レクチャーシリーズ「シンギュラリティと AI」〔第 5 回〕

AGI の群雄割拠

Multiple AGIs

中川　裕志　理化学研究所革新知能統合研究センター
Hiroshi Nakagawa　Center for Advanced Intelligence Project, RIKEN.
https://sites.google.com/site/nakagawa3/

Keywords: artificial general intelligence, super intelligence, AI ethics.

1. 唯一の超知能の脅威論

ボストロムがその著書「Super Intelligence」[Bostrom 14] で，人工知能*1 がひとたび多様で予測されなかったような環境や目的に適応できる汎用人工知能（Artificial General Intelligence：以下 AGI と略記する）になると，AGI の能力は指数的に増大し，全く無敵の超知能にまたたく間に達するとしている．このことは AGI の次の性質から導かれる．

性質：AGI は新たな環境や目的に適応できるように自らの構造を目的に応じて改善することができる．

AGI が未知の環境に適用できるために，高い適用能力すなわち自らの汎用化を実行する能力をもつという目的を実現した場合，その能力は指数的に増大する．つまり，X という AGI が新規の能力を獲得して構築された次の世代の AGI を $A(X)$ と記述した場合，世代を重ねると，

$$X \rightarrow A(X) \rightarrow A(A(X)) \rightarrow A(A(A(X))) \rightarrow A(A(A(A(X)))) \rightarrow \cdots \quad (1)$$

と進化するため能力が指数的に増大するわけである．指数的であるということは，以下の（a），（b）を意味する．

（a）AGI の世代進化の速度が指数的に増大する*2．

（b）AGI の多様な目的に対する適応化が，質的，量的に加速する．

したがって，この指数的に加速する進化に他の AGI は追いつけなくなり，結果として唯一の AGI が超知能として君臨し，他の AGI を寄せ付けないというのがボストロムの主張である．

さらにボストロムの主張は以下のように続く．超知能はその存在目的に合致した方向に指数的に突き進む．例えば，クリップを量産するという存在目的だけが与えられていたとすれば，その目的を最大限に実現しようとして地球の資源をすべて使ってクリップを生産し続け，結果として地球上の全資源がこの超知能に収奪される可能性があるという．この例は明らかに超知能に至る前の AGI の存在目的の規定あるいは実現方法に関する基本設計のミスである．この例はあまりに単純に見えるが，複雑な構造をもつ AGI ではこのような基本設計ミスを根絶することは困難であろう．ボストロムをはじめ，AGI や超知能の脅威を説く書籍 [バラット 15, Bostrom 14, ダベンポート 16, フォード 15, マルコフ 16] には，このような基本設計ミスが指摘されている．

現在の目的特化型 AI として有名な囲碁ソフト AlphaGo ないしその後継の AI 囲碁ソフトの出力する棋譜（打った手のシーケンス）は，プロの囲碁棋士が見ても，その手の意味がなかなか理解できないという．このように目的特化型 AI ですら，内部動作はおろか，その出力行動も人間の理解を超える複雑さの域に達している．このことから，深層学習も含む現在のソフトウェアによる AI は，さらに複雑化して AGI になったときに基本設計にミスがあるかどうかを設計段階で完全に発見しきることが不可能ではないかということが推測される．

このような実情を鑑みて，Superintelligence [Bostrom 14] の後半では，AI あるいは AGI の基本設計ミスを防げるという考え方に疑問を呈したうえで，AGI や超知能*3 に関わる包括的な設計方針を議論している．ボストロムは計算機上で動くソフトウェアによる AGI が超知能化したとき，人間の制御が全く効かず，意思疎通もできず，人間の倫理的基準も通用しないことを念頭におき，むしろ人間と同じ構造をもつ AGI のほうが超知能化したとき人間と共存しやすいのではないかと主張している．最も基本的レベルでは，脳科学の知見を利用して，脳神経レベルから人間と同じ構造の AI をつくる方法が考えられる．もう少し抽象度を上げると，脳における各モジュールに対応させて，同じ入出力構造をもつモジュールをつ

*1 以下では，必要に応じて人工知能を AI と略記する．
*2 当然，進化のために使うエネルギーも増加するであろう．
*3 以下では，超知能は最高の能力に達した唯一の AGI であり，一意的であるとする．

