特集「ネオ・サイバネティクス」・論文

AIネットワーク状況下における集合的責任：ネオ・サイバネティクスの理論に基づく電子人間批判を交えて

Collective responsibility under AI network: Criticism of electronic person based on neocybernetics theory

キーワード：
人工知能, 倫理, 情報倫理, 集合的責任, 電子人間, オートポイエーシス
keyword：
artificial intelligence, ethics, information ethics, collective responsibility, electronic person, autopoiesis

青山学院女子短期大学，理化学研究所，青山学院大学　河　島　茂　生
Aoyama Gakuin Women's Junior College, RIKEN, Aoyama Gakuin University　Shigeo KAWASHIMA

要　約

本論文は，AIやロボットが社会に普及している状況下において，また複数のAIが通信ネットワークにおいて接続していく状況下において，いかに倫理的責任の帰属を位置づけるかを検討している。ネオ・サイバネティクスの理論に依拠しつつ，EU議会における電子人間の提言への懸念を示し，AIネットワーク環境下の集合的責任ともいうべき考え方を支持した。電子人間確立の提案は，オートポイエティック・システムでないものに人格という位置を与えることであり，それは，実情に合わないのに加えて倫理的問題を引き起こしかねない。電子人間を制度的に確立しなくとも，集合的責任の制度構築により補償は可能である。近年のコンピュータ技術の動向を鑑みるに，特定の人や組織に責任を帰属できない場合が想定される。その場合は，被害者を救済し，開発者・利用者の萎縮を引き起こさないために集合的責任の導入が求められる。ただしAIネットワーク状況下における責任のありようは，集合的責任のみだけは不足である。特定の人や組織の瑕疵が明確である場合は，そこに責任を帰属させることが望まれる。これは近代以降の慣習になっており容易に変えることが難しいうえ，開発者・利用者の故意の過失もし

原稿受付：2019年１月29日
掲載決定：2019年２月25日

くは怠慢，責任感の減退を防ぐためには，また技術を改善する動機の維持のためには必要であると考えられる。

Abstract

The aim of this paper is to consider how to position the attribution of ethical responsibility in situations where AI and robots are widely used in society and in situations where cooperation exists among multiple AIs and between AIs and other systems. Based on the theory of neo-cybernetics, we express concern about the electronic persons' proposal introduced in the European Parliament and support the idea of collective responsibility under the AI network environment. The electronic persons' proposal involves according a position as a personality to what is not an autopoietic system. It does not match the actual situation and causes ethical problems. Even if we do not establish an electronic person institutionally, we can compensate for it by establishing systems of collective responsibility. In view of recent trends in computer technology, we assume that responsibility cannot be attributed to specific people or organizations. In such a case, establishing collective responsibility is required to compensate the victims and not give rise to a situation wherein people do not favor developing or using AI. However, under the AI network situation, there is lack of collective responsibility. If the defects of a specific person or organization are clear, responsibility should be attributed to the concerned person or organization. This has been the custom after the modern era and one that is difficult to change. Besides, to prevent deliberate crimes, negligence, and decline in the responsibility of developers/users and to remain motivated to improve this technology, it is necessary to impose ethical responsibility on wrongdoers, whether they are people or organizations.

1　問題の所在

1.1　研究の背景

本論文のねらいは，AI（Artificial Intelligence）やロボットが社会に普及している状況下において，また複数のAIが通信ネットワークにおいて接続していく状況下において，いかにして倫理的責任（ethical responsibility）の帰属を位置づけていくかを考察することである。

第３次ブームが始まってからおよそ５年が過ぎ，AIが社会に組み込まれてきている。ロボットの内部にも機械学習のアルゴリズムが使われていることが多い。そうしたなかで，責任のありかたの検討が継続して行われている。

EU議会では，高いレベルのロボットに電子人間（electronic persons）という法的地位を与えることが提案され話題になった。具体的には，「Report with Recommendations to the Commission on Civil Law Rules on Robotics」（Delvaux, 2017）の59のf）で提案されている次の文言が該当する。

> 長い目でみたときにロボットに特化した法的地位を策定することはありうる。その場合，少なくとも最も洗練された自律型ロボットは電子人間の地位を得て，そのロボットが成した功績や損害の責任を引き受ける。おそらく自律的決定を行ったり，みずからの判断で第三者と相互作用したりするようなケースでは電子人格が適用される。

ロボットを電子人間とみなすことで，ロボット自体に責任の帰属を行えるようにする措置である。この文言は話題を呼び，AIやロボットの専門家や企業家，法律や医療，倫理の専門家たちが反対を唱える公開書簡を作り，署名を集めている。電子人間の是非は，これからの社会の倫理的了解を大きく変えるトピックである。

また，AI間のネットワーク化が進むことで倫理的責任の帰属先が不透明になってしまいかねないことも危惧される。AIの内部は，人々が期待する性能を上げるべく，CNN（Convolutional Neural Network）　やRNN（Recurrent Neural Network），Auto Encoderといった複数の機械学習の手法やIf-Thenの条件文が多数組み合わされ構築されている。さらに，それらのAIが相互作用してデータを交換させて群として機能することで多様なサービスが実現される。AI間が連携することで，たとえば以前万引きをした人が店舗内に入ってきたことを監視カメラの映像からAIの画像認識技術が検知し，その信号が警備ロボットに伝わり，警備ロボットが該当者を追跡するといったことが可能となる。あるいは，鉄道運行状況を監視しているAIが運転見合わせを検知し，その路線の駅にAI搭載の自動運転車を配車させるといったことが可能となる。

