［招待論文］

AI 社会における「人間中心」なるものの位置づけ
“Human Orientation” in an AI society: Consideration of AI's Code of Ethics

河島 茂生†
Shigeo KAWASHIMA

†青山学院女子短期大学 現代教養学科／理化学研究所革新知能統合研究センター
†Aoyama Gakuin Women’s Junior College / RIKEN Center for Advanced Intelligence Project

要旨

本論文は，人間と機械との同質性／異質性を整理したうえで AI の倫理綱領等を考察している．ネオ・サイバネティクスの理論に基づき人間と機械との間に違いが明確にあることを示し，「人間中心」社会の倫理的基盤を確認した．その後，AI ネットワーク社会推進会議が提案した利活用原則（案）の公平性の原則を検討した．公平性の原則で論点として挙げられている「人間の判断の介在」の意義は，人間の人生を左右する判断に関しては，機械の異常品を検知するのとは違い，機械に責任転嫁せず，人間の責任で行うことが求められるということである．唯一性を備えたオートポイエティック・システムの集合体たる人間に対する重要な意思決定については，同じくオートポイエティック・システムの集合体である人間が覚悟をもって行うべきものである．また公平性には，社会的な不安定性が増しているなかで，人々が社会的排除に陥るのを防ぐ包摂の考え方が含まれることが望ましい．

Abstract

This paper considers AI's Code of Ethics after accounting for the homogeneity/heterogeneity between human and machine. Based on the theory of neocybernetics, the difference between humans and machines is clearly shown and the ethical basis of “human oriented” society is confirmed. Next, the principle of fairness in the Draft AI Utilization Principles proposed in the conference “The Conference toward AI Network Society” is examined. The main point discussed in the principle of fairness in “Human Intervention” is that judgments that affect human life require human beings to bear ethical responsibility for them, not pass it on to the machine. Regarding important decisions that affect human beings, which are an aggregate of the uniqueness of autopoietic systems, it is necessary for humans, who are also aggregates of autopoietic systems, to take ethical responsibility. Moreover, the principle of fairness should encompass inclusive thinking, which would prevent people from falling into social exclusion while increasing social instability.

1. 問題の所在

1.1. 研究の背景

社会において工学的技術が組み込まれて日常生活に広く深く浸透していくにつれて，人間に配慮した科学技術が倫理的要請となってきた．そのため，学会等で倫理綱領が次第に作られるようになってきている．倫理綱領は，あくまでもガイドラインであり，法律のように厳密に適用条件が定義されるものではないが，技術開発のあるべき方向性を指し示す導きの糸として作られるようになってきた．倫理綱領は，自分たちが重視する価値を明文化することで，開発の場面での手引きとなり職業集団としての責任の自覚を促す機能をもっている．

1847 年にアメリカ医師会が倫理綱領を作り，それを参考にして AIEE(American Institute of Electrical Engineers，現在の IEEE(Institute of Electrical and Electronics Engineers)が 1912 年に倫理綱領を作っている[1]．その後，機械技術者や土木技術者などの集団が倫理綱領を作っていった．日本では，1938 年に土木学会が倫理綱領を策定した．戦争を挟み，1961 年には日本技術士会が倫理綱領を作っている．けれども，各種学会に広がることはなかった．それが変わるのは 1990 年代になってからであり，1996 年の情報処理学会の倫理綱領策定が契機である．情報処理学会が急いで作った理由として，世界の情報技術関係の学会で当時，倫理綱領をもっていなかった国が日本と韓国だけだったことが挙げられる．それ以降，1998 年には電子情報通信学会が倫理綱領を作り，それ以外にも技術系の学会が次々と倫理綱領を作成して普及することになった．

