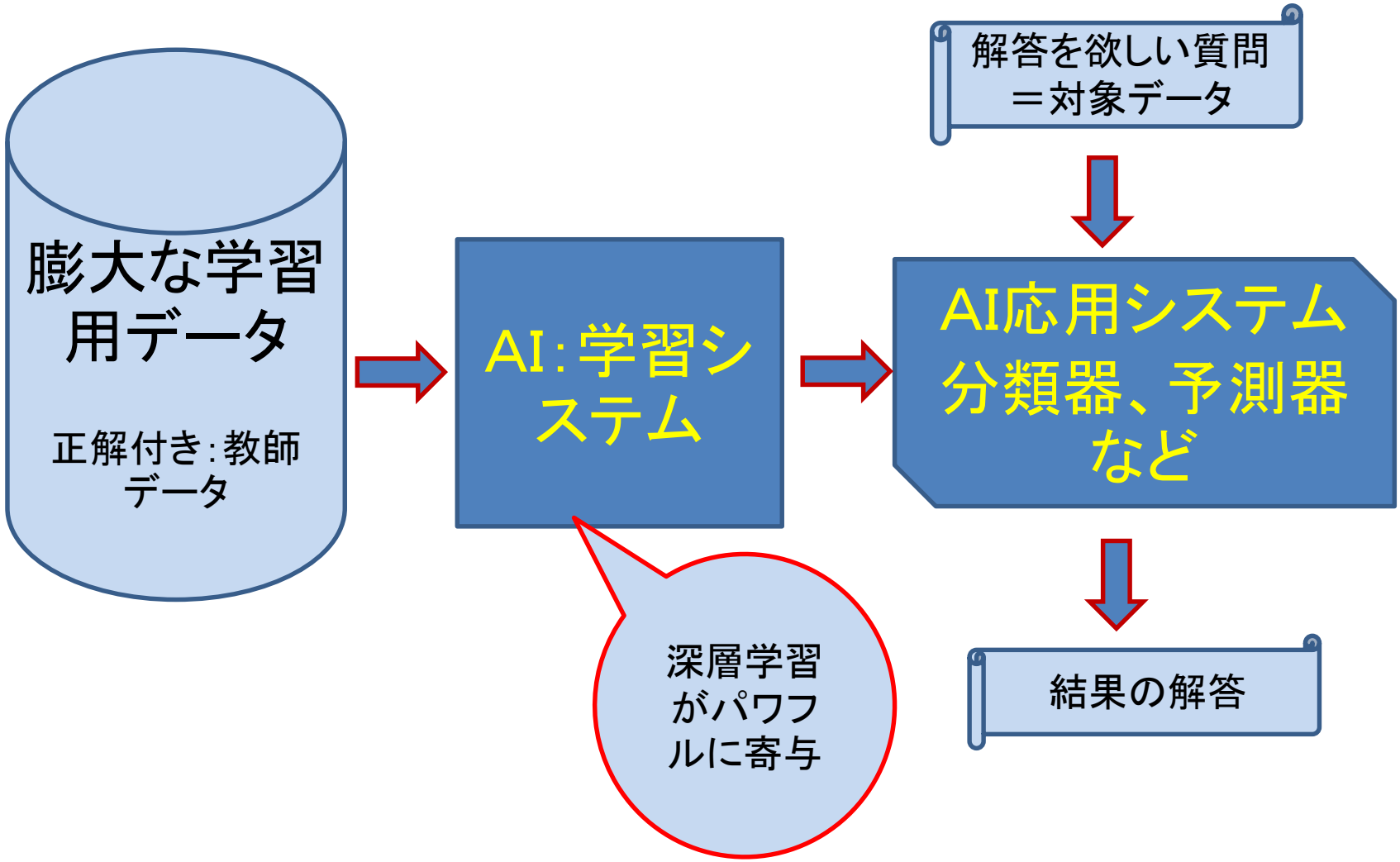
人間はループの内か外か

- IA vs AI と並行して議論されてきたのは、システムやタスク実行において人間がループの内側にいるか外側にいるかという設計上の起点

- 内側　だと　IA的
  - システムは人間の拡張。人間はシステムと共同作業
  - IAが複雑高機能化すると人間も変わらざるを得ない

- 外側　だと　AI的
  - システムは自律的に動き、人間はシステムに命令するだけ

現代AIの仕組み：学習と応用システムの関係



技術の現状は既にブラックボックス化が大きく進行中

- 人間の仕事を自律的AIで置き換えるにせよ、人間の能力を拡張するにせよ
  - 開発者に責任はあるはず
  - だが、既に関係者たちが把握しきれない状態に突入しているのかもしれません
  - 関係者： Multi-stakeholder
    - 人工知能開発者
    - 人工知能へ学習に使う素材データを提供した者
    - 人工知能製品を宣伝、販売した者
    - 人工知能製品を利用する消費者

