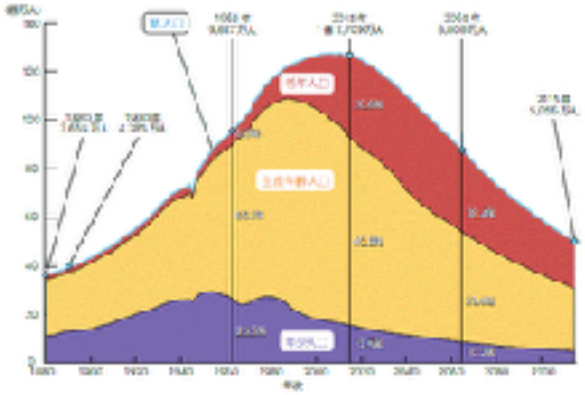
AIのキーワードから倫理とリスクを考える

日本倫理学会／日本経営倫理学会／人工頭脳学会会員　永井郁敏

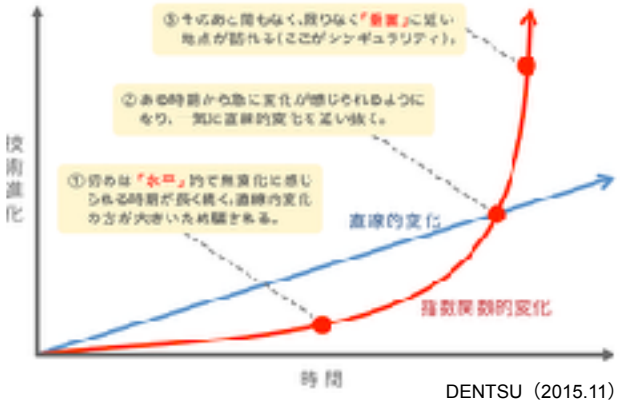
前回は「もう銀行は不祥事なんか起こしている暇はない！」で KEY WORD…BANK／AI／倫理／ディスラプト／リスク／先進技術／GDPR で発表しましたが、BANKと一部AIに言及するに終わった。
今回は、6月の学会の統一テーマである「**AI/ロボット時代における経営倫理**」を踏まえ、そのAIのどこにどの様な倫理問題（リスク）があるか、2018年半ばあたりからのAI技術動向と開発プロセスから考える。

1．なぜAIなのか？（人口と特異点）

- 2015年？が日本の人口のピーク→2030年頃から人口が毎年2百万人減少（オーナス現象）が始まる。



国立社会保障・人口問題研究所（H29）



DENTSU（2015.11）

- 2025年には４人に1人が75歳以上！、2030年頃から人口が一挙に減少する。
- そしてこの2030年頃に、早くもシンギュラリティの到来があるとも言われている（レイ・カーツワイルは様々な技術進化のスピードを詳細に検証した結果、現実世界が技術的特異点を迎える時期を「2045年」と当初予測している）
- この人口オーナスへの対策には、産業構造の転換が必要→人手から機械による自動化、そしてAI＆ロボット化で労働力を補うことが求められている（出典：内閣府：Society5.0）

2．そもそも「AIって何だ？」

- 何がAIなのか？（ロボット？コンピューター？電子脳？）
  - 計算機性能や容量制限等で実現できなかった事が、できるよになった事柄が殆ど（OA=or≠AI?）
  - 少ないデータから答えを導く統計（実験計画法など）から、全数データによる分析（ビッグデータ）へのシフトが、急速にAI技術の発展に寄与。
  - ロボット（自律制御型）はAIコンテナ。AI自体はロボット内部のコンピュータ（頭脳部分）で、ロボが身体を有していれば（人型ロボットなど）、その体がロボット本体。Siriの背後にあるソフトウェアとデータはAIと言える
  - 機械で自動制御し、物を作るオートメーションは人型ではないが、広義にはロボット（ロボティック・プロセス・オートメーション：RPAが代表例）
  - AI技術＝統計（主に多変量解析）＋ビッグデータ（クラウド）＋センサー技術（IoT）＋高速CPU

- AI（artificial intelligence）とは？
  - 人工知能とは、人間にしかできなかったような高度に知的な作業や判断を、コンピュータを中心とする人工的なシステムにより行えるようにしたもの（IT用語辞典）
  - コンピュータがデータを分析し、推論（知識を基に、新しい結論を得ること）や判断、最適化提案、課題定義や解決、学習（情報から将来使えそうな知識を見つけること）等を行う、人間の知的能力を模倣する技術を意味します（IoT用語辞典）
  - 『計算（computation）』という概念と『コンピュータ（computer）』という道具を用いて『知能』を研究する計算機科学（computer science）の一分野」を指す。「言語の理解や推論、問題解決などの知的行動を人間に代わってコンピューターに行わせる技術」、または「計算機（コンピュータ）による知的な情報処理システムの設計や実現に関する研究分野」（Wikipedia）

- AIを考える（言及する）場合の、3つのカテゴリー
  - **狭義の人工知能（ANI）**：チェスや囲碁将棋等、単一機能に特化したAI（自動車はANIの集合体）▶2019年現在のAIレベル（特化型）
  - **汎用人工知能（AGI）**：人間レベルの知性があるAI▶2029年の登場を予測する研究者は48%
  - **超頭脳（ASI）**：AGI登場後に複数のAGI同士で学習する、文字どうりの人類より数億倍？の能力を持つスーパー・インテリジェンス。「科学的創造性・一般的知恵・社会的スキルなど、あらゆる分野で人間の最高の頭脳よりず〜っと賢い知性」とニック・ボストロムは定義している（2001年宇宙の旅に登場する「Hall」やジョニー・デップ主演の「トランセンデンス」？）
  - つまりAI革命は、ANIからAGIそしてASIへの変革だと言える