人工知能(artificial intelligence，以下 AI)に関しても同様であり，AI の社会的な影響がかなり深いとこ

［招待論文］2019 年 1 月 26 日受付, 2019 年 2 月 11 日受理

ろまで広範囲に予想されるため，AI のガイドライン・報告書・提言が多く作られていった．たとえば，AI ネットワーク社会推進会議の「AI 開発原則」(案)および「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 Global Initiative の“Ethically Aligned Design”, FLI の“Asilomar AI Principles”, The White House の“Preparing for the Future of Artificial Intelligence”，House of Commons Science and Technology Committee の“Robotics and artificial intelligence”，European Parliament の“Report with Recommendations to the Commission on Civil Law Rules on Robotics”，Stanford University AI100 の“Artificial Intelligence and Life in 2030”などである．そこでは，AI の時代においても人間の尊厳を守ること，そして人間と機械との協調の大切さが述べられている．

1.2. 関連研究および研究の目的

AI ネットワーク社会推進会議の「報告書 2017」「報告書 2018」は，「人間中心」(human-centered)の社会像が幾度も唱えられている[2][3]．また，内閣府のもとに「人間中心」(human-centric)という語が入った会議「人間中心の AI 社会原則検討会議」が設けられている．そうしたところで提唱されている意義を明確化するためにも，人間と機械との違いを踏まえなければならない．そうでないと，人間と機械は同じであるのにもかかわらず，人間を中心とするというのはきわめて意味が不明瞭で浅薄なものでしかないことになってしまう．

本特集「AI 時代における人間中心の情報システム」についても同様である．人間中心の情報システムを考えるにあたって前提とすべきは，人間と機械との相違点である．そうでなくては，人間と情報システムが等式で結ばれてしまい，「人間中心」という意味が事実上，無効化する．人間と情報システムが同質であれば，「人間中心の情報システム」は「情報システム中心の人間」と読み替えても同義になるからである．

あらためていうまでもなく，これまで AI の倫理綱領等に関する研究は多く行われてきた．たとえば，江間有沙・長倉克枝[4]は，The IEEE Global Initiative の“Ethically Aligned Design”に関するワークショップを数多く開き，その論点をまとめている．また，福住伸一ら[5]は，インタビュー調査を実施して AI ネットワーク社会推進会議の AI 開発原則(案)を検討し，サービス提供やユーザ側の視点の重要性を指摘している．上村恵子ら[6]は，日本・アメリカ・欧州の AI のガイドラインを比較検討し，それぞれの文化的・社会的・宗教的背景が反映されていると述べた．上村恵子らが指摘する通り，欧州のガイドラインは AI を道具・製造物として明確に定位していることが特徴である．けれども，実際に言及されている欧州の報告書をみると，どのような点で人間と機械との間に違いが見られるかは言明されていない．すなわち，人間と機械との同質性／異質性をもとに AI の倫理綱領等を検討した研究は見当たらない．

したがって本論文では，まず人間と機械との間に乗り越えがたい差異がいまだに存在しているか否かを検討する．その後，AI にかかわる倫理綱領等のうち AI ネットワーク社会推進会議が提出した利活用原則(案)にある公平性の原則を検討する．AI 開発原則(案)は，福田雅樹ら[7]の著作により解説がなされ広く深く展開されているが，利活用原則(案)は 2018 年に出されたばかりであり提出後の考察はまだなされていない．そのため，本論文では利活用原則(案)の公平性の原則を取り上げていく．

なお人間中心主義(anthropocentrism)は，長らく動物倫理や環境倫理において批判されてきた．構造主義やポストモダン思想でも攻撃の対象となり，Michel Foucault により「人間の死」とまでいわれた．自然界の中心としての人間はもはやいない．情報システム学会では，人間中心の英語表記を human centric や human-centered ではなく，human oriented としている．日本語に直訳すると「人間志向の」「人間に配慮された」といった意味であり，バランスの取れた表現であると考えられる．