けれども，AIのネットワーク化が進むことで困った事態が起きることも想定される。たとえばGAN（Generative Adversarial Network）によって本物と間違えられそうな嘘の画像・動画が作られ，それがボットによって自動的にさまざまなネット上の場所に投稿されたら，どうだろう。しかも作成者は不明なままである。そのことによって名誉を毀損されて人生が変わってしまったとき，どのように補償を求めればよいだろう。あるいは，セキュリティ対策を施したコンピュータに対してもクラッキングできるAIが大量にばらまかれたとしよう。そのAIの開発者は分からない。とある合図で一斉に攻撃をしかけ，工場や銀行等のシステムを相次いで麻痺させる。業務が妨害され多額の損失が生まれた場合，その補償はどのような存在が担えばよいだろう。マルウェアの攻撃によりお掃除ロボットが幼児めがけて突進したり介護ロボットのパワーの強弱が変わり身体が傷つけられたりすることも推察される。AIネットワー

クにおける状況下においては，このような問題に対処することが求められる。

本論文の目的は，上記のような重要課題に関して，ネオ・サイバネティクスの理論に依拠しながら電子人間の検討ならびにAIネットワーク下における集合的責任（collective responsibility）ともいうべき考え方を考察することである。この両主題は互いに関連しており，本論文は，後述するように電子人間の導入を批判し，集合的責任の導入の意義を述べていく。

なお本論文は，法的責任（liability）ではなく，あくまで倫理的責任に議論を限定する。法的責任として扱われることが多い議題も倫理的観点から論じる。周知のように責任（responsibility）という語は多義語であるが，本論文では過去の行為の賞罰にかかわる責任という意味でも，現時点ならびに未来の行為の義務や責務にかかわる責任という意味でも用いる。また本論文でいうAIは，機械学習を中心としたソフトウェアに加え，AIを構成要素とするコンピュータ・システム（たとえばロボット）も含む。EU議会での提案についての言述は，その用語法に則りロボットと表記することもある。

1.2　関連研究および研究の目的

Perrow（1984）は，定常事故（normal accident）という語を生み出し，被害が甚大だったスリーマイル島原発事故やボパール化学工場事故などを取り上げながら，複雑かつ大規模な科学技術に依存している現代社会は事故が避けられないと指摘する。複雑なテクノロジーが緊密に結びつけられているため，それぞれ単独では起こりえない動きが生じる。安全装置や管理者も対応できず，予期せぬ大事故につながる。このPerrowの指摘は，AIネットワーク環境下においても成り立つ。先に触れた通りAIネットワークは，複雑なAI同士が通信ネットワークで連動して群となって動くからである。けれどもPerrowの研究は，複雑かつ大規模な科学技術の倫理的責任にまで及んでいない。

科学技術の個人的責任を超えた集合的な共同責任の提唱は本論文がはじめてではない。たとえばSchomberg（2009）は，不確実性が増す技術が社会に組み込まれる状況を作り出しているのは，個人の意図というよりも集合的行為であるため集合的責任を考えなければならないとした。そのうえで，集合的責任を構築するためには少なくとも公開討論やテクノロジー・アセスメント，法・政策上の基礎づけ，知識の質と予見可能性の向上が必要であると述べている。とはいえ，Schombergの論文はAIもしくはAIネットワークについては射程外である。本論文は，ネオ・サイバネティクスの理論に依拠することにより基礎的なところからAIネットワーク化の集合的責任を捉える。

AIと責任の問題については数多くの研究がある。なかでも赤坂（2018）は，不法行為法との関連からAIに法的人格を与えた場合の損害補填機能や抑止機能，制裁機能を検討しており，一定の妥当性は見出せるものの，制裁機能やAI自体の故意・過失の認定に疑問があるとしている。必ずしもAIに法的人格を付与する必要性はなく，本論文でも集合的責任の一形態として上げる無過失補償制度について支持している。とはいえ赤坂の研究は，あくまで法学的観点から論じたものであり，本論文とは理路が異なる。本論文は，人間と機械との違いならびにそれに付随する倫理的含意に基づき，電子人間なる概念に批判を述べ，代わりに集合的責任を論じていく。また赤坂は，AI・ロボットを法的人格として扱い財産権をもたせると法的責任が明確になる点を評価しているが，後述するようにそれはAI・ロボットへの責任転嫁につながると思われる。

また大屋（2017）は，加傷性と個別性の点でロボットやAIが責任を負いうる主体となるのは困難だと述べている。人間は傷つき死ぬ存在であるがゆえ処罰が効力を発揮し責任を担う。また個

体であるがゆえ責任の単位を同定しやすい。けれども，ロボットやAIはそうではない。この示唆は，理路は違うけれども，本論文においてネオ・サイバネティクスの理論に依拠しながら人間と機械との差異を論じることで一層明確になる。加えて本論文では，この先行研究で検討されていない集合的責任を扱う。

ネオ・サイバネティクスに基づいてAI社会の倫理を検討した先行研究としては，「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」（河島，2016）が挙げられる。この論文は，ネオ・サイバネティクスの理論をもとに生物と機械との違いを述べ，AIは自動化の範囲が広がっているにせよあくまで人間が作った機械であり，AI自体に倫理的責任を負わせることは困難であると結論づけている。

