研究概要 2026

伊藤孝行（京都大学大学院情報学研究科）

2026年9月29日

本稿は、第18回京都大学公共政策大学院・JIAM連携セミナー「AI時代の"人間力"〜変化の中で、自ら判断し、行動するために〜」（2026年9月29日、京都大学）における招待講演「AIエージェントによる人間の合意形成支援と今後の展開」の内容をまとめたものである。

AI時代の合意形成とエージェント統治

正当性・メカニズムデザイン・人間の挑戦

要旨

生成AIとAIエージェントの急速な発展は、社会的意思決定のあり方と、人間が担うべき役割の双方に根本的な問いを投げかけている。本稿は、筆者が従事してきたAIを介した合意形成支援とマルチエージェントシステムの研究を踏まえ、この問いに対する筆者の立場を体系的に論じる。第一に、AIを介した合意形成の正当性は、理想的な人間ファシリテーターとの比較ではなく、現実に利用可能な代替手段との比較において評価されるべきであり、さらに制度設計そのものを市民の協働に開くことによって根本的に解決されうる。第二に、アフガニスタンにおけるランダム化比較試験は、AIファシリテーターがインタラクションの促進と議論の規律化を通じて民族間の偏見を低減させうることを示唆する。第三に、AIエージェントの真のリスクは単一エージェントの逸脱ではなく、善意で設計された複数エージェントの相互依存から生じる創発的帰結にあり、その対処にはガードレールを超えたメカニズムデザインの拡張が必要である。第四に、AIが知的作業を代替する時代において人間に固有なものは、有限な身体と時間に根ざした「何に挑むかを選ぶ力」であり、教育の役割はこの力を育てることにある。これらの論点を貫くのは、AIの能力と限界を現実的に評価した上で、人間の役割を制度の設計、ローカル知識の供給、および挑戦の選択に再定義するという姿勢である。

キーワード: 合意形成支援、AIファシリテーション、マルチエージェントシステム、メカニズムデザイン、集団知性、熟議民主主義、身体性

1. 序論

1.1 問題の所在

大規模言語モデル（LLM）に基づく生成AI、およびそれを自律的に動作させるAIエージェントの発展は、この数年で質的に異なる段階に入った。バイブコーディングからエージェンティックソフトウェアエンジニアリングへと至る潮流は、専門的な訓練を受けていない者でさえ自らの問題意識をソフトウェアとして実装できる時代をもたらしつつある。筆者自身の経験でも、従来三ヶ月を要した開発が一日で完成するという事態が日常化している。

この変化は、二つの根本的な問いを提起する。一つは、社会的意思決定においてAIがどのような役割を担いうるか、そしてその役割はどのように正当化され、統治されうるか、という問いである。もう一つは、AIが知的作業の多くを代替するとき、人間に固有に残るものは何か、という問いである。この二つの問いは独立したものではない。AIの役割を規定することは、同時に人間の役割を規定することである。

1.2 筆者の立足点

筆者は、マルチエージェントシステム、自動交渉、メカニズムデザインの研究に長年従事し、近年はAIエージェントがファシリテーターとして議論を支援する合意形成支援システム D-Agree の開発と社会実装を進めてきた [1][2][3]。D-Agree は国内の複数の自治体において施策策定や市民参加に用いられ、またアフガニスタンにおいては民族間の対話を対象とするランダム化比較試験（RCT）が実施されている [4][5][6]。本稿の議論は、これらの理論研究と実証研究の双方に立脚している。

1.3 本稿の構成

本稿の構成は以下のとおりである。第2節では、AIを介した合意形成の正当性と責任の問題を論じ、比較的正当性と制度設計の民主化という二つの観点を提示する。第3節では、アフガニスタンにおけるRCTの知見を検討し、AIファシリテーションが偏見を低減させるメカニズムについて仮説を述べる。第4節では、生成AIの構造的限界としてのローカル知識の問題を論じ、合意形成における人間とAIの役割分担を明らかにする。第5節では、議論をマルチエージェント系の統治へと拡張し、ガードレールの限界とメカニズムデザインの必要性を論じる。第6節では、AIの将来に関する不確実性の下で技術といかに向き合うべきかを論じる。第7節では、人間に固有なものとしての身体性・有限性と挑戦の関係を考察する。第8節では、これらを踏まえてAI時代の人間の役割と教育のあり方を論じる。第9節は結論である。