2. 人間と機械との同質性

周知のように，コンピュータ科学が成立していくに伴って「物質・エネルギー」ではなく「情報」に着目する見方が登場している．「情報」の思想は，人間と機械との異質性よりも同質性を目指していた．すなわち, Warren McCulloch & Walter Pitts の人工ニューロンや Claude Shannon の通信理論, Norbert Wiener のサイバネティクスが 1940 年代に登場して以降，徐々に人間と機械を同一線上に位置づける考えが広がりをもって受け入れられてきている．というのも，その過程で「情報」という概念が指している内容が多様なまま，同じ「情報」という語でまとめられていったからである．つまり，Shannon のように工学的応用のために確率論的な定義を行い，意味を捨象した概念から，生物の生存に本質的にかかわる価値といったものまで区別されずに同じ「情報」という概念にまとめられていった[8]．それゆえ，機械も生物も同じ情報変換体であると位置づけられるようになったのである．人間も機械も，「情報」を取り入れて内部で処理し，外に処理結果の「情報」を返す同類の情報変換体として定位された．

「エネルギー・物質」とは違う「情報」に着目した場合，表 1 にあるように論理／ホメオスタシス(恒常性)／自己複製／学習／ニューロンの働きには違いが見られない．これらの点に着目した場合は，人間と機械は同一線上のものとして定位できる．

表 1　人間と機械との同質性／異質性

	人間	機械	機械の例
論理	○	○	コンピュータ
ホメオスタシス(恒常性)	○	○	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	○	×	

コンピュータの理論モデルである万能チューリングマシンやノイマン型コンピュータも，人間のあらゆる論理的思考は 0／1 のパターン変換で扱えるという発想に基づいている[9]．それゆえ，論理的推論でみると人間と機械は同じである．

フィードバック機構に基づくホメオスタシス(恒常性)は，人間も有しているが，機械にも備え付けられている．人間は，血流や筋肉の働きにより熱を外に出す量を調整して体温を一定に保とうとする．気温が 0 度であっても 40 度であっても体温がおよそ 36 度になるように調整されている．周囲の気温変化に対応して体内で調整が図られる仕組みである．フィードバック機構は，もちろん機械にもあり，典型的なものが冷蔵庫やエアコンである．

こうした発想は，AI という語を生み出した John McCarthy の言葉にも現れている．McCarthy は，1979 年にこれまで作られてきた機械は信念に関する信念をもつには至っていないが，「サーモスタットほどに単純な機械でも信念があるといえる」[10]と述べている．いうまでもなく，サーモスタットは温度調節を行う機械で，温度に応じて金属が湾曲し，それによって弁の開閉を行う．目標値の温度と現在の温度との差を縮めるように動作するフィードバック機構である．フィードバック機能の有無でいえば，人間も機械も有であり，その点で両者は差がない．目標値と現状値との差が「情報」であり，その「情報」に基づいて両者とも動作する．

自己複製に着目する意見もある．人間の細胞は，核分裂の際に DNA を複製する．人間は遺伝情報を複製して細胞を分裂させている．こうした複製は，コンピュータでも同じであり，ミラーリングやバックアップの際にデジタル情報を複製する．この点でも，人間とコンピュータは変わりがない．

学習についても，教師が言ったことを覚え正解を出すことが学習なら，AI の機械学習はまさに学習である．たとえば AI の教師あり学習は，正解である教師データを入力し，その特徴量を抽出させる仕組みである．「犬」というラベルが付いたデータを大量に入力し，その特徴量を学習させることで，新たな画像も犬かどうかを識別できる．こうした手法は，迷惑メール・フィルターでも使われている．したがって，学習が人間の特徴であるとすると，迷惑メール・フィルターでさえ人間に含まれてしまう．

ニューロンの働きについても, 1943 年に McCulloch & Pitts によって人工ニューロンが定式化されて以降，コンピュータで模倣できるようになった[11]．各種の入力は，重み付けされて計算され 0／1 の値を出力する．この人工ニューロンが次第に多段になり，今日の深層学習(deep learning)につながっている．

このように 1940 年代以降の科学的知見および技術的開発により，人間と機械との距離は近づいてきたといえる．

3. 人間と機械との異質性

けれども, 人間と機械との異質性を主張した理論体系もあった. それが Humberto Maturana や Francisco Varela によって定立されたオートポイエーシス論である[12].