また，「ビッグデータ型人工知能時代における情報倫理」（河島，2018）は，ネオ・サイバネティクスの理論のなかでも基礎情報学の概念装置を使いながら，機械学習を中心とするAIが普及した社会における個人的次元と社会的次元の各領域のありようならびに相互に交差する領域のありようについて論じている。個人の心の領域は，AIが直接入り込んでおらず唯一性が保持されており，それが社会的次元の倫理的基盤である。社会的次元は，個人的次元を拘束するゆえそこでの倫理性を担保することはきわめて重要であり，差別の生成・助長に注意しなければならない。さらに「AI社会における「人間中心」なるものの位置づけ」（河島，2019）では，人間と機械との同質性／異質性を整理したうえでAIの倫理綱領の基盤となる方向性を検討している。

けれどもネオ・サイバネティックスの理論を用いた上記３点の研究は，EU議会に出された電子人間を考察しておらず，また特定の人や組織に責任を帰属できない事態についても考察していない。本論文は，こうした点で先行研究と区別される。

1.3　本論文の構成

本論文は，第２章でネオ・サイバネティクスの概念装置について述べ，第３章でEU議会において提案されている電子人間について考察する。また，第４章でAIがネットワーク化し連動して動く状況下における集合的責任について論じる。最後に本論文を要約し，残された課題を述べる。

2　オートポイエティック・システムおよびそれに関連した倫理・責任

ネオ・サイバネティクスの理論的支柱であるオートポイエーシス理論は，生物と機械との区分を明確にした（Maturana & Varela, 1980=1991）。生物の特徴は，自分で自分（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，生物の必要十分条件である。たとえば細胞は，自分を構成するさまざまな物質を継続的に作り出し，その物質の産出過程のなかでそれ自体の境界を含めて自己産出していく。人間の体内でも細胞が次々と分化して変形・破壊しながら常に自分を作り続けている。いわば人間は，37兆個もの細胞を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，それぞれの集合体（つまり，個人）は唯一無二の存在になっていく。そして，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ほかのシステムがない以上，また内部メカニズムも唯一無二である以上，個別に環境を認知する。すなわち，システムが接する環境も唯一無二となる。ほかのシステムとの厳密な交換はきかない。

オートポイエティック・システムの反対概念であるアロポイエティック・システムは，それ自体によって作られるのではなく，別のもの（人間）

によって作られ，別のもの（allo）を生み出す（poiesis）。人間がエアコンや自動車の部品を精巧に作り，それらの部品を組み合わせる。摩耗した部品は人間が取り替える。エアコンがエアコン自体を作っているわけではなく，自動車も自動車自体を作っていない。自動車は，シート，タイヤ，ハンドル，ガラス，ヘッドランプ，メーターなど３万点ほどの部品からできている。それらは，人間が作り組み立てている。後述するようにAIもアロポイエティック・システムの一種である。

こうした点を考慮すると，機械よりも，我々を含む生物に倫理的配慮を行うことが優先される。言い換えれば，人間だけでなく広く生物全般が道徳的被行為者（moral patient），すなわち道徳的配慮を受けるべき対象である。というのも，AIやロボットなどの機械とは違い，生物はオートポイエーシスの帰結としての自律性を有し唯一無二であるからである(1)。この考え方は，少なくない相違点はあるが，動物倫理の方向性と同一線上にあり，さらにその範囲を拡張するものであるといえる。

一方，道徳的行為者（moral agent），つまりその行為が道徳的観点から評価される者の範囲はどうだろうか。道徳的行為者の範囲をやみくもに拡大しては，犬や猫，子ども，機械にまで道徳的責任を負わせることになってしまう。機械は，人間が作り操作し維持管理するアロポイエティック・システムであり，そのような非自律的存在は道徳的行為者になりえない。

よく知られているように近代社会になって，人間は自由意志（free will）をもち自己決定するがゆえに，その選択から生じた結果を引き受けることを原理としてきた。オートポイエーシス論では人間の心を心的システムと呼ぶ。心的システムは，オートポイエティック・システムの一種であり，心に浮かび上がってくる思考を間断なく連鎖させていくシステムである。思考には，疎外感や孤独感，愛，生き甲斐などが含まれるが，そのなかで自分の考えで判断する自由意志なるものが特権的に位置づけられ，責任もそれに付随することになった。すなわち近代以降，心的システムの自由意志がとりわけ重視されて道徳的行為者となり，自由意志をもつ個人に責任が帰属されていった。前近代は，人間だけでなく動物や植物，さらには無機物に至るまで責任が帰属された。動物裁判も行われた。ネオ・サイバネティクスの理論でいうと，無機物はさておき，動物や植物はオートポイエティック・システムであり自律的な存在である。また個人の心的システムがその人の抱えている大量のオートポイエティック・システムのほんの一部分にすぎないことを踏まえると，その心的システムの作動のほんのわずかな働きだけに責任を帰属させるのは必然であるとはいえない。とはいえ近代以降，人間の自由意志への責任帰属の習慣が広まった。この習慣は，日常的な社会生活に深く広く根付いているため，今後もそう簡単には変わらないと目される。