2. AIを介した合意形成の正当性

2.1 正当性と責任をめぐる異議

AIを介した合意形成支援に対しては、次のような原理的な異議が提起される。AIは最終的に責任を負う存在ではなく、選挙民主主義の規範を共有しているわけでもない。AIに不正をさせないためのルールを組み込めば組み込むほど、設計者の規範がシステム全体を左右することになる。しかし設計者は、たとえ優れた研究者であったとしても、選挙民主主義の手続きによって正当性を付与された存在ではない。この異議は正当なものであり、D-Agree を設計しAIエージェントを組み込んだ筆者自身が、常に向き合わなければならない問いである。本節では、この異議に対して二つの観点から応答する。

2.2 比較的正当性：現実の代替手段との比較

第一の観点は、正当性を絶対的な基準ではなく、現実に利用可能な代替手段との比較において評価すべきだというものである。これを本稿では比較的正当性と呼ぶ。

アフガニスタンにおける筆者らの実践は、この観点の重要性を端的に示している。同国において民族間の対話を人間のファシリテーターに委ねようとすれば、そのファシリテーターがどのような背景を持ち、どの民族の出身であるかという問題が直ちに生じる。特定の民族に属するファシリテーターは、他の民族の参加者から中立とは見なされない。外国人、たとえばアメリカ人やヨーロッパ人を起用した場合も、別種の偏りと不信が避けられない。このような状況においては、人間が司会するよりもAIにファシリテーションを委ねる方がむしろ好ましいという意見が、現地の当事者から実際に表明された。

この事例が示すのは、正当性をめぐる議論がしばしば暗黙に前提としている「中立で公正な人間ファシリテーター」が、現実には存在しない場合があるということである。人間ファシリテーターもまた、出自・民族・国籍に由来する偏りと不信から自由ではない。もちろん、AIが正確かつ公平にファシリテーションを行えるかという問題は残る。しかしその問いは、理想的な基準に照らしてではなく、「人間と比較したときにどうか」という形で立てられるべきである。正当性の議論は、この比較を一つの軸として進められるべきである。

2.3 制度設計の民主化：設計者と市民の分離を超えて

第二の観点は、より長期的かつ根本的なものである。前述の異議は、「設計者」と「市民」が分離していることを前提としている。設計者の規範が全体を左右するという懸念は、少数の設計者が多数の市民に対して制度を与えるという構図において初めて成立する。しかし、この構図自体が技術的に変容しつつある。

序論で述べたように、誰もが自らの問題意識をソフトウェアとして実装できる時代が到来しつつある。この技術的条件が整うならば、市民一人ひとりが自らの考える選挙の仕組みや、意思決定のあり方そのものをソフトウェアとして記述し、それらを統合していくような仕組みが構想できる。これは現時点では構想の段階にあるが、制度設計そのものを市民が協働して行うことが、ソフトウェアを通じて実現できる可能性を示している。

この構想は、設計者と市民の分離を前提とするのではなく、制度設計の行為自体を民主的な協働に開くことによって、正当性の欠缺を解消しようとするものである [17]。従来の集団知性研究が、与えられた制度の下で集団がいかに良い判断に到達するかを問うてきたのに対し、ここで問われているのは、制度そのものを集団がいかに設計するかである。これは集団知性の概念を一段高い水準へと拡張するものであり、民主的に作られた仕組みをより効率的に実現する道を開く。

2.4 技術的構想と現実政治の距離

もっとも、以上の二つの観点は、現実政治における障壁を直ちに除去するものではない。AIの規制にせよ、市民参加による新しい民主主義のルール作りにせよ、現在の選挙民主制の下で正当性を持つ政治的意思決定者がそれを認めない限り、実現しない。これは研究者が単独で解決できる問題ではなく、極めて実務的な問題である。技術的構想と現実政治の間には技術だけでは埋められない距離があり、それを繋ぐのは、政治を動かす立場にある人々との直接的な対話であり、場合によっては研究者自身が政治の場に踏み出すことである。筆者は、この種の実践を技術研究の外部にあるものとして切り離すのではなく、技術の社会実装に不可欠な構成要素として位置づける。