オートポイエーシスとは，自己で自分(auto)を作る(poiesis)ことであり，そうした働きを内部で行う単位体をオートポイエティック・システムという．オートポイエティック・システムは，生物の必要かつ十分な条件である．たとえば細胞は，自分を構成する様々な物質を連鎖的に作り出し，その物質の産出

過程のなかでそれ自体を産出している．

一方，一般的な機械をアロポイエティック・システムという．アロポイエティック・システムは，他のものによって作られ他のものを作り出すシステムである．たとえばコピー機は，人間によって作られ，故障しても人間によって修理される．そして生み出すのは，複写された紙である．コピー機がコピー機を生み出すわけではない．

第 3 次ブームの AI を牽引しているのは深層学習である．深層学習は，CNN(Convolutional Neural Network)や RNN(Recurrent Neural Network)，Auto Encoder，GAN(Generative Adversarial Network)などに分かれるが，いずれもアロポイエティック・システムであるといえる．いかなる領域に，いかなる目的のために AI を導入するのか，教師あり学習や教師なし学習，強化学習といった機械学習のどのモデルを使うのか，単語や文章の特徴量を抽出するためにどのように記号の類似度を計算するのか，データ量が少ないときに過学習をいかに防ぐか，どのようにノイズや欠損値が少ないデータを用意するか，どれほど分類精度を高めれば実用に耐えられると判断するか，高速の計算機資源をオンプレミスで準備するかそれともクラウドコンピューティングで準備するか，そうした判断と膨大な作業は人間が行っている．第 3 次ブームの AI も，アロポイエティック・システムである．

いうまでもなく人間は生物の一種である．したがって，オートポイエーシス論に基づけば，人間と機械は厳然たる差が見いだせる．この理論で考えなければ，もはや人間と機械との線引きは難しく，「人間中心の情報システム」は「情報システム中心の人間」と読み替えても差し支えなくなってしまう．それだけ人間と機械は近づいている．けれども，オートポイエーシス論に依拠して考えると，機械は生物たる条件を満たしていない．表 1 にある通り，オートポイエティック・システムとしての機械はいまだ存在していない．

4. AI の倫理の基盤

この人間と機械との区分は，AI の倫理を考えるうえでの基盤をなす[13]．もし人間と機械との違いがなければ，機械自体に権利をもたらす方向で議論が展開できるのと同様，逆に人間を機械扱いする立論にも容易に結びついてしまうからである．人間が機械であれば，冷蔵庫のように 365 日 24 時間働かせても構わない．機械とは，用立てるものであり，なにかに役に立つために作られる．そのため，故障して目的の機能を果たせなくなれば廃棄されても仕方がない．

けれども，人間はそうではない．Martin Heidegger がいうように，そもそもなにかに役立てるために作られているわけではない[14]．人間は，細胞が次々と分化して生成し，その内部を常に作り続けているオートポエティック・システムの集合体であり，唯一無二の存在である．みずから内部を存立させ外部との境界を作り出すがゆえ「主観」なるものが生成する．また，オートポエティック・システムは，内的メカニズムに沿って環境を認知するが，同じ時空間を占める他のシステムがない以上，個別に環境を生み出す．すなわち，オートポエティック・システムの内的メカニズムも唯一無二であり，それに伴いシステムが接する環境も唯一無二となる．他者との厳密な交換はきかない．したがって，人間を「役立つ／役立たない」の尺度だけで見るべきではないし，たとえ役立たなくとも社会から排除すべきではない[注 1].