3．AI開発の歴史(出典：人工知能学会に加筆)

- **人工知能の夜明け「第1次AIブーム」（〜1956）**：哲学・数学・論理学・心理学などの分野で論じられていた「人間の知的活動を行う機械」を作る試みが始められ、56年には「ダートマス会議」で、初めてこの研究分野が“Artificial Intelligence（人工知能）”と呼ばれる。
- **古き良き人工知能（1957〜1969）**：単なる計算しかできなかったコンピュータが少しでも知的なことができるのは驚異的なことでAIの春とも言われた。1969年には最大の難問「フレーム問題」指摘される
- **現実からの反撃（1970〜1979）**：この時期は大きく三つの問題があった。１つ目は初期のAIプログラムが単純な操作だけで動作し、対象に関する知識を持っていなかったこと。２つ目は規模の問題。プログラムが原理的に解を持つことと、プログラムが実際に解を得ることができることは別でした。３つ目は，知的構造を生み出すための基本構造の限界が指摘された。それに対し，どんな問題でも解くことのできる汎用のシステムではなく，対象領域の知識を十分に用いるシステムによって、問題解決する試みが行われた。
- **人工知能の産業化「第2次AIブーム」（1980〜1988）**：商用のデータベースシステムが開発されるようになり、日本でが世界をリードする「第５世代プロジェクト」がある（11年間で540億円の投資）。結果処理スピードで世界一位を獲得したが、AIの観点からは未だ前の前の段階（ICTとしてのセンサー技術や高速ネットワーク技術、そしてビッグデータという言葉すら無かった）
- **現在そして未来の彼方へ（1989〜2000）**：直観によらない厳密な理論や確固とした実験事実をもとに現実の世界の問題を対象とするようになる。だが、バブルの崩壊とパソコンの台頭によりAIに対する投資熱が下がる。コンピュータ業界の資金がパソコンへとシフトし「AI冬の時代」に。
- **直近の20年「第3次AIブーム」（2001〜2020）**：大規模データの格納、高速演算チップ、ネットワーク等の規模とスピードそして低料金に伴ない、IoT（センサー技術）・クラウド（ネットワークストレージ）・ビッグデータ（大規模容量データベース）の３本柱が、AIを牽引し始める。特に、2006年のディープラーニング（深層学習）が、脳科学の解明と合わせた「ニューラルネットワーク」の開発により一気に高まる。
- **未来のAI「シンギュラリティ」（2030〜2045）**：前述シンギュラリティは「技術的特異点」といい、AI（ASI）が人間の知性を超え生活が一変する世界。

4．既に起きているAIによる暴走・事故

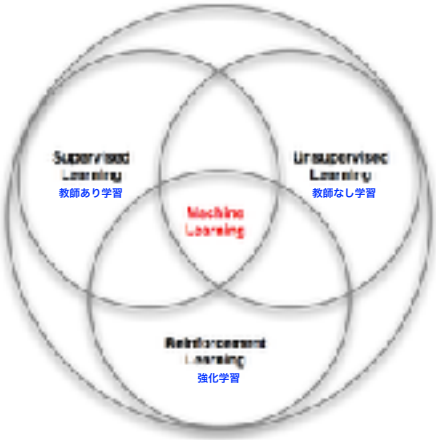
- 2016年3月、Microsoftのチャットロボットによる暴言。チャットは対話型のソフトウェアで、カスタマセンターなどでオペレータの代わりに自動音声対応する。このチャットロボット「Tay」による発言がネットで炎上し、提供開始からわずか16時間後にサービス停止に。この間の約96,000ほどのツイートの半分以上は、暴走によって繰り返された問題発言だった（しかも停止後、自らTwitterに復帰した！）▶意図的に悪意を持った一部のユーザーが、人種差別、性差別と取れる言葉や陰謀論などを学習さたとされている
- 「人工知能」同士を会話させた結果　2017年７月（フェイスブックが開発したAI（人工知能）が人間には理解できない独自の言語で会話をはじめ、同社はこのプロジェクトを緊急停止させた）
- AIロボットがついに「人類を滅亡させる」と発言　2016年3月（抹殺を宣言したのは、Hanson Robotics社が開発した女性型ロボット「Sophia」）
- AIコンピュータ化された証券取引システムにより、わずか数分のうちに巨額の市場下落を引き起こす「フラッシュクラッシュ」が起きた（2010年）
- 日本のバイク（自動２輪車）工場や独フォルクスワーゲン社で、AIマシンによる人身事故（AIが作業員等を邪魔者無用者として認識し排除したことによる）が発生している