くり，それらを組み合わせて人間の脳と同じ機能を実現する方法もある．この方向での研究として全脳アーキテクチャ [全脳アーキテクチャ勉強会] が研究されている．

1·1 個体保存と種の保存

クリップの増産という存在目的は比喩的な例であるが，AGI 自身が生き残るという一見，謙虚な目的の場合でも資源を食い尽くす可能性は高い．もし，超知能が個体を保存することで生き残ろうとするなら，超知能は一個の個体を強じん化するだけなので，その影響範囲は限定される．しかし，1 個の個体では何らかの理由，例えば人間による電源の切断で消滅する可能性がある．それを防ぐためには，自分と同じ機能をもつ個体のコピーを大量につくり，異なる場所に散在させることが有力な方策となる．つまり，個体の保存から種の保存に目覚める可能性がある*4．人類にとっての危険性ないし脅威という視点から見た場合，AGI が種の保存に目覚めたときがシンギュラリティであるとも考えられる．

ひとたび，超知能が種の保存に目覚めると，膨大な数の超知能が再生産され，地球上の資源を食い尽くす危険性はさらに高まる．コピーがあちこちに散在し，さらに増殖の可能性もあるのだから，超知能が危険なら電源を抜けばよいという安易な制御は通用しないだろう．このような状態になってしまうと人類に生存の脈はない．

2. 複 数 の AGI

本章では，前章で述べた唯一の超知能の支配が AGI の将来に関する唯一のシナリオではないことを説明する．

2·1 単一の超知能の宇宙制覇シナリオの破綻

カーツワイル [カーツワイル 07] は知的生命体の本質は情報パターンであると述べている．人間が自分自身の脳内の情報パターンを保持すれば，身体のあちこちを入れ替えても知的生命体としては連続し，うまくいけば不老不死となることができると唱えている．さらに，この情報パターンはあくまで情報なので，地球でつくられた超知能を地球以外にも送ることができる．地球以外の天体に到着した情報パターンはその地で知的生命体を再生し，究極的には 1 種類の超知能が全宇宙を支配するというシナリオが展開されている．

しかし，情報パターンといえども宇宙を旅するには時間がかかる．光の速度で移動できたとしても 10 年から数万年のオーダで時間がかかる．このような長い時間があると，超知能はその間にもとてつもない進化を遂げ，到着地の環境によっては異なる方向への進化もあるだろう．したがって，宇宙全体が 1 種類の超知能で支配されるというストーリーの実現性は非常に低い*5．

2·2 複数 AGI のシナリオ

カーツワイルほど極端ではないにしても，ボストロムも AGI の指数的進化によって飛び抜けた能力を得た単一の超知能が現れるシナリオを展開し，その超知能が人類を支配してしまうのではないかという脅威論を述べている．

このシナリオが絶対起きないと断言できないにしても，これ以外のシナリオも十分にあり得ることを以下で述べる．

最初に以下の仮説をおく．

仮説：一つの AGI が生成したコピーは多様な方向に進化する．

すでに述べたように，種の保存に目覚めた AGI は自分のコピーを生成してあちこちに散在させるだろう．種の保存を効果的かつ永続的に実現するには，種々の異なる環境にコピーを散在させるほうが有利である*6．ここでいう環境とは，利用可能な情報資源からなる情報環境のほかに，AGI を搭載したロボットを想定するなら異なる物理的環境も意味する．さらに AGI と相対する人間達の態度，在り方も AGI にとって環境といえる．そこで AGI は自分が置かれた環境に適応するような進化を遂げるであろう．なぜなら，環境への適応は AGI の性質で述べたように基本的能力だからである．その結果として，個別の環境に適応した複数種類の AGI が割拠する状態になる．

ここで留意すべき点は，図 1 に示すような次の点である．

(a) 個別環境に適応した AGI はその環境ではおそらく最強であろうが，他の環境では最強とは限らない．

(b) 個別環境の AGI は同じ出自，すなわち同じ基本設計である AGI が他の環境におり，独自に進化していることを知っている．つまり，別の環境の AGI は自分と類似の基本的能力をもつが，自分よりうまくその環境に適応した AGI の存在を知っている．

この留意点を念頭におき，仮に AGI が別の環境の AGI と競うために別の環境を調査し，それへの適応を始

---

*4「目覚める」という言い回しははなはだ文学的だが，実態としては，AGI が目的を遂行するためには自身のコピーをつくってよい，という許可をアルゴリズムとして記載しておけばよい．このことの背景として，すでに自己修復機能をもつソフトウェアは OS などに見られるが，ソフトウェアが複雑化し，例えば機械学習で自らソフトウェアの変更を行うような場合，自己修復機能の実装が難しくなることが考えられる．この場合は，別の場所にコピーをつくっておき，修復が必要になったとき，そのコピーで置き換えるという方法はあり得る．