➢したがって、事故時の責任の所在を法制度として明確化しておく必要がある時期になってきています

透明性とアカウンタビリティ

- ブラックボックス化への対応策
  - IEEE　EADversion2 法律編によれば
- 事故時の責任の所在を法制度として明確化しておく必要がある
- 具体的には
  - 透明性　と　アカウンタビリティ（説明責任）
  - より根本的にはAIに委譲できない権利：戦争、死刑など人の命にかかわるものを確定しておくことが大切

透明性

- AIにおける学習で使われた入力教師データ　とAI応用システムに投入された入力データを開示できる
- AI応用システムの出力結果
- AI応用システムにおける大雑把なデータの流れ
- AI応用システムを開発主体、出資者、

Accountability:説明責任　誤訳？

- accountabilityを日本では説明責任と訳しているが、どうもこれが誤解の元らしい。（大屋先生）
- 「説明をする責任」ではなくて
- 「責任をだれがとるかを説明する」と言ったほうがむしろ正しいようだ。

- ちなみにIEEE EAD ver2：General Principle にはそのように説明されている。
- そして、transparency は accountabilityを支援する手段
- → AIのどの部分をtransparencyの対象にするかも自ずと見えてくる。

説明責任

- 正しい考え方
  - 入力データから結果出力の妥当性、公平性、正統性をAI応用システムの利用者が納得できる形で説明すること
    - 単なる透明性に基づく開示では不十分
    - 一般利用者には納得できない専門的な説明×
      - もちろんAIの動作を理解できれば理想的だが、一般利用者、多くの場合は専門家にも、なかなか困難
    - ➢ 責任者を明確に指摘できること

➢ **しかし、この説明責任に対応できる理解可能な説明を作り出すことは至難の業**

➢ そこで登場するのがトラスト

説明からトラストへ：利得と損害

- 多数に上る過去の類似例の入力データと結果出力を示して、当該結果の妥当性、公平性、正統性を納得してもらう方法

- AI応用システムの利用履歴を開示して信用を得る

注意

- 技術的にみたトラスト
  ＝　故障せずに安定して動くこと

- AI倫理の文脈でのトラストはむしろ「社会科学的な信用」という概念に近い

トラストはどうやって確保するか？

- 過去の学問的成果の集積を信用してもらう
  - 数学、物理学、医学、…
- 専門家をトラストしてもらうライセンス制度
  - 専門家のスキルのトラスト：医師国家試験、司法試験など
  - ライセンスする側（国など）へのトラスト
- それでも事故は起きる
  - 補償制度の確立：保険など
- これら全部を統合したシステム体系がトラストの基本

トラストの補足

- 利用者がサービス提供側をトラストするという局面ばかり考えてきたが

- サービス提供側が利用者をトラストするという問題もある
  - 利用者認証（多くはネットワーク越し）
  - Self Sovereign Identity
  - Bad userではないことを推定

AI兵器:ドローンの例

- 陸海空の3領域＋宇宙（情報の取得ないし伝達）＋サイバー
- AI:支援手段
  - 敵味方の戦況認識
  - 状況に応じた作戦ないし戦略の選択支援、
  - 攻撃目標の識別

- 目標（米国）:自軍の兵士の損失をできるだけ減らす
  - ベトナム戦争で死傷者大→政治問題化、社会問題化
  - 兵士が戦闘現場に行かずに遠隔操縦で敵の戦闘員を攻撃できる兵器 → 攻撃用ドローンが開発,実戦投入

倫理的問題

◆ 兵士が敵兵を殺傷してよいのは自身も敵と同程度の死の危険に向き合うから

1. 遠隔操縦であるゆえに上記の敵兵殺傷を許す戦争の倫理は不成立

2. 遠隔操縦であるため、敵の戦闘員と民間人の区別が付きにくい。AIが搭載された自律型攻撃ドローンではこの区別をAIがしなければならない。

3. 遠隔操縦ないしは自律型であるため、民間人を装うテロリスト相手ではなおさら区別が困難

4. そこで、常時、顔認識などの個人認識と個人ごとの行動履歴を収集し、行動履歴パタンの集合からなるビッグデータからデータマイニングによって敵の戦闘員であるかどうかを推定する。