3. AIファシリテーションの効果：アフガニスタンにおけるランダム化比較試験からの示唆

3.1 実験の概要と初期値の解釈

筆者らは、アフガニスタンにおいて、AIファシリテーターを介在させた議論と介在させない議論を比較するランダム化比較試験を実施し、民族間の偏見の度合いがどのように変化するかを測定した [4]。この種の実験結果を解釈する際に注意を要するのは、実験開始時点における群間の差の扱いである。本実験では、AIあり群の方が開始時点での偏見が高く、AIなし群の方が低かった。この差はランダム化によって生じた単なる確率的な差であって、介入の効果を示すものではない。ランダム化比較試験においては、参加者が無作為に割り当てられるため、出発点において偶発的な差が生じることは避けられない。評価すべきはその後の変化、すなわちAIを導入したことによって偏見の度合いが低下していったという点である。

3.2 偏見低減のメカニズムに関する仮説

なぜ人間だけで話し合う場合と比較して、AIが介在すると偏見が軽減されるのか。筆者は、現時点で二つの要因を作業仮説として提示する。

第一の要因は、インタラクションの促進である。話し合いの中で参加者が発言すると、AIがそれに応答する。この単純なやり取りだけでも、応答があることによって他の参加者がさらに入力を行うという連鎖が生じる。仕組みとしては極めて単純であるが、何らかの反応が返ってくることがインタラクションを促進し、インタラクションそのものが豊かになる。偏見の多くが接触と対話の欠如によって維持されることを考慮すれば、インタラクションの量と質の向上が偏見の低減に寄与するという経路は十分にありうる。

第二の要因は、ファシリテーターによる監視の効果である。エージェント、すなわちファシリテーターが議論を見ているという状況においては、参加者は無責任な発言や逸脱した発言をしにくくなる。議論が荒れることが抑制され、議論自体が良い方向へと進む。これは、人間のファシリテーターが果たす規律化の機能を、AIが中立性の疑念を伴わずに果たしていることを意味する。前節で述べた比較的正当性の議論と照らし合わせれば、この機能は、人間のファシリテーターが現実には果たしにくい機能である。

この二つの要因、すなわちインタラクションの促進と議論の規律化が組み合わさることによって、偏見の度合いが低下していったというのが筆者の仮説である。

3.3 研究の現状と今後の課題

このメカニズムの解明はなお分析の途上にある。現在、次の実験、そしてその次の実験が進行中であり、さまざまなことが明らかになりつつある [20]。上記の説明は確定した結論ではなく、現時点における作業仮説として位置づけられる。二つの要因の寄与を分離して測定すること、および効果の持続性を検証することが、今後の課題である。

3.4 適用領域の現状と議会への展開

D-Agree のこれまでの適用対象は、基本的に、何らかの施策を策定する、新しい商品を開発する、学校内で新しい企画を立案するといった場面である。そこでの参加者は、市民、企業の社員、学校であれば学生と教員であり、彼らが意見を出し合い、それをまとめていく過程において用いられてきた。議会そのものにおいて導入された事例はまだ存在しない。筆者は、議会における導入を次の段階の重要な展開と考えており、システム面の運用と管理を全面的に担う用意がある。

4. ローカル知識の問題：生成AIの構造的限界と役割分担

4.1 生成AIが苦手とするもの

行政における合意形成には、住民の納得感、信頼、地域の歴史といった、AIには最適化できない要素が含まれるとしばしば指摘される。筆者はこの指摘に同意するが、その理由を曖昧な「人間らしさ」に求めるのではなく、大規模言語モデルの構造的特性から説明すべきだと考える。

ChatGPT や Claude に代表される生成AIが苦手とするのは、ローカルな問題、すなわちローカルな知識や情報に基づく推論や応答である。卓上の例で言えば、いまこの場にあるものが半分になってしまったがどうすべきか、というような状況は、AIの学習データには存在しない。学習データは大規模であるが、それは世界についての一般的な知識であって、特定の場所、特定の時点、特定の人々に関する微細な事情ではない。

この限界は、行政における合意形成においてとりわけ顕著に現れる。たとえばある市町村において、特定の踏切を廃止して高架化したいという課題があるとする。この人とこの人は賛成しているが、この人はこういう理由で反対している、といった事情は、AIに問い合わせても分からない。合意形成の対象となるのは、このような個別事案が大半であり、それゆえに困難が伴う。