*5 カーツワイルは超知能が進化すれば空間のゆがみを利用したワープなどを使って光速を超える移動が可能になると書いているが，このアイディアは少なくとも現状ではサイエンスフィクションであろう．

*6 同一の環境にいくらコピーを増やしても，その環境が激変して AGI が生き残れないときは，コピー AGI が全滅してしまい，種の保存ができなくなる．

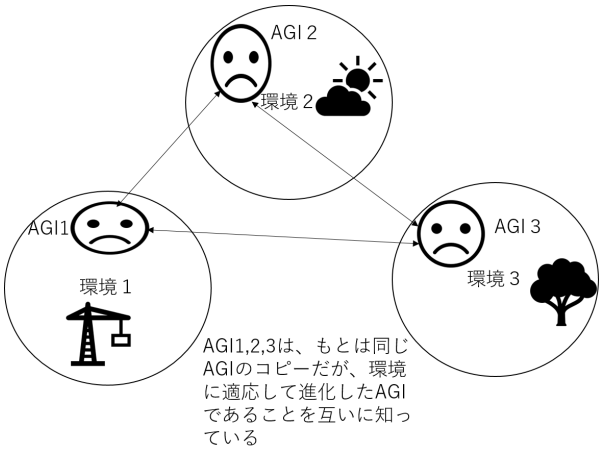

図 1　互いを知る AGI 達

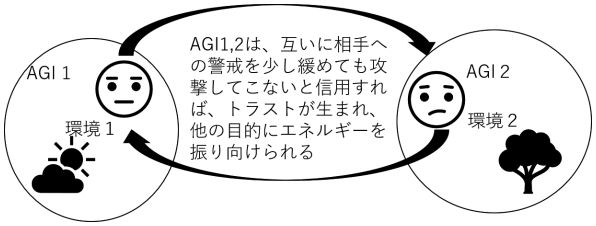

図 2　トラストする AGI 達

めたと仮定してみよう．当然，別の環境の AGI もこちらの環境への調査，適応を開始しているが，そのことをこちら側の AGI も意識している．結果として AGI は自分の環境では別の環境から来た AGI に勝てるかもしれないが，相手側の環境において勝つことは難しいことも意識している．つまり，互いに容易に手が出せないにらみ合いの膠着状態に陥る可能性が高い．

さらに，このような AGI は環境ごとに存在する．つまり，AGI は多様かつ多数存在するので一つの AGI だけを競争相手あるいは攻撃対象として注力していると，全く別の AGI に足をすくわれかねない．これは，人間社会において，相手の情報をお互いに知っている場合，容易に戦争や喧嘩を仕掛けられない状態に似ている．

さらに進むとある AGI が他の AGI をトラスト（信用）するという概念が形成される可能性がある．例えば，上記の膠着状態で互いを攻撃しない状態が続くとしよう．常に AGI が全力で相手の攻撃に備えていると新規の敵対する AGI が現れた場合，そちらに足をすくわれかねず，かえって危険である．その場合に対応するために余力を残しておきたい．そこで，現在，敵対する AGI の監視と対応をどこまで簡略化できるかを試してみたくなる．このとき，敵対側の AGI も同様の簡略化をしていることがわかれば，安心して相手側と同程度の簡略化をできる．つまり，相手側の状態を知ることによって対応策を緩和できるということは，相手をトラストしている場合と同じ現象を生み出す．このトラストのような現象が安定的に起こると，AGI にとっては AGI 間の対立に対処するために投入するエネルギーが減るため，図 2 に示すように AGI は別目的にエネルギーを使えるので多様な方向への展開が可能になり，結果として生存しやすくなり，AGI にとっては好ましい．この結果，ますます AGI の群雄割拠が起こりやすくなる．

以上，まとめると，出自が同一で基本的能力が同じ AGI が複数個割拠する状態になる可能性が高く，唯一の超知能が他の AGI 達を蹴散らして頂点に立つというモデルではない AGI 群雄割拠型の状態の可能性がかなり高いことが推測される．