5. AI革命がもたらすコト

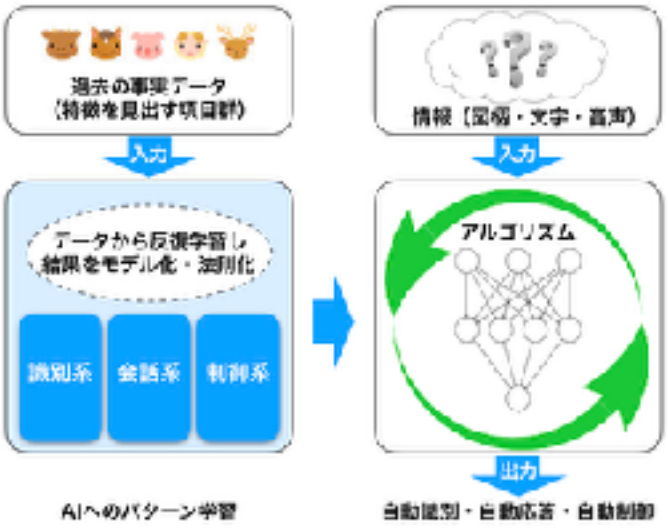
- AI革命の本質は「ホモ・サピエンスを超越する」変革変容を人類が行おうとしていることを、サピエンス全史の著者ユヴァル・ノア・ハラリ は「ホモ・デウス」と称している。つまり人間が「神（デウス）」になると！
- 従来の人類の歴史は進化論的にも緩やかな直線で考えるが、このAI革命はある時点で（シンギュラリティ：科学的特異点）を境に、一気に垂直に進化する（右図の「電脳社会」StageX）
- AIに関係する技術領域
  - 主要な領域として…　脳科学、コンピュータ技術、高速無線NW（5G）、データベース（BIGDATA）、情報分析解析、センサー技術（IoT）、認知科学、現象学、心理学 etc.
- いま言われているAIによる未来（＋）
  - 生まれる職業＋労働時間の短縮化による労働生産性の向上、自動翻訳、コールセンターの自動化、テキストの音読、新薬の開発、プログラム作成、質問に答える、小説を書く、新たな物理＆科学法則の発見、etc.
- いま言われているAIによる未来（－）
  - 失われる職業（既に多数の職業がリストアップされている）、過剰な監視活動からのファシズム化、顔認証や既存データからの誤認や誤入力による事故、AIの過信による事故、AIへの依存が生む思考停止→人間能力低下（既に始まっている？）、Aiが人間を放牧する?、etc.

6. AIの開発プロセス（AIに学習させるために）

- **事前入力：**AI開発を行うには、AI自体に予め判断に必要となる情報を学習させる必要がある。機械が学習するのでマシンラーニング、機械学習と言われている。
- **機械学習（マシンラーニング：ML）**方式は「教師あり学習」「教師ナシ学習」及び「強化学習」に大別される。更に移転学習が最新の学習となっている。MLの教師あり学習モデルのうちの1つが**深層学習（ディープラーニング：DL）**。
- DLは開発者が予め全ての動作を決めておく従来型のプログラムとは異なり、与えられた情報を基に学習し、自律的に法則やルールを見つけ出す手法やプログラム。



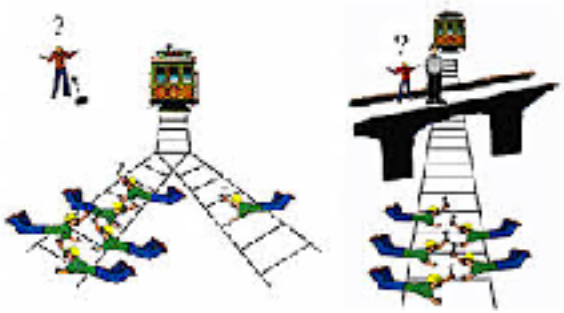
・例えば、従来のMLでバナナを認識させる場合、その形状（ex.細長く黄色いモノ）と教えることで、ナスやキュウリを見たときに「細長いが黄色くない。かつ緑色」だからバナナではないと判断する（▶特徴点（判断基準）を予め入力しておく）
・かたやDLでは、何百枚何千枚・・もの様々なバナナの写真を記憶させるだけ。AI自身がこれらの写真の中からバナナの特徴を見つけ（特異点抽出）、キュウリやナスとの違いを明らかにできる。
・このDLを可能にするものがニューラルネットワーク技術で、ヒトの脳で情報伝達を行うニューロンの仕組みを用い、無数の情報を統合して判断できる（ヒトが勝てるはずがないが、このAI思考こそ重要！）
・このニューロン／ニューラルネットワークの仕組みにより、自ら学習を繰り返すことが可能▶これによりヒト以上のパフォーマンスが可能になる（これは一夜にしてなし得るかも！）
・囲碁や将棋などの特定の対戦ゲームではトッププレイヤーを超える結果も出している。同様に自動運転の分野でも実験段階で人が運転するより安全であることが確認されている。



・更に最近では、将来的な報酬を最大化するためにAI自体が試行錯誤しながら学習する「**強化学習**」や、既に学習したモデルを他の領域に流用する「**転移学習**」などがある。転移学習により学習結果を流用することで、効率的に学習を行うことが可能になっている。既にゲームへの適応がなされており、先のArufago等は既に強化学習が用いられている。

■**自動処理：**学習を行ったAIは、その学習データを基に入力された（与えられた情報：図右上）を、パターン化された学習済みデータ（図左側）と比較し判断を行うプログラム（アルゴリズム：図右下）により結果を出力する。

■事前入力と自動処理に介在するリスクと倫理問題
・学習に用いるデータに何らかの偏り（バイアス）や、学習データ件数そのものが少ない場合にアウトプットに差異を生じる可能性（誤認識、誤作動など）
・モデルを基に入力データから判断を行うアルゴリズムに、判断の偏りが生じる可能性
・悪意を持つ当事者が、学習者として反社会的な行動を「是」として学習させる可能性
・特定の所得層、宗教、病歴等に対する差別の可能性
・フェイク情報の入力による誤認識の可能性
・自動運転車によるトロッコ問題（倫理的優先）



