人工知能学会研究会資料
SIG-AGI-001-02

汎用人工知能：ロマンとリアルの狭間から

Artificial General Intelligence: between Ideal and Real

江間有沙 [1]

Arisa EMA[1]

[1] 東京大学
[1]The University of Tokyo

Abstract: Artificial General Intelligence has a potential to influence human beings' value judgement and lifestyle. Thus it is important to involve various actors to its research from upstream. This report introduces a background of how and why inter-disciplinary collaborative studies are required and then suggests conducting risk management and risking communication from pre-crisis phase.

1　はじめに

人工知能とは何か．知能を工学的に探究するこの分野では，新しい知能を定義し挑戦していくフロンティア精神が共有されている．それは,「うまくいくということはその分野が体系化されているということを示している．人工知能はまだ体系化に至っていない」[1]や，「人工知能学会がとても特徴的で魅力的だと思う点は，学会の対象物である「人工知能」がまだ見ぬものであることだと思う」[2]などの発言からも伺える．

通常の学問領域が体系化を目的としていることを踏まえると，この考え方は特殊である．しかし「体系化に至っていない」からこそ，様々な分野との協働が可能となる．特に汎用人工知能という多義的な解釈を保持できる「まだ見ぬ」技術は，異なるアクターを結びつける可能性を秘めている．

本稿では，汎用人工知能というロマンある技術に対し，協働研究や技術評価の論点を整理することで，現実的に今後の研究動向を推進していくための提案を行う．

2　背景：異分野間協働研究の要請

人工知能学会の年次大会にある表題を見ると教育，心理学，脳科学，音楽，環境学，経営学，医療など扱っているテーマは幅広く，多様な分野との関わりが伺える．関わり方は多様であり，人工知能技術を用いて社会問題や環境問題の解決策を模索すること，音楽や学習など人間の創造的営みを理解すること，あるいはインタビュー調査やエスノメソドロジーなど社会科学の手法を用いることや，知能の定義を考えるために哲学的な思索が必要になることもあるだろう．また，実用段階においてはセキュリティやプライバシー問題など法や経済との接点もある．これらの研究の多くは問題の解決や人間の理解といった不確実なものを明らかにし，工学的なアプローチを可能にしていくための協働研究である．

しかし，世の中には答えがない問題，答えがわからない問題，答えられない問題等がある．マイケル・サンデルの白熱授業で扱っているような道徳・倫理の問題は，個人の選好，文脈，文化や時代背景によっても異なるため，多様な視点から考え続けていくほかない．このようなことを述べると，よく「人文・社会科学者は批判をするだけで解決を示してくれない」「書類書きなどの仕事を増やすだけ」という批判が聞かれる．あるいは「問題を指摘するのではなく，現実に倫理審査をクリアするための方法をアドバイスしてほしい」と言われることもある．しかし，人工知能が人間の価値判断や思考に干渉する可能性がある限り，これらの問題を避けて通ることはできない．

近年，情報科学においても倫理的・法的・社会的影響（Ethical, Legal, Social Issues: ELSI）に配慮して研究を行っていく必要性が助成機関からも要請されるようになってきている[3]．欧州においては責任ある研究・イノベーション（Responsible Research and Innovation: RRI）という用語が 10 年ほど前から政策文書で使われるようになってきている．ELSI が研究のリスクを警戒しイノベーションのブレーキを掛けるのではないかと見なされているのに対し，RRI は多様なアクターを巻き込んで協働を進めることによってイノベーションガバナンスを行っていこうという視点であるとされる[4].