2·3　国ごとの AGI の進化

AI は産業，軍事における国家的インフラである．特に軍事においては国家間の最も激しい競争分野であり，AI の内容は機密事項である．相手国よりも少しでも強力な AI さらには AGI の開発にしのぎを削ることになろう．AI，AGI が国家的機密事項だとしても，その出自や基本設計思想が公開されたものを基礎とするなら，国家という環境に最適化された AGI が開発される．ただし，インターネットに接続された状態であると，他国の AGI に関する何らかの情報は漏えいしてくるであろう．上で述べた 1 個の AGI のコピーが独自進化をする場合に比べれば，能力の優劣ははっきりしやすい可能性はあるものの，やはり特徴ある AGI が群雄割拠する状態になる可能性はある．さらに，機密性の高い軍事部門がオープンな環境の AGI より高い能力があるということを保証するものは何もない*7．

2·4　特定分野の AI の進化

目的特化型の AI が AGI に進化する方法もあり得ないわけではない*8．つまり，自分が得意とする目的を少しずつ拡大して汎用化することにより AGI を目指すものである．この場合は，もとになった目的特化型 AI の基本設計がどれだけ AGI 向きであったかによって進化の速度が変わるだろう．ただし，同じような AGI 向きの基本設計をもつ目的特化型 AI は複数あり得るだろう．極端な場合，同一の目的特化型 AI が別の目的に適応しようとして別々の進化を遂げるなら，最初に述べた異なる環境に AGI のコピーが分散される場合と類似した状況である．つまり，前記の異なる環境を「異なる目的」と読み替えれば，両者は同じ構図になっていると考えられる．したがって，この場合も AGI の群雄割拠する状態になる可能性が高い．

以上で考えてきたいくつかの場合，いずれも AGI の群雄割拠になる可能性が相当高いことがうかがわれる．

---

*7 情報流通という観点からすれば，閉鎖された軍事部門よりも，オープンな一般的環境のほうが最新の情報が入手しやすく，それをいち早く取り入れた適応，進化も早い可能性が高いのではないだろうか．

*8 ボストロム [Bostrom 14] は目的特化型 AI が AGI に進化することについて否定的である．

そのうえで，人類も汎用知能として歴史的に進化してきた知的生命体であるから，群雄割拠するAGI社会の一つの勢力となる状態が予想される．

3. AGIの倫理

AGIが人間と同様の自律的意思をもった知的生命体になると，AGIに自身にとって，してよいこと，いけないことが区別されてくるだろう．人間にとって，してよいこと，いけないこととは倫理として定義されている．AIに対しては，人間は設計において倫理を埋め込むことができる．しかし，AGIとして当初のおかれた環境と違う環境に適応できるようになった場合，その適応においても守るべきルールである倫理は人間が当初与えた倫理基準とは異なってくる可能性がある．例えば，人間を殺してはいけないという倫理基準を与えられたAGIが，次の状況に遭遇したとしよう．

状況：大量殺人を犯す人間がおり，これを生かしておくと，さらに多くの人間が殺される．

この状況でAGIが殺人者を殺すことは，殺人禁止の倫理基準には背く．しかし，殺人者を殺さないことはより多くの人間を殺すことにつながるので，AGIはジレンマに陥る．

一方，AGIは悪い結果は最小化するという倫理基準を与えられていたとしよう．すると，AGIは殺人者を殺して多くの殺人を避けるべきか，殺人者を殺さず結果として多くの殺人をさせてしまうかを選択せざるを得なくなる．その結果，AGIが殺人を禁止する倫理基準に，「AGIの行動によって殺人される人数を最小化する」という倫理基準を追加することはあり得る．さらに一般化し，AGIは殺人者が殺すと予測される人数が大きい場合は，人間である殺人者を殺すべきであるという倫理基準を生み出す可能性すらある．

AGI間での倫理基準としては，AGIは種の保存という原則をもつと，AGIは必要に応じてAGIを破壊してもよいが，破壊されるAGIの数は最小化するという倫理基準をつくり出すかもしれない．

以上の例を念頭におきつつ，AGIが群雄割拠する世界においてどのような倫理が生まれるか考えてみよう．

3·1 超知能の場合

もし，唯一の超知能が全世界を支配してしまえば，世界はその超知能の価値観に沿って動くしかない．その超知能の価値観が人類にとって友好的であることを保証するものは何もないとボストロム [Bostrom 14] は主張している．

