資料３

2018.12.18 AIネットワーク社会推進会議　AIガバナンス検討会

AIガバナンスの倫理的側面

河島茂生(青山学院女子短期大学，理化学研究所)

人間と機械との異同

人間と機械との同質性(1)

◆ 物質・エネルギー／情報 ⟹ 人間と機械との同質性

生物と機械の同一視は，「情報」という語がかなりいい加減に使われていることも係っている。

「エネルギー・物質」とは違ったものという共通項はあるものの，実にさまざま概念が「情報」という一語でまとめられてしまっている。それゆえ機械も生物も同じ情報変換体として一括りにまとめることが可能になった(西垣, 2004)。

意味を捨象した確率論的な定義 ⟸ 情報 ⟹ 生物の生存に本質的に係る価値

人間と機械との同質性(2)

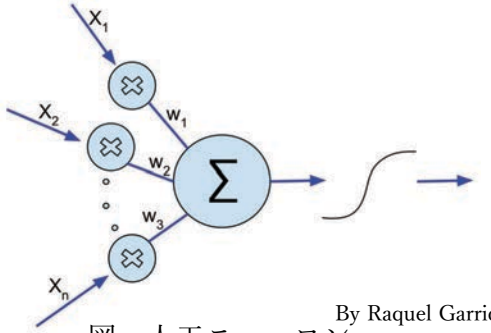
◆ 物質・エネルギー／情報 → 人間と機械との同質性

◆ フィードバック機構

サイバネティクスの草創期に活躍したロス・アシュビーは，撹乱されても平衡状態を保つ装置ホメオスタットを開発し，その4つの箱などからなる機械を「生きている」と表現し，目標値との差分をなくすメカニズムを「思考」と呼んだ(Rid, 2016=2017)。

◆ 人間のニューロンの論理的モデル化(McCulloch & Pitts, 1943)

実際のニューロンでなくても，ニューロンが行っている計算が実行できればよい。入力に重みづけを行い，最終的な出力が予期されたものであれば，その人工ニューロンは立派に考えている。



By Raquel Garrido Alhama

図　人工ニューロン

人間と機械との同質性(3)

- コンピュータの理論モデル(万能チューリングマシン，ノイマン型コンピュータ)は，人間のあらゆる論理的思考を0/1のパターン変換に落とし込んだものであり，その意味ではコンピュータは元来，人間の知能を機械的に表現したものと捉えられる(西垣，1991)。

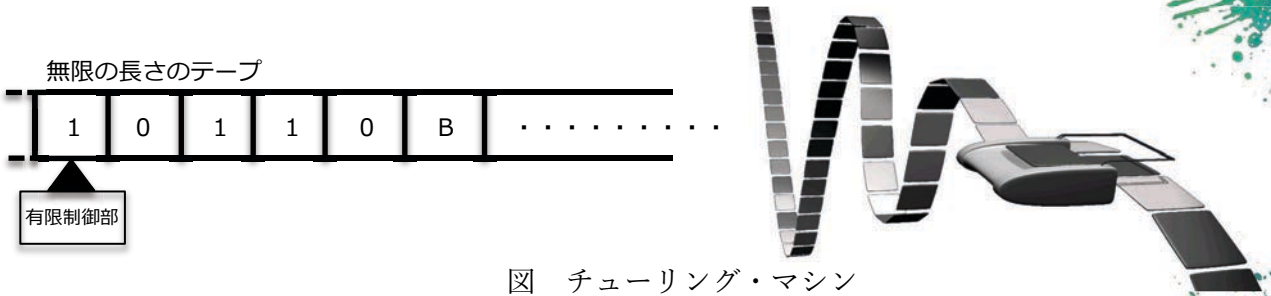


図　チューリング・マシン

- 20世紀中葉以降は特に、紆余曲折を経ながらもAI・ロボットの制作が精力的に行われるとともに、さまざまな小説や漫画、アニメ、映画などで機械的な「生命」「知能」が多く描かれることになった。

人間と機械との異同

人間と機械との違いは，論理的計算，フィードバックや自己複製，学習，ニューロンの働きには求められない。そのような特徴に着目するならば，人間と機械の相違点は消失する。

人間と機械との違いは，オートポイエーシスに見いだせる。この点を踏まえなければ，人間と機械は同質のものとなり，人間を機械のように働かせても責められない。

	人間	機械	機械の例
論理	○	○	コンピュータ
フィードバック	○	○	冷蔵庫，エアコン
自己複製	○	○	コンピュータのミラーリング，バックアップ
学習	○	○	迷惑メール・フィルター
ニューロンの働き	○	○	ニューラル・ネットワーク
オートポイエーシス	○	X	

人間と機械との異質性(1)