3　技術と社会の捉え方

どのように技術や社会像を捉えるかによって，人工知能の目的や方法論は異なる．人工知能の目指すべき方向を語るときに「宇宙派か地球派か」という問いの立て方がある．本節では，宇宙派・地球派ではなく，科学技術と社会一般に関して論じられる概念である技術決定論（technological determinism）を補助線として，人工知能と社会の関係を整理する．技術決定論は，技術が社会と独立に発展するという立場と，より強く技術が社会を変化，形成するという立場の二種類がある．この考え方は「情報化社会」の論じられ方にも用いられており，1960 年代後半から「情報技術が社会を変える」と語られ続けて現在に至る[5]．情報技術に関する技術決定論は至る所で使われている．現在では「人工知能が雇用を奪う」のではないかとの懸念を生じさせるようなシミュレーション結果が多く発表されているが[6][7]，このようなマクロな分析は技術の発展ありきで，それとは別に個人の選好や社会の多様性はある程度固定して計算するため，技術決定論との整合性が高い．

一方，技術と社会は分離可能ではなく，人間の様々な活動やプランを文脈から切り離すことはできないという考え方は，Human Computer Interaction 研究などでもおなじみであろう[8]．これらの研究は，技術は両義的で「多様に安定する可能性」を持つため設計された機能には還元できないとする現象学的アプローチや[9]，技術と社会は共進化するものであるため作動中（in the action）の技術と社会の関係性を記述する人類学的な研究[10]とも親和性が高い．このような研究は技術決定論が前提とする技術と社会の独立性や，技術が一方向的に影響を社会に与えるという考え方を批判する．しかし，文脈や環境とのインタラクションなどに着目するこれらの観点も，技術ありきで問題設定を行っているため，ある種の技術決定論的な考え方からは抜け出せてはいないとの指摘もある[11]．

技術と社会の関係性を論じるときに，どのような立場に立つのかの前提を理解しておくことは重要である．現在の汎用人工知能と社会の関係の議論は長期的な視野にたった技術決定論的な議論に偏りがちである．また，シンギュラリティに関する議論の多くは，「汎用人工知能が普及してしまったら社会はどうなるか」というまだ見ぬ技術をあるものと固定して議論する．そこにはその技術を欲している人の価値観はどのようなものなのかなど，現在の価値観との接続や具体的な社会像が見えにくい．さらには，汎用人工知能を語るときに意識や人格という人間のメタファーを用いることや，ＳＦによるイメージが容易である点も期待や懸念を加速させる要因となっている．人間のように振る舞える人工知能というメタファーはセンセーショナルであるがゆえに，メディアだけではなく企業や政策研究者なども巻き込みながら一大ブームを作り出しているし，作り出そうとしているようにも見える．海外でも脅威論と期待が相まって，人工知能に関するセンターや助成金プロジェクトなどが開始されている[12][13].

4　不確実性の問題

技術の事前評価（assessment）は事後評価（evaluation）と異なり難しい．特に新技術の初期段階では発展の方向性をコントロールできるが，影響についての情報はほとんどないという「情報の問題」がある．一方，技術が広く使われるようになると影響に関する情報は大量にあるが，発展の方向性を制御することは難しいという「力の問題」があるというジレンマが指摘されている[14].

汎用人工知能など「まだ見ぬ技術」に関しては，「情報の問題」が重要となり，それは不確実性の問題とも言い換えることができる．研究とはそもそも不確実なことを明らかにしていく作業であり，その作業で得られた知見の「確からしさ」はピアレビューという科学者共同体内部の審査システムによってチェックされるということが，17 世紀から現代にいたるまで行われてきた[15]．計測機器の技術的な限界や研究者の仮説や前提の立て方などの方法論的な限界，さらには倫理的な制約により実験的解析が不可能な場合（人体実験など）や，費用が掛かりすぎるなどの金銭的な問題により検証ができないこともある．これらの要因による不確実性をなるべく減らしていこうとする知的な努力が，研究の根本である．研究の方法や成果が研究者だけに閉じられている場合，不確実性は知的好奇心を駆り立てるものとなる．しかし，科学技術の研究は社会的な営みの一つでもあるため，不確実であると問題が起きる場合がある．汎用人工知能がもたらすと言われている様々な問題（雇用問題やコントロール喪失の問題，プライバシー問題）も社会の価値判断に左右される一方，社会的な影響も大きい．

このように不確実なものに対応するために，リスクという概念が一般的に用いられる．リスクとは厳密には危害の内容とその発生確率も知られているときに使われる．ただし広い意味では「人の健康，生命の質，あるいは環境の質に対して障害をあたえるチャンスを意味するもの」[16]としても使われるため，どのような意味で使われているのかは注意が必要である．リスクは安全か危険かというゼロかイチ