人間と機械との同質性を主張する声が 1940 年代以降徐々に強まっているが，軽はずみに同質性ばかりを主張するのは実情に合わず，また倫理的に大きな問題を引き起こす可能性がある．

5. 公平の原則

ここでは，AI ネットワーク社会推進会議の AI 利活用原則(案)における「公平性の原則」について検討する．AI ネットワーク社会推進会議は，AI 開発原則(案)および AI 利活用原則(案)を公開している[2][3][注 2]．両原則(案)とも，人間と AI との違いを基盤とした人間中心の AI ネットワーク社会を実現するために作られているといってよい．AI 開発原則(案)は，図 1 の通り 9 原則からなる．AI 利活用原則(案)は，図 2 にあるように 10 原則である．

公平性(fairness)の原則は，AI 利活用原則(案)の 8 点目にあり，AI の利用にあたって「個人が不当に差別されないよう配慮する」ことを指している．主な論点として，「AI の学習等に用いられるデータの代表性への留意」「アルゴリズムによる不当な差別への留意」「人間の判断の介在」が挙げられている．いずれも AI の判定をそのまま真正なものとして受け取らずに批判的に検討を求める論点である．

このなかで注目に値するのは，「人間の判断の介在」という AI による自動化に抗する論点が入っていることである．この「人間の判断の介在」は，GDPR(EU 一般データ保護規則)の 22 条 1 項の「自動処理にのみ基づく決定に従わない権利」と軌を一にするが，AI 利活用原則(案)では 1 点目の「適正利用の

原則」の論点にも入っており，複数の論点のなかでも重視されている項目であるといえる．

（主にAIネットワーク化の健全な進展及びAIシステム
便益の増進に関する原則）
① 連携の原則
（主にAIシステムのリスクの抑制に関する原則）
② 透明性の原則
③ 制御可能性の原則
④ 安全の原則
⑤ セキュリティの原則
⑥ プライバシーの原則
⑦ 倫理の原則
（主に利用者等の受容性の向上に関する原則）
⑧ 利用者支援の原則
⑨ アカウンタビリティの原則

図 1　AI 開発原則(案)

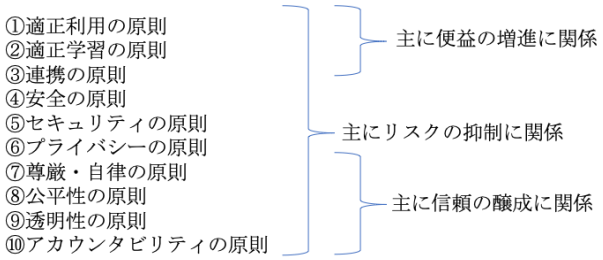

図 2　AI 利活用原則(案)

再三指摘されているように AI は，しばしば人間が組み込まれていることで非効率になっている業務に導入して効率性を上げるために有用である．たとえば異常品検知は，1 日に数万個作られるうち 1 個不良品が出るか出ないかという精度であっても，きわめて精度を厳格にしなければならないものであると，人間の熟練者による膨大な目視が要されてきた．きわめて負荷の高い作業である．そこに AI を導入することによって，人間の目視による検査を工程から外し，自動化して効率性を上げようとしている．しかしこの「人間の判断の介在」という論点では，AI ネットワーク社会が人間中心になるためには，「利用する技術の特性及び用途に照らして，どのような場合に，どの程度，人間の判断を介在させることを期待することが適当か」[3](p.63)を議論すべきであると述べられている．あえて自動化せずに人間の判断を組み込むべき領域はあるのだろうか．あるとするなら，その理由はなにか．

たとえば，人事や入試，融資といった人生を左右する意思決定を AI に全面的に任せてもよいのだろうか．人事や入試，融資などで AI による意思決定の支援はすでに行われており，今後，その数は増えていくと予想される．意思決定の支援にとどまらず，AI によって完全自動化し人間の介在をなくしてしまえば，大量かつ迅速に処理できるため，人は他のことに力を注ぐことができる．大きな利点である．

けれども AI に完全に意思決定を任せた場合，AI への責任転嫁が起きる恐れがある．サイバネティクスの父であり情報倫理の鼻祖としても知られる Wiener は次のように述べている．

> 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[15](p.58-59)