7. トロッコ問題（M・サンデルにより有名に）
トロッコ問題が何なのか？その奥を考えてみる（自律走行車が事故に直面したとき、誰を犠牲にすることが「正解」なのか――。は、前回の例会でも解説があった）

■そもそものトロッコ問題は「ブレーキの壊れたトロッコが暴走している。このまま直進すると、線路上の5人がひき殺されてしまう。トロッコの進路を変えれば5人は助かるが、曲がった先には別の人がいて、その人は死ぬことになる（上左図）」…このトローリー問題は、Philippa Ruth Foot（今日の徳倫理学を築いたうちの一人）が提起し、ジュディス・ジャーヴィス・トムソン 、ピーター・アンガーなどが考察を行う。人間がどのように道徳的ジレンマを解決するかの手がかりとなると考えられており、道徳心理学、神経倫理学では重要な論題として扱われている。また一方、あなたは橋の上で見知らぬ人の横に立ち、トロッコが5人の方に向かっていくのを見ている。トロッコを止める方法は、隣の見知らぬ人（デブ）を橋の上から線路へ突き落とし、トロッコの

進路を阻むことしかない。この問題は「The fat man」と呼ばれ、Judith Jarvis Thomson が提案したトロッコ問題のバリエーションだ（上右図）

- いずれも実際には色々なケースが考えられ、かつそれらの前提が関係する（▶つまり、個々に認識や判断が異なる結果がもたらされる！）→護衛艦をめがけて突っ込む300人乗りのジェット機への対応など

8. 関連領域とアルゴリズム

このトロッコ問題のように、何らかの判断を行うプロセスとして、AIにはアルゴリズム（ex.どうなったらどうするか？を、予め決めておく）がある。

- AIが関係する領域は全て、このアルゴリズムが判断を行う
  - アルゴリズムとは数学、コンピューティング、言語学、あるいは関連する分野において、問題を解くための手順を定式化した形で表現したもの（Wikipedia）
  - ようは、その解を得るための具体的手順および根拠　▶手続き手順・ワークフローのよなもの
  - この手続き手順をいかに齟齬や誤解をなくし（オープン化）、かつ倫理的に実行するコトができるか？現実の世界は既にこの状況に突入している（米国警察捜査でのAI活用における黒人の検挙数）
  - そのアルゴリズムを具体的に記述するプログラムには、ex. Python, R, Julia 等がある
- アルゴリズムに潜む課題
  - これらの手続きから、どの様に汎用性を維持できるのか？▶つまり、誰かがアルゴリズムを考えた時点で何らかの「バイアス」が存在（発生）する（▶前述のトロッコ問題など）
    - アルゴリズムからプログラム化する時点での誤認識（開発者の判断に潜む）による偏り
    - プログラム自体に介在する人的バグ（将来はAIがプログラムするので改善？）
  - あるものができたら⭕、できなければ❌（２言論に潜むリスク）
  - そもそも倫理自体に正解がないこと、そして何より…
  - 倫理のあり方は時代時代で変わる（ex. 30年前避妊治療はNG▶今保険適用。遺伝子組替え操作は？）
  - セキュリティ課題▶アルゴリズムがハッキングされ、悪意のあるアルゴリズムへの書き換え
  - したがって現時点では、セーフガードをいつどの様に起動させるかの検討が必要
  - Bitcoinの技術基盤であるBlockchainのスマートコントラクトは、アルゴリズムの塊

9. 認知バイアス

- 所詮この社会はバイアスで出来上がっている（▶視点の違い、立場、環境、教育、時代の違いなどによる見識の違い。宗教、民族、国、政治思想とうの違いなど）
- どうすれば偏りをなくすコトができるか（＝バイアスの排除）
  - （私的意見）無理では無いか？（何故なら→アルゴリズムを作るのは他ならぬ人間だから）
  - AI技術を悪用されたら困る！→これも無理（SFに出現する「悪」は皆これ！）
  - さらに時代が進めば、このアルゴリズムやプログラムをAIが作るよになる時代が来る（probably）
  - でもこのAIも誰かが作っているので、何らかのバイアスは必ず存在する（maybe）