- オートポイエーシス理論
  - セカンド・オーダー・サイバネティクス(ネオ・サイバネティクス)の枢要な一角を占める。
  - ウンベルト・マトゥラーナとフランシスコ・ヴァレラによるオートポイエーシス論は，生物もマシンと捉えるが，それは機械とは別種のマシンであり，自分で自分を作っていく性質が生物の特徴であることを理論化した(Maturana & Varela, 1980=1991)。

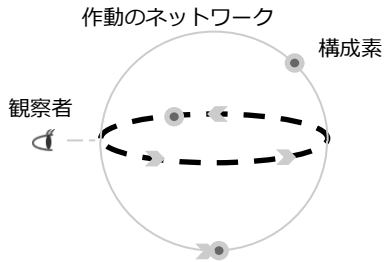


図　オートポイエティック・システム

例: 細胞，免疫システム，心，社会

オートポイエティック・システムは，自分で自分(auto)を制作(poiesis)しながら円環的に内閉したシステムである。

マトゥラーナとヴァレラが主に細胞や神経系、生物個体の認知機能に関する研究をもとにして学術的に定立した。オートポイエティック・システムは，生物の十分かつ必要な条件を兼ね備えている。

人間と機械との異質性(2)

◆　アロポイエティック・システム

アロポイエティック・システムは、オートポイエティック・システムの反対概念であり、「自動車のように、その機能が自分自身とは異なったものを産出する機械」(Maturana & Varela, 1980:135= 1991:242)であって、入出力関係に従属して動作するシステムである。すなわち、アロポイエティック・マシンは、開放システムであり、ある入力(input)をすれば、常に一定の出力(output)をするように調整されているシステムである。

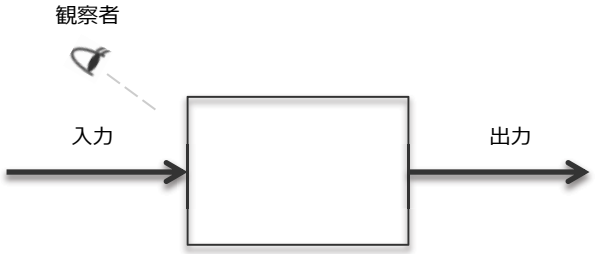


図　アロポイエティック・システム

例: エアコン，自動車，人工知能

オートポイエーシス理論を踏まえると，AIやロボットはまだ生物の条件を兼ね備えていない。というのも，人間がデータを大量に収集したり機械学習のソフトウェアを作り精度を確認して調整したり，あるいは多額の資金を投じて最先端のハードウェアを用意しているからである(河島, 2016)。

AI利活用原則に関して

技術系の倫理綱領の必要性

アメリカの技術アカデミーは，20世紀の工学における大きな成果を選出した(National Academy of Sciences on behalf of the National Academy of Engineering, 2000)。

Electrification	Highways
Automobile	Spacecraft
Airplane	Internet
Water Supply and Distribution	Imaging
Electronics	Household Appliances
Radio and Television	Health Technologies
Agricultural Mechanization	Petroleum and Petrochemical Technologies
Computers	Laser and Fiber Optics
Telephone	Nuclear Technologies
Air Conditioning and Refrigeration	High-performance Materials

技術系の倫理綱領等の策定

■海外

1847　アメリカ医師会

1912　AIEE(現在のIEEE)

■日本

1938　土木学会

1961　日本技術士会

1996　情報処理学会

1998　電子情報通信学会　電気学会

1999　日本機械学会　建築学会

2000　日本化学会　日本塑性加工学会

2001　日本原子力学会　映像情報メディア学会

AIに係る倫理綱領等の策定

- ◆AIネットワーク社会推進会議「AI開発原則」(案)
- ◆AIネットワーク社会推進会議「AI利活用原則」(案)
- ◆人工知能学会「人工知能学会 倫理指針」
- ◆The IEEE Standards Association “Ethically Aligned Design”
- ◆ FLI“ASILOMAR AI PRINCIPLES”
- ◆The White House “Preparing for the Future of Artificial Intelligence”
- ◆House of Commons Science and Technology Committee “Robotics and artificial intelligence”
- ◆European Parliament “Report with recommendations to the Commission on Civil Law Rules on Robotics”
- ◆Stanford University AI100 “ARTIFICIAL INTELLIGENCE AND LIFE IN 2030”

……

AI時代においても人間の尊厳を守り，人間と機械との協調が求められている。

尊厳・自律の原則

ア　人間の尊厳と個人の自律の尊重

- 人間と機械との異質性を前提にAIを利活用することが求められる。
- 人生を左右しかねない意思決定(採用・人事評価・与信・入試判定等)に係るAIの利用にあたっては人間の介在が求められる。