AI は，「AI の学習等に用いられるデータの代表性への留意」「アルゴリズムによる不当な差別への留意」が「公平性の原則」の論点になっていることから分かるように，必ずしも真正な値を示すわけではない．第 3 次ブームを支える AI は，ビッグデータ型であり，以前よりも扱うデータ量は格段に増えているけれども，それでもデータのカテゴリを限定して計算処理している．というのも，あらゆるデータを機械学習に入れると，計算量は莫大に膨らみ，フレーム問題が起きてしまうからである．また，いかにセンサーをはりめぐらせようとしても，大量にデータを集めても，どの角度からどのレベルで計測するかという問題が必ず生じ，世界そのものには到達しえないからである．

上記の論点に加えて，AI はうまく人の策略に使われてしまう可能性が否定できない．AI の計算結果の偏りがあることに気づいても，それが自分にとって都合がいいからと，意図的に気づかないふりをして重要な意思決定を AI のせいにする人も出てくると思われる．

さらにいうと，AI 技術者が好きなように AI の処理結果を操作できてしまう恐れすらある．AI の担当者が読み込む変数を変え，機械学習のメカニズムも調整し，自分にとって都合のよい結果が出るまで繰り返す．その後，「AI がこういっているのだから」と判定結果を出す．そうなると，その判定結果を見せられた相手は容易には反論ができない．たとえ説明を求めたとしても，相関関係の数値だけを見せられてしまい，異議を唱えることが難しい．

AI の判定結果は客観的で科学的な信憑性を帯びるが，その実は人間が操作可能であり，意識的せよ無意識的にせよ，いろいろな思惑が入り込んでいる．繰り返し述べるように AI は，あくまでアロポイエティック・システムであり，人間が予期せぬ変数間の値を示すことがあるものの，人間が作っている．

自分の都合のよいように AI の判定を変えることが可能である．したがって，AI 自体に責任転嫁する態度は避けなければならない．

当然のことながら人間が介在しても，人間がコンピュータよりもきちんとした判断ができるとは必ずしもいえない. けれども人間が必ず介在することで, AI に責任転嫁するような誤った態度は避けられる．その決定により人生の進路が変わってしまいかねない人をきちんと人間がみて判断し，たとえ責任の一端であっても人間が引き受ける．オートポイエティック・システムの集合体たる個人は，かけがえのない存在である．同等の唯一性を備えた人間が覚悟をもって対象者の人生を左右する決定を行う．人間のほうが AI よりも優れているか否かではなく，人間が必ず介在することによって責任転嫁を防ぐ．それが無責任社会を導かない措置として重要であると考えられる．

実際には，AI の処理結果と違った判断を行うことは困難を伴う．というのも，AI の結果と同じ判断を最終的に行うのであれば，その帰結が芳しくなくとも「AI の判定と同じだったのだから」と釈明できる．しかし AI の判定とは異なった判断をすると，その意思決定の責任はその人に降り掛かってくる．法的には注意義務違反として処罰される恐れも出てきかねない．

しかし，たとえ手間がかかり効力が未知数であっても，人生を左右する意思決定については人間の介在がなければ，前述の通り自分で責任をとらずに AI のせいにする無責任社会を招きかねない．

それゆえ人生を左右する意思決定にあたっては，人間の介在を必須の要請とすることで人間の関与を明示化し，責任の一端を担う仕組みを整えることが望まれる．そうでなければ，人間の介在が不可視化され，AI が完全自動で真正な値を示しているといった虚構が形成されてしまう．上司や裁判官などを AI に全面的に置き換えてしまえば, 公明正大で全体最適化された判断が行われるといった妄言が生じてしまう．AI 神といった誇大妄想が生まれてしまう．