オートポイエーシス理論によれば社会は，社会システムと呼ばれ，細胞や心的システムとは違った別種のオートポイエティック・システムである。つまり，社会は「生物的」であり，コミュニケーションが後続のコミュニケーションを喚起しながら存立する自己産出を特徴としている。個人の心的システムには還元できない社会システム内での自律性が生じる。コミュニケーションは意味的な関係で連鎖する。質問は回答を導き，商品・サービスの提供は金銭の支払いを条件づける。依頼は受託／拒否の返事を強いる。けれども，コミュニケーションは不確実性を帯びており，いかなるコミュニケーションが連鎖していくかを統制できない場合も多い。質問の意図で発した言葉が命令や嫌味に受け取られるなど，自分の思いとは離れてまったく別様に解釈されることが起きる。あるいは身に覚えがないことでも噂が広まってしまい収束できない状況も起きる。とりわけ社会システムの規模が大きくなると，コミュニケーションは非

人称的な側面が強くなっていく。特定の誰かが日本の経済や法律，学術のありかたを隅々までコントロールできるわけではない。むしろ各種のルールを作り，属人的な要素を廃していく傾向にある。

現代社会では，コミュニケーションの連鎖がコンピュータ技術に媒介されることが多く，2013年以降は特に機械学習を中心としたAIが入り込んでいる。コミュニケーションはデジタル技術に取り込まれると意味が捨象され０／１のパターンに符号化されるが，そのコンピュータの処理結果がまた人間によって観察されコミュニケーションの素材となり有意味化する。現代社会ではこのプロセスが迅速かつ膨大な量になっている。すでに社会とデジタル技術が分かち難く結びついた「“人間＝機械”複合系」（西垣，2008）となっている。

社会のなかで責任が生じる。ある問題を伴う出来事があったとき，誰の責任であるかは必ずしも自明ではない。それゆえ，責任の度合いとともに，いかなる組織や人に責任を被せるかの議論がしばしば起きる。たとえばスポーツの試合中に暴行があったとき，それが暴行をした選手の意図的行為なのか，それとも監督やコーチの命令にどうしても従わざるをえなかったのかが検討される。前者であれば選手の責任となり，後者であれば監督・コーチの責任となる。決算の書類に虚偽があったときfそれは社長の意図的行為なのか，それとも部下が社長を貶めるために裏で動いたのか，さまざまな噂が飛び交う。前者であれば社長の責任であり，後者であれば部下の責任である。すなわち，心的システムの自由意志のありようが問題になっているのである。もちろん十分に予見可能であるにもかかわらず，それにかかわる対策をとっていないのであれば非難に値する行為とみなされる。責任の所在は，さまざまなコミュニケーションを積み重ねながら定位されていく。すぐさま責任の所在が明確にならないこともあり，帰属先または分け方をめぐってしばしば争いが生じる。一度決まっても，後続のコミュニケーションで覆されることもある。当然ながら，ある人の行為が強制されたものであり避けられない状況下であった場合は，自分の意志で決定することができなかったとみなし，倫理的責任が軽減されたり，あるいは免除されたりする。

ただし社会システムの責任は，特定の人や組織に帰属できる場面ばかりではない。コミュニケーションの連鎖が急激に進み，誰も止められないことが起き，特定の誰かに帰責することが難しい場面が生ずる。たとえば株式市場で買い注文が殺到して市場全体の株価が上がったとしても，特定の個人のおかげではない。持ち株が上がった人全体がいわば分散的に責任を負い利益を受けるかたちである。新聞の発行部数が減っても，特定の個人の責任ではない。国家全体の出生率が上がらなくても特定の個人が責めに帰すわけではない。これらは，自然人である個人ではなく，社会システムの動きのゆえであり，いわば株式市場や業界，国家といった社会システムの問題なのである。先ほど触れた通りコミュニケーションは非人称的でありうる。

３　電子人間に対する批判的検討

前述したようにEU議会では電子人間としてロボットを扱う提言がなされている。この提起は，ロボットの法的責任を考えるために編み出された案であり，高機能化するロボットの責任割当の複雑さを解消するために強制加入の保険制度を確立して補償基金を設立することや，EUで登録簿を用意し，個々のロボットと基金との対応関係を可視化しておくことなどと関連して提出されている（Delvaux, 2017）。

ところが冒頭で触れたように電子人間の案は反対署名にまで至っている（Robotics-openletter.eu, 2018）。その反対の理由は，技術的な面でも倫理的・法的な面でも困難があるからである。反

対署名の文書によれば，技術的な面でいうと電子人間に関する提案が出てくるのは，最先端のロボットでさえ実装されていない機能を過大評価し，予見不能性や自己学習能力についても表層的な理解にとどまっていることによる。また，サイエンス・フィクションやいくつかのセンセーショナルな報道発表によって歪められたロボット観に基づいていることによる。法的な面に関しては電子人格は，自然人のモデルから導きだすことはできない。また，法人のモデルからも導きだすことができない。というのも，法人の背後には人間がいるが，電子人格はそうではないからである。

これに加えて，反対署名の文面にはないが，多くのロボットは通信ネットワークにつながっておりデータ収集・送信，ソフトウェア更新などを絶えず行っているため，ロボット単体に責任を帰属することにはプラグマティックな観点からも疑問である（大屋, 2017）。

ロボットに独立した法的人格を与えようとする立論は，なにも上記の提言だけではない。たとえば，ChopraとWhiteは，ヒューマニストを怒らせる恐れがあるとしつつも，人工的行為者（artificial agent）は人間と同様の自律的な意思決定を行うため，法的人格に値すると書いている（Chopra & White, 2011）。これに反対する哲学者は，人工的行為者には「何かが欠落している」と言うのだが，それは排他主義的な発想であると述べている。