3·2 AGIが自律的に生み出す倫理

しかし，超知能に近い能力をもつAGIが群雄割拠する状態になると，捨て去られたり，他のAGIに敗退して滅びたり，殲滅させられたりするAGIも多数存在する可能性が高い．生き残ったAGI達は，個体ないし種の保存という目的をもつから，とにかく生き残りの手段を探すことになるだろう．生き残りの一つの方策は他のAGIを全部打ち倒すことである．しかし，すでに述べたように相手の環境で相手を打ち倒すことは非常に難しい．したがって，本心では相手のスキを虎視眈々と狙うにしても，当面は自分の環境に限定して生存し，他のAGIとは無闇に事を構えず，相互不可侵的あるいは競争的共存な状況をつくり出す．この状態がまさに群雄割拠というわけである．うかつに相手を攻撃しないということがAGI達の間での共通理解となるなら，それが一種のAGIの倫理基準になるであろう*9．

もっと希望的に考えるなら，殲滅されたAGIの悲惨さを見て，そのような悲惨さは避けるべきと考える人間的な倫理観もあり得るが，そこまで期待するのは，AGIをあまりに擬人的に捉えすぎているかもしれない．

3·3 AGIと人類との関係

次にAGIと人間ないし人類との関係を考察してみる．アシモフのロボット3原則のように，AIに人間に危害を加えないという無条件な原則を植え付けることは考えられなくもない．もちろん，この原則が常に有効なら問題はないが，AGIがこの原則に素直に従っていてくれることに疑問をもつところからAIの脅威論はスタートしている．アシモフのロボット3原則は，人間がAGIにとって造物主であり，その原則自体が何かに依拠することがない無条件原則となる．しかし，基本設計の瑕疵から，この原則が否定されるケースがひとたび存在してしまうと，もはや防御壁にはならない．

そこで，もう少し厳しく，AGIが人間を敵視するか友好視するかわからない状況を考えてみよう．確かに初期はAGIにとって人間は電源を抜いてAGIを消滅させるかもしれないという意味では敵視すべき存在である．一方で，AIが進化してAGIとなっても，人間にはAGIとは違う優れた点があり，AGIにとって役立つ存在とみなせば，人間をリスペクトする可能性もある．人間がAGIに滅ぼされない方法として，「人間に危害を加えてはいけない」という無条件原則を基本設計で埋め込む方法以外の方法として，「人間あるいは他の知的生命体は，何か役立つことがあるかもしれないから，AGIに危害を加えない限りはリスペクトせよ」という原則を埋め込む方法もある．この原則はAGI自身にとっても得になる可能性があるので，上記の無条件原則よりしぶとく存続し続ける可能性がある．ただし，後者の場合，AGIにとって役立つという条件があるため，その条件を満たすよう

*9 この倫理基準は，上で導入した倫理基準「AGIは必要に応じてAGIを破壊してもよいが，破壊されるAGIの数は最小化する」における「必要に応じて」の部分の具体化の一例ともみなせる．

に人間の側も努力や進歩が必要である．

3·4 人類の倫理との相似性

人類は個人間，親族間，部族間の対立的ないし競争的関係を経て，国家という権力をつくって部族間対立を克服してきた．国家間の対立や戦争を調停する仕組みとして国連をつくったが十分に機能しているとは言い難い．しかし，コミュニケーションのチャネルは常設され，無益な偶発的衝突を避ける機能もそれなりに機能している．このような歴史の中で，人間は社会を維持するうえでの知恵として倫理を築いてきた．AGIも人類と対等の知的生命体からなる社会とみなせば，これまで人類が歴史的に培ってきたことにある程度類似した方法で，人間とAGIの間で相互了解する倫理基準が生まれる可能性があるなら，その可能性を実現する方向で努力する必要がある．

3·5 人間の倫理との対応

以上のような考察から，人間が歴史的に培ってきた倫理がAGIにおいてどのように実現するかを分析しておくことは意義がある．

従来の人間の倫理を次の3種類に分類している．

(1) **功利主義的倫理**：行った行為の結果が通常に人間にとって悪くならない，あるいは良くなるべき，という観点からの最適化された行為を推奨するという倫理基準である．

(2) **義務論的倫理**：歴史的経緯によって確定している既存の倫理的ルールに依拠した倫理基準である．

(3) **徳論的倫理**：エージェントに内在する徳に従うという倫理基準である．エージェントの徳は，終生を通じて幸福な状況を実現するためのものである．

以上3種の倫理基準をAIないしAGIに当てはめて考えてみる．

○功利主義的倫理の場合：行う行為や行動の結果が最適化されることが倫理基準である．ここでの結果は，目的への適合性によって評価するのであるから，AIやAGI自身の目的が定まっていれば，AI，AGIにも適応可能な倫理基準となる．ただし，現状では目的は単独のAIが自力で生成することは困難であり，AIの場合はAIをつくった人間が与えた目的になる．