公平性の原則の論点には現時点で入っていないが，もう一点議論に値するテーマがある．それは，スコア社会あるいはロボット・AI による技術的失業を想定し，社会への包摂(inclusion)を含めた原則にしていくことである．たしかに公平性は，辞書的な意味では偏りがなく中正であることを示しており，AI ネットワーク社会推進会議の報告書もそれに基づいて原則を整えている．けれども，これからの社会は不安定性が増し，社会的排除が一層進むことが懸念される．全人的な点数化により，点数の低い人への差別が起き，その人自身も自尊心を保てなくなり悪循環に陥ってしまいかねない．あるいは技術的失業により住居を退去せざるをえず家庭も仕事で築いた人間関係も崩壊してしまう人が出てくる．生物的特性によって差別しないことは当然だが，単なる能力主義をも超えた公平性の原則にしていくことが要請される．

能力主義的な公平性だけを重視すると，歪んだことに陥る．というのも，社会には不条理さがあるにもかかわらず，「社会的に成功している人はその人が努力したからであり，逆にうまくいかない人はその人が怠けているから」考えると，うまくいっていない人を過剰に批判することにつながってしまうからである．その人がどれだけ努力しようとしても，教育を受ける機会がなかったり，病弱で満足に仕事ができない状況だったりすることもある．そうした人までも批判してしまうことになってしまう．社会心理学では公正世界信念(belief in a just world)なるものが知られている．その尺度で捉えられる公正は，リバタリアン的な能力主義の考え方に相当するものであり，公正世界信念が強ければ強いほど社会的に恵まれない立場にいる人たちに厳しくあたってしまう．

John Rawls は，公正(fairness)としての正義を掲げ，無知のヴェールという概念装置により，格差が容認される条件は社会的にもっとも弱い立場の人の便益につながるときであるとした[16]．無知のヴェールが掛けられていると，自分がどのような状況にあるか分からず，もしかしたら最悪の状態にいるかもしれない．そのようなときに，人は最悪の状態が相対的に好ましいものを選ぶとした．Rawls のいう公正は，各人の対等な権利とともに，社会的な包摂が含まれている．もちろん AI ネットワーク社会推進会議の各報告書[2][3]には包摂という語が一語ずつ入っているものの，より明示的に強調するため，公平性の原則に社会的包摂を含めていくことが求められる. そのことによって 7 点目の「尊厳・自律の原則」とのつながりも出てくると見受けられる．

人間は誰もがオートポイエティック・システムの集合体であり，社会のなかで言語的行動をとり共に道徳的共同体を形成している．他者のかけがえのなさを尊重し，誰もが社会的排除に陥らないようにしていくことが望まれる．

6. 結語

本論文は，人間と機械との同質性／異質性を整理したうえで AI の倫理綱領等を考察した．ネオ・サイバネティクスの理論に基づき人間と機械との間に違いが明確にあることを示し，「人間中心」社会の

基本的な心がけを確認した．その後，AI ネットワーク社会推進会議が提案した利活用原則(案)の公平性の原則を検討した．公平性の原則の論点「人間の判断の介在」の意義は，人間の人生を左右する判断にかんしては，機械の異常品を検知するのとは違い，機械に責任転嫁せず，人間の責任で行うために求められるということである．唯一性を備えたオートポイエティック・システムの集合体たる人間にかかわる重要な意思決定については，同じくオートポイエティック・システムの集合体である人間が覚悟をもって行うべきものである．また公平性には，社会的な不安定性が増している状況下で，人間の社会的排除を防ぐ包摂の考えが含まれることが望ましいことも述べた．

最後に本論文に残された課題について述べる．本論文では，AI ネットワーク社会推進会議の利活用原則(案)のうち公平性の原則に話題を絞ったが，その他の原則には触れていない．さらには，周知の通り AI にかかわる倫理綱領等が次々と作られており，それらについての検討も行っていない．今後は，そうした倫理綱領等にも議論の射程を広げていく必要があると思われる．また，人間中心の AI 社会とするためには，倫理綱領を整えるだけでなく多くの技術的・社会的課題が山積している．AI システムのありかたは人間社会のありかたに密接に関係している．現場の情報技術者に加えて，専門外の人たちとの対話の場を広く用意して相互に討論し，AI ネットワーク社会のありかたを共に考えていく必要がある．