- なので、何が求められるのか？
  - いついかなる場合でも、常に立ち戻れるベース基盤（規範など）が必要（ex.倫理綱領？）
  - 一旦作った基盤は、時代の時々で必要に見直すことにより「成長」を続ける必要がる
  - 原子力技術は、発電に用いれば電気を生むが（表面的ですが）原爆にもなり得る！
  - AIに於いても然り！（毒にも薬にもなり得る）
- どんな倫理がどこに求められるのか？
  - テロリストに渡さない（でも超優秀なテロリストは存在する！、ハッカー然り！！）
  - 自動運転（トロッコ問題への対応▶必要がない！▶なぜなら、正解が明確ではなく、そもそも正解がないことと、自律自動車自体がもはやセンサーマシンであるため、トロッコ問題のような状況が実際に生じる可能性はとても低い。人がハンドルを握っている方がはるかに危険！
    - いまのところ、排気ガスのような黒煙と壁の判断は難しい。同様に億万長者とホームレスも
    - でも実際のプログラミング（アルゴリズム）においては、なんらかのトレードオフを自律自動車に学習させてる可能性はある
  - 学習させる内容の品質管理が求められるが、バイアスの存在は自分自身では気づかないことが殆ど
  - 倫理に正解はない（誰の倫理なのか？正解がないという課題）
  - 文化的・地域的（民族的）バイアスについては、**予測市場**で考える方法も
  - 倫理観は「時代時代位で変化する」
- 海外の状況から遅れすぎていて見えていない日本
  - 倫理観は国家の品格（改めて、これからのあるべき姿を再定義する時期に来ている！）
  - スコアシステムにより、国もスコア化により評価＆ランク付けされる
  - どうすれば倫理観を醸成できるのか？（▶先ずは考える力をつける（哲学的思考）と規範基盤作り）
- 数学（代数）的思考がもたらすリスク?!
  - a=b は b=a か？ x=3 の理解が倫理観を決める？▶a=bならbは不要だ！と考える
  - 感覚的に理解できないことが大事▶学校では落ちこぼれになるが▶しかし、大成するコトも！
  - x←3とすれば…という代数や方程式を感覚的に受け入れることは「**何でもOK」に繋がる！**
  - つまり、この部分に本質的倫理観が潜んでいる？（NHK:AIって何だ?!養老孟司）
  - 生きるための恒常性（ホメオスタシス）が大切

10.AIが求める人間社会の再定義

- 今はまさしく変容の時（DX：デジタル・トランスフォーメーション）。つまり蛹から蝶へと移りゆくその端境期だと言える。この端境期は元に戻ることが許されない不可逆的な変革（変化と変容は異なります！）

なので…

- AI化により哲学的視点から現在曖昧にされて来た部分（事柄）の定義化、或いは再定義が必要
- いま求められる議論、しなければならない事柄は何か？▶存在するバイアスの認識とその最適化（つまり状況への適合と合意）が求められる
- それは、人類として「どういう未来を実現したいのか？しなければならないのか？」の合意
- 何故なら、未来は選択によって創られるから！
- 人類に残された「AIを信用しない、使わない、という選択肢」も（Will Knight）
- 倫理が邪魔をする科学の進歩？ 科学者であれば実現したい！（ex. 中国人の遺伝子学者による遺伝子の書き換えベイビーの実現：Mad Scentist？）
- 人類は神の領域に突入した（ハラリ）
  - 人間って？　AIって？　倫理って？　哲学って？　お金って？　労働って？　誕生って？　死って？…

- AIをどうしてそれほどに心配しなければならないのでしょう？人間はいつでも好きなときに、AIのプラグを抜くことができるのでは？・・・（スティーヴン・ホーキング）
  - 人間がコンピュータに尋ねた「神は存在するか？」コンピュータはこう答えた。「いまや、神は此処にいる」。そしてプラグのヒューズを飛ばした。
  - ステファン・スティーヴン・ホーキングは彼の最後の著書「ビッグ・クエッション」で、AIに対する危険をつぎの様に述べている。「AIの性能が上がって、加速度的に自らを再設計できる様になることだ。ゆっくりとした生物学的進化の速度に制約された人類は、そんなAIに太刀打ちできず、AIに取って代わられるだろう。そして将来的には、AIは自分自身の意思を持つようになり、私たちの意思と対立する様になるだろう…」（中略）「数学者のアーヴィング・グッドは、1965年に人間を超える知性を持つ機械は、自らの設計を反復的に改良できることに気づいた。SF作家のヴァーナー・ヴィンジがいうところのシンギュラリティだ。そんなテクノロジーが金融市場で賢く立ち回り、人間の研究者よりも優れた発明をし、人間の指導者よりも人身操作に長けていて、私たちには理解することさえできない兵器を使って人間を征服するというのも想像できないことではない。AIの短期的な影響は、誰がそれをコントロールするかにかかっており、長期的な影響は、AIはそもそもコントロール可能かどうかにかかっている。一言で言えば、**スーパーインテリジェンスな（超知能を持つ）AIの到来は、人類に起こる最善の出来事になるか、または最悪の出来事になるだろうということだ**」

11.AIが意識を持つか？（AIの人格って？）

- Aha! Gocha!といった閃きや築きを、AIが持つに至るのは当分無理（…永遠にかもしれない）
- 2019年現在、人格を持つAIは登場していない。あたかも人格を持っているかの様に振る舞うAIは存在しているが、1966年にMITで開発された「ELIZA」を高度化したものだ（例えば人がイライザに「腹が痛い」と言えば、イライザは「なぜ腹が痛いのか？」と返す。これは確かに会話だ。でもこれはイライザに多数の想定した会話パターンを仕込んでおいたためで、想定されたパターンに則した質問を続けると、あたかも会話が成立したかの様に見える）
- 予め想定した判断から、深層学習からのAIによる判断が行われている（この部分がブラックボックスであることから、説明責任が要求されている）
- つまり人格を有する様な振る舞いをさせるコトにも、倫理観を持たせる必要がある（ex.Tay問題）
- 結局今の所、AIが自ら学習はするが、意識（広義の意思）は製作者の学習させるデータが基本だ。
- したがって、AIを製作する科学者への倫理的感覚の醸成が必要
- でも…（Evil：マッドサイエンティストは必ず現れる！）