その欠落している「何か」は，ネオ・サイバネティクスの理論に基づけば，オートポイエーシスである。たとえ最新のAIが搭載されていてもロボットはアロポイエティック・システムであり，反対署名のいうようにロボットを電子「人間」として扱い，そこに帰責することには異議を唱えざるをえない。

エアコンや自動車と同じくAIも，AIがAIを作っているわけではなく，アロポイエティック・システムである。AIの第３次ブームを牽引している深層学習でも同様である。従来のSVM（Support Vector Machine）などの手法に比べて自動化の範囲は増した。大量のデータから特徴量の抽出を自動で獲得し，間違った出力があれば出力に近い側から調整する誤差逆伝播法も使われているからである。けれども，AIを導入する目的や領域を決め，教師あり学習や教師なし学習，強化学習の手法の選択を行うのは人間である。また単語や文章の特徴量を抽出するための記号類似度の計算方法を考え，実用に耐えられる分類精度の値を決めるのも人間である。大量の教師データを用意するのも人間であり，CNNを使うのかRNNを使うのか，何層のニューラル・ネットワークにするのかを決めるのも人間であり，検証用のデータを用意してテストするのも人間である。GPU（Graphics Processing Unit）などのハードウェアも人間が用意している。AIがハードウェアも含めて自分で自分を作るようになるには，相当の技術的ステップが要される。オートポイエーシスの有無で考えると，第３次ブームのAIもアロポイエティック・システムであり，いまだ生物と機械は異質である（河島，2016）。AIは，ネオ・サイバネティクスの意味では自律性を備えていない。

また人間とAI・ロボットを同列に扱えば，大きな倫理的問題を引き起こしかねない。人間のようにAIを扱うという考えは，逆に人間を機械のように扱ってもよいという考えを導きかねない。人間も機械も同じであれば，同じように扱っても差し支えないからである。これは，人間が機械のように365日24時間働くことを要求されるということでもある。コンピュータであれば３年〜５年ほどしたら処理スピードが遅くなるため不要物としてしばしば捨てられるが，人間もそれと同じような扱われ方をされるということでもある。いうまでもなく，人間の尊厳を損なうことである。

加えて，ロボット自体が独立してみずから判断を下せるからそれ自体で責任を負えるとすれば，それを開発・利用した人や組織は免責される。

SharkeyやNavejansが手厳しく批判するように，ロボットへの法的人格の付与は，機械の起こした動作に対して製造者の責任を消去する（Delcker, 2018）。そうなれば，製造者の瑕疵が明確になるケースであっても，その責任を問えない事態に陥る。ロボットを組み立てたりデータを提供したりする場面で誤りがあっても，その責任を追及できない。2007年に南アフリカで訓練中に自動制御兵器が誤作動し９人が死亡，15人が傷を負った事件で，南アフリカ政府は機械の故障だとして製造者を非難した（Hosken, 2008）。そのような非難ができなくなりかねない。かつてWiener（1964=1965）が指摘したように，人間が設計・製造した機械自体に事故の責任を押しつけるという完全なる責任逃れに陥るといえる(2)。そのような責任逃れの考え方からは，今後の被害者数を減らすための方策が生み出されてこないと推察される。電子人間という提案は，実態から乖離しており倫理的問題をも惹起しかねない。

AIやロボットが起こした事故に対する補償基金が電子人間とあわせてEU議会で提言されている。しかし補償基金は，電子人間とは別個の論点であり，電子人間なるものを確立しなくとも設立できる。補償基金設立にあたり電子人間は必要条件ではない。というのも，過失が不分明な場合には集合的責任を導入すれば機能するように想定されるからである。あくまでAIやロボットを作り利用している人間社会側の責任として定位することで，不必要にAIやロボット自体に責任転嫁する事態を防ぐことが可能である。

付言しておけば，EU議会で提案されている電子人間の考え方では，AI・ロボットが結婚したり投票したりすることは想定されておらず，自然人よりも法人と似た観念が仮定されているように見受けられる。ただし法人自体は，社会組織でありオートポイエティック・システムである。すなわち，その内部のメカニズムに則り環境を認知しながら公式的な決定を連続的に実施し存立している。オートポイエーシスの有無でいえば，AI・ロボットと社会組織ではやはり差がある。

4　AIネットワーク化における集合的責任

近代になってからは個人の自律性，つまり心的システムの自由意志がその論理的根拠となり，基本的には個人に責任を帰属させてきた。各人の心的システムはみずからの意思によって判断し行為する。その帰結については責任を引き受ける。倫理的行為においても同様である。近代以降，人間の心的システムにおける自由意志に倫理の基盤を置き，そのことから瑕疵があったときに個人に責任を課し，社会の安定を図ってきた。被害を生んだのは個人の行為であり，その人を非難することで被害者の苦しみを解消することが求められた。

こうした事態は，そう簡単には変わらないと想定される。近代社会に生きる我々は，それぞれの人が自己決定する権利をもっており，多かれ少なかれ各人が自由意志によって判断を下していると考えているからである。したがって過失が個人の判断に求められる場合，個人に帰責することはこれまで通り続いていくと予想される。意図的に悪意をもって人に損害を与える人が完全にいなくなることは考えられない。金銭でサイバー攻撃を請け負う人もいる。ドローンを操作して，あるいは自動運転車を乗っ取り，他者を殺害することもありうるだろう。IoTのプログラムに設計上のミスがあれば，その責任を問われるのはメーカーであると予想される。欠陥による危険を認識していても措置を講じていないのであれば，社会的に期待されるべき義務を果たしていないと解される。必要なセキュリティ対策を講じていない人もいる。AIが普及した社会が無責任社会になってはならず，開発者・利用者の故意の過失もしくは怠慢，責任感の減退を防ぐためには，また技術を改善する動機の維持のためには，明確なるミスについてその倫理的責任を追及し続けざるをえない。特定