謝　　辞

本論文は，科学研究費補助金若手研究(B)「人工知能・ロボット・サイボーグの倫理的問題に関する理論的かつ実証的研究」(平成 29 年度−平成 31 年度，代表: 河島茂生，研究課題番号: 17K12800) の助成を受けた研究に基づいたものである．

（注）

[注 1]基礎情報学は，視点移動の操作により，人間がかけがえのない存在でありながら，社会のしきたりに沿った行為を求められていることを整合的に理論化した[8].

[注 2]本節の内容は，会議の公式見解・発表ではなく，一構成員である筆者の個人の見解である．

参考文献

[1] 札野順, “新しい時代の技術者倫理,” 放送大学教育振興会, 2015.

[2] AI ネットワーク社会推進会議, “報告書 2017,”. http://www.soumu.go.jp/main_content/000499624.pdf, 2019.1.1 参照.

[3] AI ネットワーク社会推進会議, “報告書 2018,”. http://www.soumu.go.jp/main_content/000564147.pdf, 2019.1.1 参照.

[4] 江間有沙, 長倉克枝, “「倫理的に調和した設計」の論点整理,” 情報法制研究, No.4, 2018, pp.3-14.

[5] 福住伸一, 神野真理子, 稲垣香澄, 安浩子, 広明敏彦, 前田春香, 水上拓哉, 佐倉統, “AI を活用したサービスにおける ELSI 的観点の新たなガイドライン項目の抽出,” https://confit.atlas.jp/guide/event-img/jsai2018/3H1-OS-25a-04/public/pdf?type=in, 2019.1.1 参照.

[6] 上村恵子, 小里明男, 志賀孝広, 早川敬一郎, “日米欧の地域特性に着目した AI 倫理ガイドラインの比較,” https://confit.atlas.jp/guide/event-img/jsai2018/3H1-OS-25a-01/public/pdf?type=in, 2019.1.1 参照.

[7] 福田雅樹, 林秀弥, 成原慧編著, “AI がつなげる社会,” 2017.

[8] 西垣通, “基礎情報学,” NTT 出版, 2004.

[9] 西垣通, “デジタル・ナルシス,” 岩波書店, 1991.

[10]McCarthy, J., “Ascribing Mental Qualities to Machines,” 1979. http://cs.uns.edu.ar/~grs/InteligenciaArtificial/ascribing.pdf, 2019.1.1 参照.

[11]McCulloch, W. S., Pitts, W., “A Logical Calculus of the Ideas Immanent in Nervous Activity” The Bulletin of Mathematical Biophysics, Vol.5 No.4, 1943, pp.115-133.

[12]Maturana, H. R., Varela, F. J., “Autopoiesis and Cognition,” D. Reidel Publishing Company, 1980. (河本英夫訳, "オートポイエーシス," 国文社, 1991.)

[13]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, Vol.5 No.2, 2016, pp.53-69.

[14]Heidegger, M., “Die Technik und die Kehre,” Verlag Günther Neske, 1962. (小島威彦, Armbruster, L.訳, "技術論," 理想社, 1965.)

[15]Wiener, N, “God and Golem, inc.,” M.I.T. Press, 1964. (鎮目恭夫訳, "科学と神," みすず書房, 1965.)

[16]Rawls, J., “A Theory of Justice,” Harvard University Press, 1971. (川本隆史, 神島裕子, 福間聡訳, "正義論: 改訂版" 紀伊國屋書店, 2010.)

著者略歴

河島 茂生（かわしま しげお）

2010 年東京大学大学院学際情報学府博士後期課程修了. 2016 年青山学院女子短期大学現代教養学科准教授. 2017 年 6 月理化学研究所革新知能統合研究センター客員研究員．AI ネットワーク社会推進会議 環境整備分科会・影響評価分科会・AI ガバナンス検討会構成員，現在に至る.