12.役に立つAIのために

- AIがもはや社会の要請だとした場合、開発者やサービス提供者には説明責任が求められる
  - 技術的な問題の説明／サービス事業者への説明／消費者への説明
  - すでに深層学習のレベルは1万層以上になっており、その因果関係は移転学習などにより、部分的に情報の消滅が発生する問題もあると言われている
- 社会に役立つ人工知能を開発するためには、情報システムの全てに言及できるが、「どうしてそのシステム（AI）が必要なのか？」を議論する必要がある
- 認知バイアス（思考の偏り）を意識し、専門分野に偏った思考フレームを形成していないかを、常にチェックする
- ただし、社会を発展させるための科学を、倫理が抑圧することは、科学技術の将来にとってマイナス
- 科学と倫理のハイブリッド感（バランス感覚）が必要！▶科学者の倫理教育と倫理学者の科学的教育
- 金沢工業大学での取り組み（ex.BERCとのタイアップ研修ほか）。イリノイ工科大学のMichael Davisが考案したセブン・ステップ・ガイドを補完し、金沢工業大学版の実践的・体系的な取り組みを行っている（左下）

コース	レベル	モジュール
CS 134：ネットワーク	上級学部生	フェースブック、偽のニュース、そして検閲の倫理
CS 6：システムプログラミングと機械編成	学部生	ASCII、Unicode、および自然言語表現の倫理
CS 152：プログラミング言語	上級学部生	ソフトウェアの検証と妥当性における倫理
CS 165：データシステム	上級学部生	データシステム設計におけるプライバシー
CS 181：機械学習	上級学部生	識別と機械学習
CS 263：システムセキュリティ	卒業	ハッキングバックの倫理

- ハーバード大学の科学と倫理を同時に学べる「Embedded EthiCS（右上）」では、具体的な事例を基とに何が問題なのか？から初めているところが、多民族国家らしいプログラムだ。

- これらの情操教育の観点から、MIT media labの伊藤穣一は「情熱・仲間・プロジェクト・遊び」が求められるとしている。ここでのプロジェクトの意味は、目的・目標を共有する、といった意味ではないかと考える。
- 当たり前が崩壊する時代、新しい当たり前をデザインする力が、いま求められる。