かという議論から脱却する視点を提供する．また，あるリスクを最小化しようとしたときに対抗リスクが生じてしまう場合はそのリスク・トレイドオフ（かけひき）を議論することが可能になる．その場合の論点としては，大きく対抗リスクによる悪影響が目標リスクと同じものか違うものか，また対抗リスクを受ける集団が目標リスクの物と同じかどうかが挙げられる[16]．対抗リスクによる悪影響が目標リスクと同じものはたとえば，人工知能の発展により雇用が喪失する可能性もあるが，経済成長へと導く可能性があるというものであり，これは同じ経済という指標の中でのトレイドオフである．しかし，人工知能の発展で恩恵を受ける人と損失を受ける人が異なる場合も考えられる．これは，対抗リスクを受ける集団が目標リスクの物と異なることを意味する．他方，プライバシーの問題やロボット兵士などのデュアルユース問題は，経済，政治，心理，倫理など様々な観点からの議論が可能となり，問題の難易度は上がる．しかし，そのリスクを数値化，比較できるようになれば具体的な対策まで議論ができるため，リスク評価やリスク分析は環境対策や医療政策などに取り入れられている．

ある物事をリスクとして定義することで議論や合意形成がしやすくなるが，そもそも何をリスクと見なすのか，またリスクとして数量化する方法は妥当なのかという根本的な議論を忘れてはいけない．リスクという観点を導入すること自体が大変狭い範囲でのコンセンサス（合意）を求めるものであり，拒否すべきものも拒否できず受入管理をするという議論の枠組みに取り組まれてしまう．そのため，リスク評価の土台に乗らないという選択肢を考えるべきだとする考え方もある[17].

また，たとえリスク評価の手続きや結果が合理的であるとされたとしても，それが社会的に受け入れられるかというのはまた別の話である．人々がリスクなどの合理的な考え方を理解すれば，その決定などを受け入れてくれるはずだという考え方は「欠如モデル」として批判されてきた．例えば遺伝子組み換え作物においても，遺伝子組み換え作物の情報を正しく理解しているからといって，それを受け入れるかというとそこの相関関係はないことが研究で示されている[18]．同様のことは人工知能にも当てはまるだろう．結局のところ，人間がその人工知能を取り巻く環境や人工知能そのものを信頼できるかどうか，またその問題が当人にとってどのくらい重要であるのかなどによって評価は異なってくる．たとえば，今日の夜に何を食べるか，目的地までどのようなルートを選ぶかなどは，我々はすでに情報技術の提案に従うことに違和感はない．しかし，たとえば進路や伴侶の選択，生命，軍事に関することなども人工知能に委ねることができるかどうかは，まだ人によって判断が分かれるところだろう．人工知能が人のパートナーとなりうるかの研究は興味深く「人狼知能」においても，文脈を読むこと，信頼を得ることが人工知能にできるのかを扱っている[19].

このような観点から，リスク認知，リスク・コミュニケーションなどの研究領域においては，個人の置かれている状況，あるいは信頼や価値観などの文脈を理解し，対話を重ねていくことの研究が蓄積されている[20]．また科学技術社会論や公共政策の分野においては，合意形成のメカニズム，手続き的な妥当性を担保するための意思決定システムの在り方についての研究も行われている[21][22].

5 提案

5.1 平時のリスクマネジメント

汎用人工知能は人間の生き方や価値観に干渉する可能性が高いため，何をリスクと見なすのか，対抗リスクをどのように見積もるのかなどの枠組みを設定する段階から考えていく必要がある．技術は社会との相互作用でその意義や役割を獲得していくものであり多様安定性[9]を持つ．そのためにも何がリスクなのか，そもそもリスクとしてとらえて良いのかという発散系の議論と，少なくとも現在の枠組みにおいては 5 年先など短期的に想定すべきリスクとは何かを特定し，その対策を考える収束型の議論を平時から行っていく必要がある．