の人や組織の過ちまで，後述する集合的責任として補償することは人々が納得しないと考えられる。許し難い過失や故意があった場合でも，それを追及できないとするならば被害者は不満を募らし鎮静化には至らない。開発者やそのAIによってサービスを提供するものが責任を負い，役割にあった行為を遂行すると予期できるからこそ，人々はAIネットワークを信頼して受容することができる。コミュニケーションが遂行されるなかで，どのような人・組織に帰責していくかが定まっていくと推量される。複数の人や組織が関与している場合は，話し合いの末，7：3などのように事後的に責任が配分されていくだろう。あらためていうまでもなく，倫理的責任の帰属は財産的な制裁だけにかかわるわけではない。職業活動の統制に通じる懲戒にもかかわる。

重要な課題は，前記したように，複数のAIが連携して動くなかで個人の悪意や過失が同定できない場合に起こる。あるいは，どの個人が行った行為なのか特定できないケースに起きる。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。不注意や管理の不徹底だけでなく，どんなに予防策を講じても事故は起こりうる。おかしな挙動が絶対に起きないことは考えにくい。

AIは，専門性の高い分野であり，その内部機構はただでさえ複雑であり，今後もその複雑性は増すと考えられる。MicrosoftのAI「Tay」の騒動に端的に表れているように機械学習を使っているため，データが変動すると出力も変わる。個々のAIは対策を施してサービスを実施するだろうが，そうしたAI同士がネットワーク化し連動するなかで，予期しない動きが生じ未解明の事象が出てくることは避けられない。協業／分業に伴うリスクもある。aというAIシステムの開発者は，ほかの複数の会社が開発・運営しているb，c，d，・・・といったAIのデータを取り込み，独自のメカニズムで処理して結果を出すようにプログラムを実装したとしよう。このときaの開発者は，b，c，d，・・・が正しいデータで妥当なデータ処理をしていることをあてにできるからこそ，自分の作業に専念することができる。aの開発者は，b，c，d，・・・の中身やデータの適切性について検証する時間もなければ権限もない。したがって，たとえ擬制であってもb，c，d，・・・が適正な動きをしていると信じなければならない。もちろん，b，c，d，・・・のAIの不具合が噂になったり，データ・フォーマットが崩れたりしているとaの開発者は気づく機会を与えられる。しかし厳密なる検証はし難い。それゆえ誤ったデータが一度流通すると，それが連鎖していく恐れがある。これらはAI群のネットワーク化に内在するリスクである。また，ネットワーク通信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。インターネット広告では，広告主や広告代理店，配信事業者でも，どのようなウェブサイトに広告が表示されているかを正確に把握できず，思惑から外れて猥雑なウェブサイトや政治的にきわめて偏ったウェブサイトに掲載され，またアドフラウド（広告詐欺）にもあっている。AIがネットワーク化して群として機能するようになると，責任の所在の検証が難しくなることは，それほど頻繁に発生するとは考えにくいとはいえ，容易に想定される。AIの相互作用といった予期しえない動きまで開発者や利用者の義務の範囲内に入れてしまうと，過失がないにもかかわらず責任が課せられてしまう事態に陥ってしまう。

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。過度に責任を負わせようとすると，開発者側

や利用者側への圧力が高まる恐れがある。たとえ事故が重大であったからといって，そこから開発者側･利用者側の瑕疵を演繹することはできない。過度な責任追及は，開発者・利用者側への負担を増し，かえってAIシステムがもたらす利益を確保することが難しくなる。

逆に過失が同定できないからといって，被害者が救済されない事態を招くことも望まれない。特に被害者側が過失を立証しなければならない場合，AIネットワークの複雑さを鑑みると，その過失の立証はきわめて難しいといわざるをえない。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くことができない。人間は誰もが唯一無二のオートポイエティック・システムの集合体であり，かけがえのない存在である。誰もが社会的排除に陥らないようにしていかなければならない（河島, 2019）。すなわち, 開発者や利用者側への活動を阻害せず，かつ被害者を救済する制度の創設が要請される。