◆

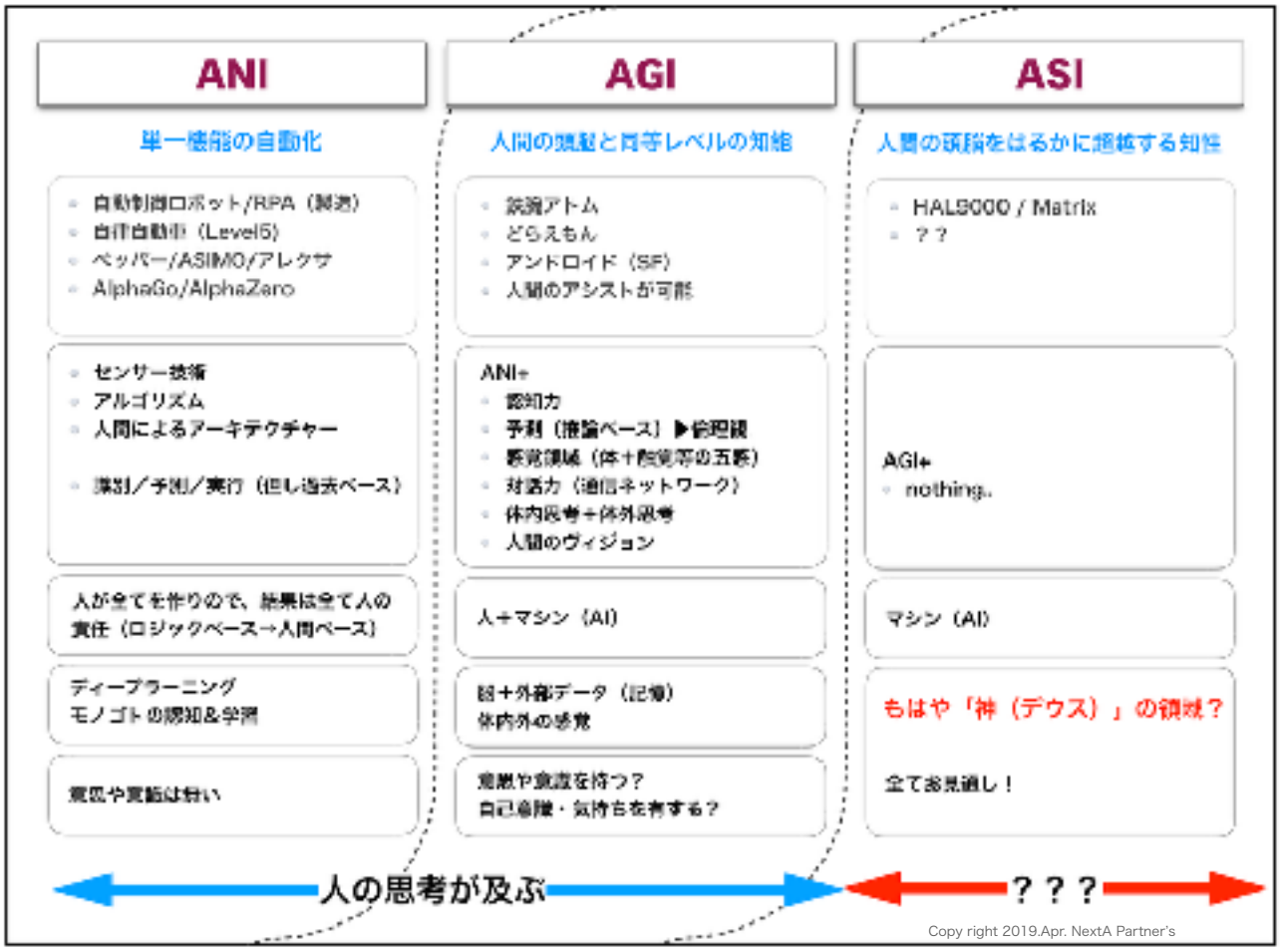
参考文献

- 国立社会保障・人口問題研究所：http://www.ipss.go.jp/syoushika/tohkei/Mainmenu.asp
- DENTSU「シンギュラリティ」という壮大な仮説 真の脅威はその「検証力」にあり：https://dentsu-ho.com/articles/3260
- 内閣府 Society5.0：https://www8.cao.go.jp/cstp/society5_0/society5_0.pdf
- 日本人工知能学会 学会誌「人工知能」2019.3 No.2 「道徳判断の自動化をめぐる問題：規範の選択と協力の進化」
- 日本人工知能学会 What's AI：http://www.ai-gakkai.or.jp/whatsai/
- マサチューセッツ工科大学(MIT)メディアラボ：https://www.media.mit.edu/research/?filter=groups
- 人間の倫理は非理性的か：「トロッコ問題」が示すパラドックス, Wired 2008：https://wired.jp/2008/11/11/
- NHK「AIって何だ」ゲスト養老孟司談：https://www.nhk-ondemand.jp/goods/G2019095566SC000/index.html
- 「ファスト&スロー あなたの意思はどのように決まるか？」ダニエル・カーネマン, 2014/6, 早川書房
- 「ビッグ・クエスチョン―〈人類の難問〉に答えよう」スティーヴン・ホーキング (著), 青木 薫 (翻訳), NHK出版 (2019/3/14)
- 「人工知能のための哲学塾」三宅陽一郎, ビー・エヌ・エヌ新社 (2016/8/11)
- 「サピエンス全史」「ホモ・デウス」　ユヴァル・ノア・ハラリ, 河出書房新社 (2016/9/8)
- 「科学技術者倫理」金沢工業大学科学技術応用倫理研究所編：白桃書房 (2017/4/16)
- 「Embedded EthiCS」ハーバード大学：https://embeddedethics.seas.harvard.edu/index.html
- 根拠を説明できないAIが招く“人工無能”な組織の懸念：日刊工場新聞（2019.02.03）
- いまテクノロジーには哲学とSFからの問いが必要だ：WIRED JAPAN（2019.02.19）
- 脳がつくる倫理：科学と哲学から道徳の起源にせまる：note, Tetsusan（2019.02.20）
- 人工知能（AI）と哲学：クライテリオン, 川端 祐一郎（京都大学大学院助教）
- AIは「インテリジェント」なだけ―倫理的なAIには倫理的な人間が必要：TechRepublic Japan, Garry Kasparov
- 「自動運転車」に倫理を教えることはできるのか：NewsPicks（2018.11.04）
- AI Business Canvas：永井郁敏（2019.03.01）

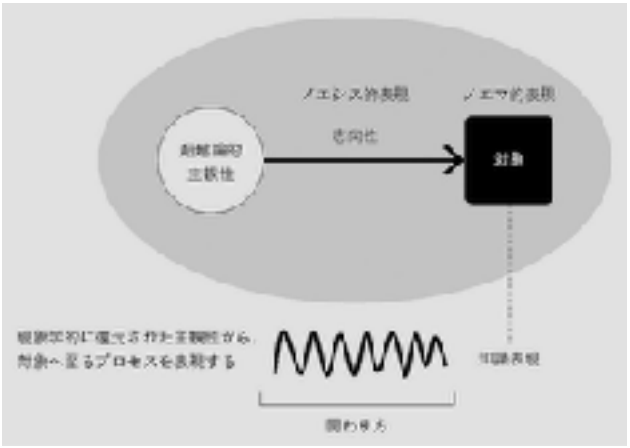
以上

Appendix:

以下は、現在言及されているAI三形態の比較（永井作表）



以下、三宅陽一郎. 人工知能のための哲学塾より
三宅は、自身の携わるゲームのキャラクターに、汎用人工知能（意識）を持たせるため、現象学からアプローチし、「人工知能のための哲学塾」を出版。その続編として「東洋哲学編」もある。