4.2 合意形成における人間とAIの役割分担

この限界から導かれるのは、AIの利用を断念すべきだという結論ではなく、合意形成における人間とAIの役割分担に関する明確な原則である。すなわち、ローカル知識の供給源はあくまで人間であり、AIの役割はそれを引き出し、構造化し、やり取りを促進することにある。D-Agree の設計思想はまさにこの原則に基づいている [1][7][8][18][19]。D-Agree においてAIは、結論を最適化して提示する主体ではなく、参加者が持ち寄るローカルな知識と意見のやり取りを促進し、整理するファシリテーターとして位置づけられている。第3節で述べた偏見低減のメカニズムが、いずれもAIの知識ではなく参加者間のインタラクションに関わるものであったことは、この役割分担と整合している。

4.3 ローカル知識の統合に向けて

このようなローカルな事情をAIにいかに取り込み、AIをうまく活用して合意形成へと導くかが、今後の重要な研究課題である。これは容易な問題ではない。しかし、必ずしも新しいものばかりではないにせよ、外部知識の参照や対話を通じた知識の獲得など、さまざまな技術が登場してきている。それらを組

み合わせることによって、この限界は克服できると筆者は考えている。重要なのは、克服の方向がAIにローカル知識を「持たせる」ことではなく、人間が持ち寄るローカル知識をAIが「受け取り、活かす」ことにあるという点である。

5. マルチエージェント系の統治：ガードレールの限界とメカニズムデザイン

5.1 単一エージェントの制御：ルールベースからデータ駆動へ

AIエージェントの暴走を防ぐために、人間が事前に禁止事項を指示し、要所にチェックポイントを設けるというアプローチが広く採られている。しかし、チェックポイントを設けたり禁止リストを作成したりしても、起こりうる事態のすべてを予見することはできないため、すべてを停止させることはできない。この限界は、AIの構成原理の転換に由来する。

かつてのAIはルールベースであった。すなわち、やってよいことだけをコンピュータに教え、やってよいことだけを出力させようとするアプローチである。このアプローチにおいては、システムの振る舞いは設計者が記述したルールの範囲内に原理的に収まる。これに対して現在のAIは、ありうるすべてのデータからルールを抽出する。その結果、やってはいけないこともまた、学習されたルールの中に必然的に含まれてしまう。研究者が現在行っているのは、特定の目標に対して適切に動作するよう強化学習を施し、望ましくないルールを採用しないように調整することである。しかしそれでも、望ましくない振る舞いを全て排除することはできない。これに対して講じられているのがガードレールであり、逸脱した行動を取らないようにできる限りの制約を課すことが行われている。これが、単一のエージェントの振る舞いを矯正することに関する現状である。

5.2 善意の目標が悪い帰結を生む構造：エージェント間の相互依存

筆者がより深刻な問題と考えるのは、単一エージェントの制御を超えた、複数エージェント間の相互依存から生じる問題である。善意で設定された目標が、最終的に我々にとって良くない結果を生む事例は数多く存在する [21][27]。

その構造は次のように定式化できる。あるエージェントには善意であるタスクを指示し、別のエージェントにも善意で別のタスクを指示する。しかしそれらの行動が互いに依存していたり、一方のエージェントの行動の結果が他方のエージェントの行動の前提となっていたりする場合、両者とも善意で設計されたにもかかわらず、最終的に悪い結果が生じうる。

具体例を挙げる。あるエージェントは、毎日株価の情報をSNSに投稿している。これ自体は単に情報を提供しているだけであり、悪意はない。別のエージェントは、SNSの中から株価情報を探し出し、自身のユーザーに対して特定の株を買うべきだと勧める。ここで後者のエージェントが、前者の投稿の中から質の悪い情報を拾ってしまう可能性がある。両者の行動の間には依存関係があり、一方の行動と他方の行動が組み合わさったときに何が起こるかは、最終的には予測できない。これは複雑系の問題である。

この構造が意味するのは、個々のエージェントをいくら善意で設計し、いくらガードレールを強化しても、系全体の帰結を制御できないということである。単一エージェントのレベルでの対策は必要ではあるが、十分ではない。

5.3 メカニズムデザインの考え方

このような状況に対する原理的なアプローチとして筆者が提示するのが、メカニズムデザインの考え方である。これはミクロ経済学において数理的に分析されてきた理論体系である [24][25][26]。その基本的な発想は、エージェントに対して仕組み、すなわち制度を与えることにある。ただし単に仕組みを与えるだけではなく、その仕組みに従って正直に行動することが各エージェントにとって損にならないように設計する。そのようなメカニズムを構築しておけば、各エージェントは正直に行動するしかなくなる。