したがって悪事を働いたものが特定できない場合にせよ，十分予見されえず非意図的な場合にせよ，過失が同定できないならば，AIネットワークが組み込まれた社会システムそれ自体が一種の道義的責任を担い，損害を被った人に補償していく制度の構築が望まれる。これは，言い換えれば社会システムを道徳的行為者と定位する社会制度であり，社会システムの問題として受け止めるということである。AIネットワークも，社会のコミュニケーションを半ば担い，人々のコミュニケーションを機械情報に変換して機械情報を動的に連鎖させるため，現代社会の社会システムの一部を形成している。上記で述べたように，社会システムのコミュニケーションは，そもそも不確実性があり非人称的な面を抱えている。そのような特徴をもつ社会システムに組み込まれてAIネットワークは動作する。個々のAIはアロポイエティック・システムであり，入力と出力の対応関係が定められた範囲で収まるように調整されている。けれどもAIがネットワーク化し不確実性のある社会システムに組み込まれて群として動いたときに開発者・利用者の予期せぬ動きが生じ，コントロール不能に陥ることが考えられる。善意で開発しても，わざと悪用する意図がなくとも，注意していても開発者・利用者が完全に統御できるものではない。集合的責任の制度構築により，開発者・利用者の負担を過度に増加させなくとも被害者を救済する途が開かれる。開発活動・利活用の保護および被害者の保護を考慮すると，こうした制度は，過分な責任追及を行う度合いを相対的に低くし被害者救済も図れるため，バランスの取れた帰結をもたらすように推察される。過失のありかが分からないケースにおいては誰かの罪を問うというよりも，AIネットワーク化の産業活動全体，つまりAIを開発・維持しデータを収集・整理・解析しAI同士を連携させる一連の作業全体に関係していると受け止めることが求められる。またそうしたAIネットワークを基盤技術としそれと渾然一体となっている人間社会全体の問題であると受け止めることも議論の範疇に入る。

制度としては，税金や保険，業界の組合，業者からの拠出，利用者の一部負担などの財源による補償が考えられる。すなわち税金の税率を増やして社会保障を強化したり，加入を義務とする保険を設けたりして，いわゆる無過失補償制度の確立を目指すことも一手段である[(3)]。

すでに「自動運転に係る制度大綱」（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, 2018）において次のように述べられている。「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」

（p.18）。つまり，明らかな瑕疵が見出せる場合にはその人が責任を問われるが，そうでない場合には政府が補償的責任をとるということである。同様の指摘は，「自動運転における損害賠償責任に関する研究会 報告書」（国土交通省自動車局，2018）にもみられる。

集合的責任は，なにも身体への傷害に対する金銭的な補償だけにかぎられない。インターネット上ではフェイクニュースが氾濫している。AIは，フェイクニュースの検知や伝播過程の分析にも使われるが，フェイクニュースの作成や流布にも使われる。匿名化技術によって作成者や送信元が特定できないように加工され，プラットフォーム企業のサービスに流される。そのデータが人生を左右しかねない意思決定を支援するAIに取り込まれ，そのほかのAIにも伝播していく。たとえ不快な嘘，不正確で誤解をもたらす情報が出回り名誉が傷つけられて人生が狂わされても，怒りを向ける相手が分からず対処できない。謝罪広告も出してもらえない。そうしたときに，公益財団法人等の運営組織を共同で立ちあげてフェイクニュースが事実誤認であることを示し，合わせてAIのデータの書き換え要請を行うことで被害者を救済する方法もありうる方策である。

誤解を避けるために付け加えておくと，非難を向ける先が不分明である際に集合的責任が求められるのであって，個人や法人の瑕疵が特定できる場合は，これまで通り個人や法人に責任を帰属すればよい。ここでいう集合的責任とは，集合知の対といってもよく，個人や組織を超えた大きな規模（業界や国家，国際社会など）における責任を指している。また，集合的責任はアカウンタビリティではない。というのも，予期できない事故が起き，また被害者もしくはステークホルダーが納得する意味を伴った説明ができない場合に機能することを想定しているからである。

加えて集合的責任は，AIネットワーク化社会における個々人がそれぞれ罪（guilt）を背負うということではない(4)。Arendt（1987=2007）は，罪は単独の個人の行為に関連づけられるものであって，悪いことをしたという自覚が罪にあたるとした。集団に罪があるといってしまうと，「わたしたちのすべてに罪があるのだとしたら，誰にも罪はないということになってしまう」（1987=2007: 195）と述べている。あくまで集合的責任は，過誤の罪を引き受けることを指しているのではなく，補償的責任である。社会システムの便益を増進させるために導入されているAIネットワーク化に付随するリスクであり，予見可能性の低い出来事により人生に重大な悪影響を被った人を社会的に排除しないための措置である。たとえ無辜であってもAIネットワーク化社会を成立させている人々が分散的に背負わなければならない代価である。

5　結語

本論文は，ネオ・サイバネティクスという学術的礎をもとに，電子人間の提言への懸念を示し，AIネットワーク環境下の集合的責任ともいうべき考え方を支持してきた。

電子人間に対する提案は，アロポイエティック・システムであり非自律的なものに人格という位置を与えることであり，それは，実情に合わないのに加えて倫理的問題を引き起こす。

本論文は，個人的・組織的責任と集合的責任の両立を支持する。近代以降の社会は，個人の内にある心的システムの自由意志を特権的に位置づけ，そこに倫理的責任を帰属させてきた。特定の人や組織の過誤であることが同定されているにもかかわらず，社会で補償していくとなると，人々は違和感を抱き不満を覚えると想定される。また個人的・組織的責任をなくせば，開発者や利用者は悪意をもったり注意を怠ったりしてしまうことも考えられる。したがって個人的・組織的責任をできるかぎり追及していくべきである。しかし複

雑なAIを含んだコンピュータ・システムがネットワーク化し互いにデータをやりとりして動くと，どうしても特定の人や組織の過誤が判然としないことが起こりうる。このような特定の人や組織に責任を帰属できない場合，被害者を救済し，開発者・利用者の萎縮を引き起こさないために集合的責任の導入が求められる。すなわち個人的・組織的責任だけでは限界がある。個人の内面の倫理観を高め技能を向上させるだけではなく，不確実性が内在化している社会システムの責任として引き受けることが求められる。個人的・組織的責任を可能なかぎり追及し，それでも難しい場合は最後のセーフティネットとして集合的責任に沿った制度を準備せざるをえない。