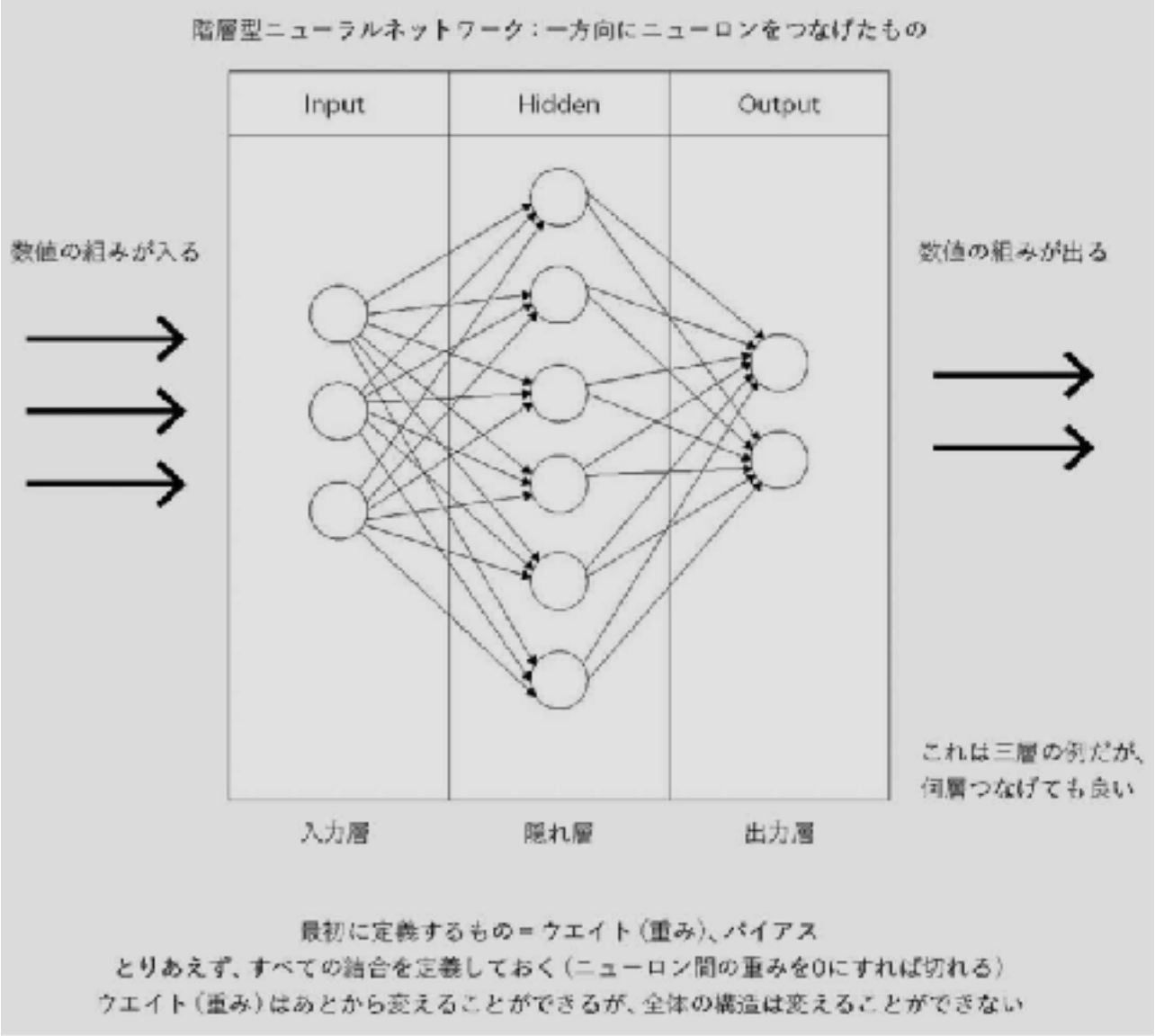


◉単なる知識表現(Knowledge Representation)
–グラス20センチメートル
–杏酒がグラスに入っている
–レモンがグラスに入っている
–氷がグラスに入っている
–マイナス５度
–グラスを持つことができる(アフォーダンス)
–飲むことができる(アフォーダンス）アフォーダンス＝環境が動物に対して与える「意味」のこと

◉志向表現（Intention Representation）
（上記の知識表現に加え…）
–この輝きは黄金に似ているなぁ
–去年シカゴで飲んだリキュールと同じ味だ
–レモンはシチリア産かなぁ
–グラスの表面はツルツルで亀の甲らの様だ

三宅は、下線で追加された部分がまさしく志向的な部分で、個人的な経験や感情に基づく記憶だとしている。

2015年に、イギリスのDeepMind社によって作られたDQN（DeepQNetwork）が「スペースインベーダー」や「ブロック崩し」といったAtariのクラッシックゲームを一人でプレイし、人間が手を加えなくてもハイスコアを出す方法を学習できるレベルに達したことが話題になりました。このDQNは、ニューラルネットワークの一種になります。ニューラルネットワークはシンボルを使って構築される人工知能、端的にいえば、プログラム言語によってロジックが組まれた人工知能とはまったく違う、独立した大きな分野を形成します。第一夜のコラムでもふれたように、「コネクショニズム」と呼ばれることもあります。生物の脳がニューロンと呼ばれる神経細胞から構築されていることは一八世紀から知られていました。それぞれのニューロンは「軸索」という複数の足を持っていて、それによって他のニューロンと接続されています。そして、そのニューロンのネットワークの上を電気信号が伝わっていきます。つまり、細胞でできた電気回路なのです。ただ、それぞれのニューロンは一定の電気がたまるまで電気を発信しないので、電子回路よりゆっくりとした電気回路になります。（同コラム「ニューラルネットは汎用人工知能を実現するか？」三宅陽一郎. 人工知能のための哲学塾 (Japanese Edition) (Kindle の位置No.776-787). Kindle 版. ）



つまり、人工知能には１つの現象から、様々な志向の違いによる部分を取り込んでいくことが求められる。その実現手法の１つであるパーセプトロンを構成するニューラルネットワーク（この場合はInput/Hidden/Outputの３層。４層以上の物を現在「深層学習：ディープラーニング」と称する）。