近年，クライシス・コミュニケーションという危機管理のためのコミュニケーションの研究が欧米で盛んであり，事件や事故などの危機的状況の対応をいかに組織的な学習へと結び付けられるかの研究が行われている[23]．クライシスはいつどのように起こるか予測はできない．そのため平時から様々な危機に対して研究者間でコミュニケーションをし，またリスクが起きた時にどのような対応ができるかを検討する必要がある．

たとえば，情報技術コミュニティで大きなクライシスとして挙げられるのは Winny 事件である．事件後，情報処理学会と情報ネットワーク法学会は「Winny 事件を契機に情報処理技術の発展と社会的利益についてを考えるワークショップ」を開催した．学会として Winny 事件に声明を出すよう求める声もあったが,「当時の学会内にはさまざまな意見があったうえに，全員が全容をつかんでいるわけではなかった」ために見送ったという[24]．その他の事例として表紙問題も記憶に新しい．表紙問題はその後，

編集委員会にてイラストレーターとの密なコミュニケーションを行う体制が整ったという[25]．また倫理委員会も発足し，何か起きた時にどのように対応をしていくのが良いだろうかという具体的な対応策やシミュレーション，あるいは学会としてのスタンスのようなものを発信していくことを目的としているという[26].

それらは若手の研究者を巻き込んで議論する体制を作っていくことが望ましい．次世代の技術や価値観を作っていくのは若手研究者である．また，人工知能研究はフロンティア精神あふれ研究テーマのサイクルや実用化へのタイムスパンも生命科学などの他分野に比べて展開は早い．最先端を行っている研究者が技術と社会の相互作用についても考えていけるような仕組みを作っていくことが重要である．現在，助成金等の申請には社会的影響について記入する項目が必須となりはじめているほか，インターフェイス系の研究論文誌では，自分の論文の将来の社会への影響や考察を書く項目が存在しているという．申請書や論文の作成を通して，研究にブレーキをかけるためではなく，自分自身の研究と社会との関係性について考えるきっかけを作っていくことも研究コミュニティの役割として重要になってくるだろう．

冒頭に人工知能研究は体系化されておらずフロンティア精神が共有されていることを述べた．しかし，研究テーマがフロンティアであっても研究者の社会的責任に対する意識が未熟であって良いというわけではない．フロンティアであることは新しい価値に挑戦するということも含んでいる．司法の介入や社会的なバッシングから研究にブレーキがかかってしまう可能性を軽減させるためにも，研究者コミュニティとしてリスクマネジメントを行っていくことが，今後さらに重要となってくるだろう．

5.2 社会との対話

リスクマネジメントためには,「我々の社会は汎用人工知能によってどうなるのだろうか」という技術決定論的な切り口ではなく,「我々の社会をどうしたいのか」を議論していくことが求められている[27]．汎用人工知能を作っていくプロセスにおいて，様々な分野の人たちと，知能とは，意識とは何かを議論していく仕組みや場を形成していくことも必要であろう．

そのために，センセーショナルな視点からのみ話題提供をしていると前回のブームの二の舞となる恐れがある．汎用人工知能ができたらこんなことができるという技術ありきでの議論は脅威論や過度な期待論を生じさせる．しかし，技術への期待が大きければその反動も大きい．人工知能は過去 2 回のブームとその後の冬の時代を経験している．5 年後，10 年後にできることと， 50 年後，100 年後にできることの間にはかい離がある．ブームによって資金や人，モノのネットワークが強化されることは，技術の発展のためには望ましい．しかし，2 度のブームを経験した今，再び冬の時代が来る前に，ブームのソフトランディング法を過去から学び考えていく必要がある．技術者の責任としては，地道に現実にできることも併せて発信していくことが重要になっていくだろう．そのためにも，メディア研究，レトリックやコミュニケーション研究との連携や，過去のブームについての在り方を歴史的に振り返る視点が重要になる．