最後に本論文で残された課題について述べる。本論文は，理論的な基礎研究であるゆえ，具体的な制度について考察していない。実際には，どのように特定の個人や組織に帰属できる責任と集合的責任とを区別し運営していくのか，制度の構造は多様でありうる。集合的責任を認定する仕組みをいかにするか，補償の金額は定額にするのか，それとも個別に対応していくのか。補償や運営組織の財源はどのようにするのか，どのような領域を適用範囲とするのか。あまりにも数が多くなってしまった場合，どのように対処していくのか。あらためていうまでもなく，これらは相互に連関しており総合的な検討が必要である。

また，社会システムの責任として引き受ける場合，営業秘密に抵触しない範囲でAIの技術が公開されていなければならず，透明性を最大限に図り，たとえライバル企業であっても共同で解決策を練ることが強く期待される。調査機関を設け，事故の調査を行い再発防止に努めなければならない。さらにAIネットワークの技術者や企業といった職業集団内に限ってコミュニケーションするだけでなく，それ以外の多くの人たちとの対話の場を用意し，可能なかぎり利害関係者が納得しあって進めていく必要があると目される。

謝辞

本論文は，中川裕志先生（理化学研究所）からいただいたご示唆を踏まえ執筆した。深く感謝申し上げる。また本論文は，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

（１）社会システムを観察する立ち位置からいえば，個人は他者や機械と交換可能であり，特定個人への思いやりだけを追い求めると平等性を欠く事態に陥ってしまう。ただし，個々人への配慮の次元を忘れてしまえば，社会システムの倫理性の確保の基盤は失われる。こうした議論については「ビッグデータ型人工知能時代における情報倫理」（河島, 2018）を参照。

（２）AIネットワーク社会推進会議の利活用原則（案）における公平性の原則で論点となっている「人間の判断の介在」は，個人の人生を左右する重要な意思決定に関してAI自体への責任転嫁を防ぐために求められる事項である（河島, 2019)。

（３）すでにフランスでは医療事故事例において損害賠償制度と併存するかたちで無過失補償制度が導入されており，このほかスウェーデンやニュージーランド，デンマークでも導入されている。日本でも産科医療補償制度が整備されている。

（４）Arendtは，宗教の影響によって，倫理や道徳が集団的なものから個人的なものへと変わったといい，集合的責任は政治的なものであると述べている。このように個人の次元だけに倫理をとどめてしまうのは本論文の立論との相違が見て取れる。

参考文献

赤坂亮太（2018）「不法行為法における AI の法的人格に関する検討」, 2018年度人工知能学会全国大会発表資料, <https://confit.atlas.jp/guide/event-img/jsai2018/ 1 F 2 -OS-5a-03/public/pdf?type=in>
Accessed 2019, January 12.

Arendt, H.（1987=2007）Collective Responsibility. Boston College Studies in Philosophy,（26）, pp.43-50.（中山元訳「集団責任」,『責任と判断』筑摩書房, pp.195-208.）

Chopra, S. & White, L. F.（2011）A Legal Theory for Autonomous Artificial Agents, University of Michigan Press, Michigan, 252p.

Delcker, J.（2018）Europe Divided over Robot 'Personhood',
<https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/> Accessed 2019, January 12.

Delvaux, M.（2017）Report with Recommendations to the Commission on Civil Law Rules on Robotics
<http://www.europarl.europa.eu/cmsdata/113782/juri-final-report-robotics.pdf> Accessed 2019, January 12.

Hosken, G.（2008）Army Blames Gun's Maker for Lohatla, IOL News,
<https://www.iol.co.za/news/south-africa/army-blames-guns-maker-for-lohatla-387027> Accessed 2019, January 12.

河島茂生（2016）「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（２）, pp. 53-69.

河島茂生（2018）「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.

河島茂生（2019）「AI社会における「人間中心」なるものの位置づけ」,『情報システム学会誌』14（２）, pp.21-28.

国土交通省自動車局（2018）「自動運転における損害賠償責任に関する研究会 報告書」,
<http://www.mlit.go.jp/common/001226452.pdf> Accessed 2019, January 12.

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（2018）「自動運転に係る制度大綱」, <https://www.kantei.go.jp/jp/singi/it2/kettei/pdf/20180413/auto_drive.pdf> Accessed 2019, January 12.

Maturana, H. R., Varela, F. J.,（1980=1991）Autopoiesis and Cognition, D. Reidel Publishing Company, Dordrecht, 146p.（河本英夫訳『オートポイエーシス』国文社, 320p.）

西垣通（2008）『続 基礎情報学』NTT出版, 219p.

大屋雄裕（2017）「外なる他者・内なる他者」,『論究ジュリスト』（22）, pp.48-54.

Perrow, C.（1984）Normal Accidents, Basic Books, New York, 386p.

Robotics-openletter.eu（2018）Open Letter to the European Commission on Artificial Intelligence and Robotics,
<http://www.robotics-openletter.eu/> Accessed 2019, January 12.

Schomberg, R. von（2009）Organising Collective Responsibility, Keynote lecture at the first annual meeting of the Society for the Study of Nanoscience and Emerging Technologies, Seattle, 11 September.

Wiener, N.（1964=1965）God and Golem, inc., M.I.T. Press, Cambridge, 99p.（鎮目恭夫訳『科学と神』みすず書房, 149p.）