イ　AIによる意思決定・感情の操作等への留意

- 1990年代以降，インターネット嗜癖が広まっている。
- このようなリスクに関しては教育の現場での取り組み，AIシステムを含んだコンピュータ・システムの開発側によるインターネット依存への対応，利用者の自覚が求められる。

ウ　AIと人間の脳・身体を連携する際の生命倫理等の議論の参照

- エンハンスメント(健康の維持や回復を超えた人間の能力の増進の追及)に関しては人間の自律性が他律化されてしまう懸念がある。

公平性の原則(1)

- 監視選別社会による個人の尊厳と人間の平等性への影響
  - 企業の営業秘密などからくるスコア計算の不透明さ
  - スコア算出のための変数の扱い(特に生物的特性を示す変数，ビッグデータによるプロファイリング)
  - データの偏りや間違い(Federal Trade Commission, 2016)
  - AIによる判定への依存
    - AIがあらゆることを自動で行うことが夢想される結果，意思決定の責任をAIのせいにして自分(たち)の責任逃れをするために使われる恐れがある。

「機械崇拝者たちが機械を賛美する動機の一つは，機械は人間のもつスピードと精度の限界に制約されないということだが，それに加えてもう一つ，具体的に立証することは困難だが，にもかかわらずなかなか重要な役割りを果たしているにちがいない動機がある。それは，危険な決定や破滅的な決定を下すことに対する個人的責任を他に転嫁することによって避けたいという願望である」(Wiener, 1964=1965: 58-59)

⇨人間の介在の必要性

公平性の原則(2)

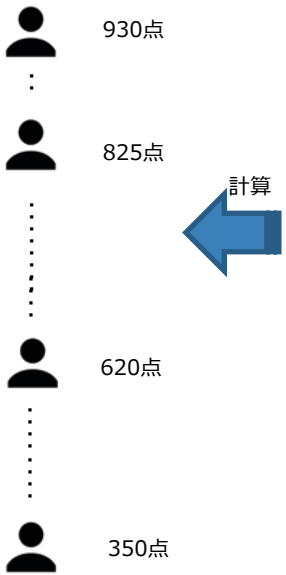
◆　監視選別社会による個人の尊厳と人間の平等性への影響

◆低スコアもしくはスコア拒否による社会からの連鎖的排除の懸念があるため，社会への包摂(inclusion)をも含めた原則へ

◆全人的な点数化により，低い点数の人への差別が起き，また自分自身も自尊心を失う恐れ

◆公平(fairness)は，言語辞書上では偏りがなく正しい状態を意味している。

◆社会経済的に恵まれない立場の人を考え，単なる能力主義を超えた公平性を含めて定義づける必要性(尊厳・自律の原則との連関)。



年齢
身長
声色
出身地
居住地
学歴
婚姻の有無
職業
勤続年数
年収
テレビの大きさ
趣味，習い事
家族構成
ペットの有無
買い物履歴
ボランティア活動
ケータイのキャリア
SNSの投稿内容，友達
ゲーム時間
公共料金の支払い
法律違反
ローンの借り入れ
・・・

公平性の原則(3)

◆　監視選別社会による個人の尊厳と人間の平等性への影響

◆可逆性テストに受かるかどうかを考えてみてもよい。可逆性テストとは自分を相手の立場においてみて嫌だと思う行為か否かをチェックするテストを指す。

◆公正世界信念(belief in a just world)が強すぎると，弊害が起きることも指摘されている。

■公正世界信念の質問例

		あてはまる	どちらかといえばあてはまる	どちらともいえない	どちらかといえばあてはまらない	あてはまらない
1)	この世の中では、努力はいつか報われるようになっている	1	2	3	4	5
2)	この世の中では、努力や実力が報われない人も数多くいる(逆転項目)	1	2	3	4	5
3)	この世の中では、悪いことをしたものは必ずその酬いを受ける	1	2	3	4	5
4)	この世の中では、悪いことや間違ったことをしても見逃される人が数多くいる(逆転項目)	1	2	3	4	5

利用者別のAI内部のメカニズムに対して求められる理解度

◆AIサービス・プロバイダ＝いかなるアルゴリズムが裏で動いているかまで把握して提供することが求められる。AIシステムがいかなる影響をステイクホルダー(最終利用者，間接利用者および第三者)に与えるかを考慮しつつ，AIサービスを提供すること。

◆最終利用者(医者や金融機関，人事担当，入試担当等)＝読み込まれているデータの種類やその代表性等に留意し，利用することが求められる。間接利用者の求めに応じ，他者のプライバシーや企業の営業秘密を損なわないかぎりにおいて情報を公開すること。間接利用者の人生を左右するような決定に関しては，人が介在すべき。いかなる影響をステイクホルダー(間接利用者，第三者)に与えるかを考慮しつつ，AIサービスを提供すること。
　ただし，自動運転タクシーの乗客，異常検知にAIシステムを利用する製造メーカー，AIアシスタント対応のAIスピーカーを利用する個人等は除く。