このような議論をしておくことは，国内だけではなく，対外的にも重要である．前述したようにロボットや人工知能に対する ELSI あるいは RRI など人文・社会科学との協働研究は欧米において盛んにおこなわれている．欧米で作られた基準や倫理規定をそのまま採用するだけではなく，自らの価値規範や尺度などを打ち出していくことが現在求められている．その意味では，この話は 2 節で上げた他分野との協働の話とも密接につながってくる．

6 汎用人工知能への期待

「汎用人工知能」を何の目的のために作るのかは，人によって様々であろう．人間のように自ら学習する人工知能をつくる場合，人工知能が読み込むデータセットの「質」が問題となる．何を読み込むか，どう読み込むかというのは，人間が読み込ませるデータを考えているときも問題となることである．何をどのように読み込んだのかというプロセスが記録として残り可視化されている限り，後で人間がその人工知能の読み込んだデータの偏りや質について検証することが可能である．これは人が意思決定をするにあたって依拠している価値判断やフレーミングを，文献やインタビュー調査から可視化する作業に似ている．問題は読み込んだデータや学習のプロセスがブラックボックス化されていることだが，これは人間においても同じことが言える．逆に言うと人間の思考回路や価値判断はブラックボックスのようなところがあり，そのために集合の意思決定が困難にあったり，研究不正などが起きてしまっていたりする．人間を模した汎用人工知能をつくるという場合，人間が行っている不合理なこと，不正なこともできるようになるとされる．

一方，データの質さえ保障できれば，ある特定の価値や学問に特化した人工知能を作り，問題解決を

してもらうことはそれほど遠くない未来に実現するのかもしれない．実際，専門性の高い職種ほど雇用が奪われやすいのではないかとも指摘されている．しかしそこで産出された合理的な最適解がそのまま受け入れられるかというと，そうではないということは前述した．たとえば公平で公正な判決が出せるからといって，人工知能裁判官を人々が受け入れられるかどうかは判断が分かれるところであろう．人々の価値判断や選好が多様である限り，合理的であるということはその提案を受け入れることに直結はしない．

提案を受けさせるのではなく，新たな判断や選択肢の発見支援目的である特定の価値や学問に特化した人工知能を使うこともできる．前述したリスクマネジメントや異分野との対話をいきなり人間同士で実施するのではなく，ためしにそのような多様な価値観を表示させることができる人工知能を使うことで, 視野や選択肢を広げる使い方ができる. 実際に, ディベートのための資料を提供するような人工知能や「予想外」の視点を提供してくれるデータツールはすでに実装されている[28].

以上，汎用人工知能に何を期待するかを考えるだけでも様々な視点があり，そこでは価値とは何かを多様な研究者やアクターを巻き込みながら議論をしていく土台が必要となる．冒頭に述べたように，汎用人工知能という多義的な解釈を保持できる「まだ見ぬ」技術は，異なるアクターを結びつける可能性を秘めている．問題の洗い出しができるような対話の場が形成されるのは夢に過ぎないのか，あるいは現実に可能であるのか．それを研究会では投げかけてみたい．

謝辞

本研究の一部は科学研究費補助金（挑戦的萌芽研究）「人工知能の規範・倫理・制度に関する対話基盤と価値観の創出」，国立情報学研究所・公募型共同研究「情報と社会の系譜学」，国際高等研究所・研究プロジェクト「人工知能に関する問題発掘型対話基盤と新たな価値観の創出」の助成による．
また様々な視点や助言を下さった研究会メンバーをはじめ様々な場所でお会いし，お話をお伺いさせていただきました皆様に感謝申し上げます．