5. 3.で敵の戦闘員であると推定された人物に対しては、その人物から攻撃される前に先制攻撃をすることができる。

自律型AI兵器

◆ 自律型AIドローンが誤って非戦闘員を攻撃したときの責任

1. AIパッケージの開発者。たとえば、プログラム言語や機械学習パッケージの開発者

2. 上記のパッケージを組み合わせて、ドローンに搭載するAIシステムを開発した人あるいは組織

3. このようなAI兵器の仕様を決めて発注した国防省のような軍事組織

4. このようなAI兵器の実戦投入を決めた軍首脳部、ないし実施部隊

5. このような軍事組織を作成、運用している政府

6. このような政府を選んだ国民

群をなす自律AI兵器

- 全ドローンが揃って四方八方から攻撃
  - 単体ドローンに比べて群ドローンははるかに大きな破壊力
  - IEEE EAD ver2 では自律AI兵器群の禁止を推奨

- ドローンのような安価な機材にインターネット経由で入手容易なAIシステムを組み合わせる武器を国際的な統制が効かないテロリストが導入したらどうなるか？

- 特定通常兵器使用禁止制限条約（CCW）を通じて、兵器を使う段階は国際的に制約
- 兵器製造規制は各国任せになるという国際政治状況

IEEE EAD　ver2：
自律兵器システムの定義の再構築

- 倫理規定には往々にして重要な抜け穴
  - 例：人道問題になり得る兵器を開発し兵士に与える
- 自律兵器 (AWS, Autonomous Weapon Systems) の概念定義は混乱
- 自律兵器は、秘密裡に帰属不明な状態で運用されがち
- 自律兵器の結果責任をあいまい化されがち

IEEE EAD:
自律兵器システムの定義の再構築

- 自律兵器開発の合法化は中期的に地政学的危険性を招く先例となる
  - 自律兵器同士で発砲し合い意図しない紛争
- 自律兵器に頼って構築した戦略バランスは、ソフトの進化で一夜にして崩れたりする恐れがある
- 人が監督しなければ、余りにも簡単にはずみで人権侵害が起き、緊張が高まる

IEEE EAD:
自律兵器システムの定義の再構築

- 自律兵器の直接的、間接的顧客は多様性に富んでおり、兵器拡散や誤用の問題を複雑な問題にする

- 自律兵器は紛争を急速に拡大する
  - 人より反応が速いため、自律兵器同士で対峙すると人よりも急速に紛争が拡大

- 自律兵器の設計保証を検証する標準の欠如

- 自律兵器と、準自律兵器の倫理上の境界の理解は混乱しがち

人工知能と軍事

- 人工知能の軍事利用は避けることができない
- →なぜなら
  - 人工知能が部分的にせよ軍事利用されているのはほぼ確実
  - 技術として公開されている部分が多いため、どこの国でも容易に技術キャッチアップできます
  - 核兵器とちがって、全ては情報ないしデータで表現できるので、インターネットによる拡散を防げません

人工知能と軍事

– テロリストでも独裁国家でも敵国でも、容易に人工知能技術を入手し軍事利用できます。

- 核兵器技術はソ連が自前開発したのではなく、あちこちからスパイを使って入手したらしいとのこと（バラット本１４章）。
- ましてやインターネットのこの御時世、簡単に拡散

➢IEEEの倫理基準では、テロリストや独裁国家にAI兵器が渡ってしまった場合に悪用されないような設計をせよ　と技術者に求めているが非現実的

人工知能と軍事

- 人工知能に倫理性を埋めこめばよいという意見もありますが
  - 戦争に使うロボットの場合、自軍のロボットが倫理的に動いても、
  - 敵軍のロボットが倫理を無視する場合には、必敗です。
- 人工知能研究予算の多くはDARPAから
  - 冷戦のころはロシア語の文書を大量かつ容易に解読しようとして、大きな機械翻訳の予算が投入された
  - Siriも開発予算の大きな部分はDARPA
  - DARPAの予算だから成果は当然、軍事技術に転用されています。

人工知能倫理と軍事

- 昔
  - 軍事技術が民生に転用
  - 例：機械翻訳　常に米国の敵国の言語を英語に機械翻訳する技術開発にDARPAの予算が投入された
    - ロシア、日本（経済的競争相手）、アラブ、中国
- 現在、ひょっとすると昔から
  - 民生技術が軍事に転用
  - 例：ドローン、素数理論（暗号）

- デュアルユースへの態度はしっかり考えるべき時期