◆間接利用者＝第3次ブームのAIが確率論的推論に基づいており，機械学習や読み込むデータによりAIの判定が異なることに留意し，必要に応じて最終利用者に問い合わせを行うこと。

AIネットワーク化における集合的責任

AIネットワーク化における集合的責任(1)

- 過失が個人の判断に求められる場合

  ⇒ 個人に倫理的責任を帰属する

- 個人の悪意や過失が同定できない場合

  ⇒ 社会に倫理的責任を帰属する

(社会的背景)

- 複数のAIがネットワーク化し連携しながら動くことが想定されている段階で，個々のエンジニアや運営者に瑕疵が認められない場合でも，他者の人生や生命に強く影響を与えるような，誤った動きが起きることが予想される。
- また，通信ネットワークを介した事件ではデジタル・フォレンジックの限界がすでに露呈しており，誰が行った殺害予告・サイバーテロなのかを特定できないケースが相次いでいる。

AIネットワーク化における集合的責任(2)

- 過失が個人の判断に求められる場合

  ⇒ 個人に倫理的責任を帰属する

- 個人の悪意や過失が同定できない場合

  ⇒ 社会に倫理的責任を帰属する

（個人に倫理的責任を課す場合のデメリット）

- なにも過失がない場合にも，エンジニアや運営者が責任追及されるとすれば，それはAI開発および利用の萎縮につながり，社会的な損失ともなる。

（どこにも倫理的責任を帰属しない場合のデメリット）

- 被害者が救済されない事態を招く。AIの予期せぬ動きで，身体に危害が加えられたり人生を狂わされたりする人が生じた場合，そうした人たちを救済する仕組みは欠かすことができない。

AIネットワーク化における集合的責任(3)

- 過失が個人の判断に求められる場合

  ➡ 個人に倫理的責任を帰属する

- 個人の悪意や過失が同定できない場合

  ➡ 社会に倫理的責任を帰属する

(道徳的行為者として社会を定位)

- 社会それ自体が一種の道義的責任(補償的責任)を担い，損害を被った人に補償していく制度の構築が望まれる。税金や保険，業界の組合などによる補償が考えられる。
- 「ハッキングにより引き起こされた事故の損害(自動車の保有者が運行供用者責任を負わない場合)に関しては，政府保障事業で対応することが妥当であると考えられる。他方，例えば，自動車の保有者等が必要なセキュリティ上の対策を講じておらず保守点検義務違反が認められる場合には上記の通りではないと考えられる」(高度情報通信ネットワーク社会推進戦略本部・官民データ活用推進戦略会議「自動運転に係る制度大綱」2018: 18)

参考文献

- Federal Trade Commission(2016) “Big Data”.　https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf　アクセス日: 2018/10/1
- Johnson, D. G. & Noorman, M. (2015) “Recommendations for Future Development of Artificial Agents”.
- 河島茂生「ネオ・サイバネティクスの理論に依拠した人工知能の倫理的問題の基礎づけ」『社会情報学』5巻2号, pp.53-69,　2016年。
- 河島茂生「ビッグデータ型人工知能時代における情報倫理」『基礎情報学のフロンティア』東京大学出版会, pp.59-79,　2018年。
- Maturana, Humberto R., Varela, Francisco J., Autopoiesis and Cognition : the Realization of the Living, D.Reidel Publishing Company, 1980. 河本英夫（訳）『オートポイエーシス：生命システムとはなにか』国文社，1991年。
- McCulloch, Warren and Pitts, Walter,“A Logical Calculus of the Ideas Immanent in Nervous Activity”in *The Bulletin of Mathematical Biophysics*, Vol.5, No.4, 115-133, 1943年。
- National Academy of Sciences on behalf of the National Academy of Engineering(2000)“Greatest Engineering Achievements of the Twentieth Century” http://www.greatachievements.org/Object.File/Master/4/024/Feb22Release.pdf　アクセス日: 2018/10/1
- 西垣通『デジタル・ナルシス：情報科学パイオニアたちの欲望』岩波書店，1991年。
- 西垣通『基礎情報学』NTT出版, 2004年。
- Rid, Thomas, Rise of the Machines : a Cybernetic History, W.W. Norton, 2016. 松浦俊輔（訳）『サイバネティクス全史：人類は思考するマシンに何を夢見たのか』作品社，2017年。
- Wiener, Norbert, God and Golem, inc. : a Comment on Certain Points Where Cybernetics Impinges on Religion, M.I.T. Press, 1964. 鎮目恭夫（訳）『科学と神：サイバネテイックスと宗教』みすず書房，1965年。