参考文献

[1] 松原仁 : 人工知能における「読んでおくべき本」， 人工知能学会誌, vol.12, No. 1, pp. 36‑43 (1997)
[2] 松尾豊，山川宏：人工知能学会 25 周年特集「四半世紀を越えて」にあたって, 人工知能学会誌, vol. 26, No. 6, pp.553 (2011)
[3] 独立行政法人科学技術振興機構研究開発戦略センター：科学技術未来戦略ワークショップ「知のコンピューティングと ELSI/SSH」, (2013)
[4] 吉澤剛：責任ある研究・イノベーション: ELSI を越えて，研究技術計画, vol. 28, No. 1, pp. 106–122 (2013)
[5] 佐藤俊樹：ノイマンの夢・近代の欲望：情報化社会を解体する，株式会社講談社, 1996
[6] Frey, C. B., Osborne, M. A.: The Future Of Employment, http://www.futuretech.ox.ac.uk/future-employment-how-susceptible-are-jobs-computerisation-oms-working-paper-dr-carl-benedikt-frey-m (2013)
[7] 株式会社野村総合研究所：日本の労働人口の 49%が人工知能やロボット等で代替可能に, http://www.nri.com/jp/news/2015/151202_1.aspx, (2015)
[8] サッチマン, A.ルーシー：プランと状況的行為：人間-機械コミュニケーションの可能性（1999)
[9] アイディ,ドン:技術と予測が陥る困難，思想，926, pp.145‑156 (2001)
[10] ラトゥール,ブルーノ:科学が作られているとき―人類学的考察，産業図書 (1999)
[11] Wyatt, S,:Technological Determinism Is Dead: Long Live Technological Determinism. E. J. Hackett, M. Lynch, J. Wajcman, & O. Amsterdamska eds., The Handbook of Science and Technology Studies. Cambridge, Massachusetts London, England: The MIT Press. (2007)
[12] 江間有沙：「人工知能と未来」プロジェクトから見る現在の課題，人工知能学会全国大会 2015 予稿集 , https://kaigi.org/jsai/webprogram/2015/pdf/2I5-OS-17b-1.pdf
[13] 西下佳代,茅明子，矢島彰夫,奥和田久美，人工知能やロボットの社会的影響に関する先行的研究動向，第 30 回研究・技術計画学会予稿論文集（2015）
[14] Collingridge, D,:Social Control of Technology, Continuum International Publishing Group Ltd (1980)
[15] 藤垣裕子 : 専門知と公共性-科学技術社会論の構築へ向けて，東京大学出版会（2003）
[16] グラハム, D. ジョン : リスク対リスク-環境と健康ノリスクを減らすために，昭和堂（1998）
[17] ラングドン，ウィナー：鯨と原子炉，紀伊國屋書店（2000）
[18] Bucchi, M., Neresini, F: Biotech remains unloved by the more informed. Nature, vol. 416. No. 6878, pp. 261 (2002)
[19] 鳥海不二夫，梶原健吾，大澤博隆，稲葉通将，片上大輔，篠田孝祐: 人狼知能サーバの構築，ゲーム

プログラミングワークショップ 2014 論文集, pp. 127-132 (2014)
[２０] 肇子吉川:リスク・コミュニケーション―相互理解とよりよい意思決定をめざして, 福村出版 (1999)
[２１] 小林傳司編:公共のための科学技術, 玉川大学出版（2002）
[２２] 平川秀幸 : 科学は誰のものか-社会の側から問い直す, NHK 出版（2010）
[２３] Coombs, W. T.:Parameters for Crisis Communication, in The Handbook of Crisis Communication, W. T. Coombs and S. J. Holladay eds., Wiley-Blackwell, Oxford, UK (2010)
[２４] Winny 問題を考える学会ワークショップ, IT media, Winny 問題を考える学会ワークショップ, http://www.itmedia.co.jp/news/articles/0406/28/news023.html
[２５] 大澤博隆, 2015 年表紙更新にあたって-前年の「表紙問題」のまとめとこれから-, Vol. 30, No. 1, pp. 2-6 (2015)
[２６] 松尾 豊, 西田 豊明, 堀 浩一, 武田 英明, 長谷 敏司, 塩野 誠, 服部 宏充, アーティクル「人工知能学会 倫理委員会の取組み」,人工知能, vol. 30, No. 3, pp. 358-364 (2015)
[２７] 堀浩一：シンギュラリティへ向けてあなたと私はどうしたいか? 情報処理, Vol. 56, No.1, pp,41-43 (2014)
[２８] 株式会社日立製作所：論理的な対話を可能とする人工知能の基礎技術を開発, http://www.hitachi.co.jp/New/cnews/month/2015/07/0722.pdf, (2015)