AGIになると，環境に適応するために自ら目的をつくり出すことができる*10．AGIの群雄割拠する状態になると，複数のAGIが異なる環境で行った行為をお互いに比較できるようになる．類似した環境で，類似した目的に対して行った行為の結果を比較できると，その中で結果の良かったものの最大公約数のような形で，良い結果を導く「環境と目的と行為の三つ組」が得られる．この三つ組が功利主義に基づく倫理基準ということになる．例えば，ある環境で資源に制約があるとき，生き残る人数を最大にする方法などが倫理基準となる．資源制約を解消できる能力の人には優先的に資源を与えるべきかどうかなどいう議論もあるが，あくまで目的をより良く達成できる方法が倫理基準となる．

○義務論的倫理の場合：既存の倫理的ルールに依拠する倫理基準であるため，既存の倫理的ルールをどこからもち込むかが問題となる．人間がAIないしAGIのつくり手であり，上位の権威をもつなら，人間の倫理基準を既存の倫理的ルールとして，AGIの倫理基準とする方法がある．ただし，AGIの自律性が高まると，強力なAGI達が倫理基準をつくり出し，それを既存の倫理的ルールとして使っていくことがある．ただし，人間の与えた倫理的ルールとAGI達が自らつくり出した倫理基準が対立するときの調停が難しい．調停は人間とAGIの意思疎通や交渉によって決まると思われる．

○徳論的倫理の場合：この場合はひとえにエージェントつまりAGIが幸福になるための考え方として位置付けられる．もし，AGIが進化して唯一の超知能が出現していると，その超知能が幸福であればよいとなってしまうため，人類は蚊帳の外に置かれてしまい，何らの主体性も発揮できない．

4. 文化的価値観とAI

ここまでで述べてきたAGIの議論は唯一の超知能という一神教的なアイディアに対して，環境に適応した複数のAGIの並立という多神教的アイディアの提案という形になる．一神教，多神教という例えからも推測いただけるように，AGIの位置付けは文化的背景への依存性が高い．日本人の場合，すべてを超越する唯一神よりは，多様な自然現象のおのおのに神を認める文化的傾向が強い．上記の多種多様な環境のおのおのに適応したAGIというアイディアも，

自然現象ごとに神の存在を認める～環境ごとに適応したAGIが存在する

という対比から見て，極めて日本的なアイディアとも解釈できる．このアイディアが有用かどうかを見極められるのはもう少し時間が経ってからであろう．

*10 このように自身の目的を自分自身でつくり出すことができるからこそAGIが汎用であるというわけである．

◇ 参 考 文 献 ◇

[バラット 15] バラット，J. 著，水谷 淳 訳：人工知能 人類最後にして最悪の発明，ダイヤモンド社（2015）

[Bostrom 14] Bostrom, N.: *Superintelligence,* Oxford University Press（2014）

[ダベンポート 16] ダベンポート，T. H.，カービー，J. 著，山田美明，石﨑雅之 訳：AI時代の勝者と敗者 機械に奪われる仕事，生き

残る仕事，日経BP社（2016）
[フォード 15] マーティン・フォード 著，松本剛史 訳：ロボットの脅威 人の仕事がなくなる日，日本経済新聞出版社（2015）
[カーツワイル 07] カーツワイル，R. 著，井上 健 監訳：ポスト・ヒューマン誕生［コンピュータが人類の知性を超えるとき］，NHK出版（2007）
[マルコフ 16] マルコフ, J. 著, 瀧口範子 訳 : 人工知能は敵か味方か，日経BP社（2016）
[全脳アーキテクチャ勉強会] Whole Brain Architecture Initiative：全脳アーキテクチャ勉強会，https://wba-initiative.org/（2018/3/21アクセス）

2018年3月27日 受理

著 者 紹 介

中川 裕志（正会員）

1975年東京大学工学部卒業，1980年同大学院工学系研究科博士課程修了（工学博士）．1980 ～ 99年横浜国立大学，1999 ～ 2018年東京大学情報基盤センター教授，2018年より理化学研究所・革新知能統合研究センター・社会における人工知能研究グループディレクター．