ここには、ガードレールとは質的に異なる発想の転換がある。ガードレールは、エージェントの行動を外部から禁止・監視する。これに対してメカニズムデザインは、正直な行動が各エージェント自身の利益と一致するように制度を設計する。前者が制約の網羅性に依存するのに対し、後者はインセンティブの整合性に依拠する。すべての逸脱を予見できないという前述の限界は、前者にとっては致命的であるが、後者にとっては必ずしも致命的ではない。

この発想に基づく理論的に美しい仕組みは、オークションや、お見合いのようなマッチングをはじめとして、すでに数多く構築されている [11][13][14]。筆者らも、相互依存価値オークションや組合せオークションにおける誠実性を保証するメカニズム [11][23] や、非線形効用空間における多属性交渉プロトコル [12][15][16] の研究を通じて、この理論体系の構築に関わってきた。

5.4 残された課題：密な依存関係を持つマルチエージェント系の分析

しかし、既存のメカニズムデザインがそのまま適用できるわけではない。人間社会に見られるような、前述の依存関係が複雑に絡み合った状況は、まだ分析し尽くされていない。従来のメカニズムデザインが扱ってきたのは、比較的単純な構造を持つ状況、すなわちエージェントの行動空間と相互作用の構造が明確に定義された状況である。

この課題が今日的な緊急性を持つ理由は、複数のエージェントが人間よりも速く判断を下し、かつそれらの間の依存関係が極めて密であるという点にある。人間の意思決定が主体であった時代には、意思決定の速度と結合の密度が限られていたため、そこまでの分析を行わなくても支障がなかったかもしれない。しかしAIエージェントが意思決定の主体となる時代においては、この分析を行わなければならない。高速かつ密に結合したマルチエージェント系へとメカニズムデザインを拡張することが、筆者が現在最も重要と考える研究課題である [21][22]。

6. 不確実性下における技術との向き合い方

6.1 予測不可能性の承認

前節で論じたマルチエージェント系の統治は、AIのリスク制御というより広い問題の一部である [9] [10]。この問題について、筆者は技術者としての立場から、未来に何が起こるかを予測できる者は誰もいないという認識から出発する。この認識は、主要なAI開発企業の経営者が表明しているものと同様である。具体的には、現在のAIの発展水準を三ヶ月前に予測できた者はほとんどいなかった。したがって、今後三ヶ月の間に何が起こるかも同様に予測できない。AIをめぐる状況は極めて高い不確実性の下にあり、注意を払って対応すべきであるというのが、まず第一に言えることである。

6.2 核技術との類比

ただし、不確実性の承認が直ちに悲観論に結びつくわけではない。現在の状況は、原子爆弾をめぐるオッペンハイマーの状況と構造的に類似している。オッペンハイマー自身、そして当時の科学者たちもまた、技術がもたらす帰結を予見できないことを自覚し、注意が必要であることを訴えていた。しかし最終的には、人類の努力によって最悪の事態は回避されてきた。

6.3 準備と活用の両立

この類比から導かれるのは、楽観的に考えるならば、人類はリスクを乗り越えるための準備を整えた上で、さらにその技術を有効に活用すべきである、という立場である。不確実性の承認と注意深さの要請は、技術の放棄を意味するのではなく、準備と活用を両立させる姿勢を要請する。前節で述べたメカニズムデザインの拡張は、この「準備」の中核をなすものであり、第8節で述べる技術の体験的普及と教育は、「活用」の側面を担う。この両輪を同時に進めることが、不確実性下における研究者の責任であると筆者は考える。

7. 身体性・有限性と挑戦：人間に固有なもの

7.1 人間力をめぐる問い

ここまでの議論は、社会的意思決定とエージェント統治におけるAIと人間の役割を扱ってきた。本節からは、より根本的な問い、すなわちAIが知的作業の多くを代替するとき、人間に固有に残るものは何かを考察する。筆者は、人間力とは「何を目標とし、何に挑むかを選ぶ力」であると考えている。この力は内発的動機と呼ばれるものに近い。しかし、AIが進展し、サービス産業がニーズに細かく応える状況では、そのような内的な動機が湧きにくくなっているという懸念もある。では、内発的動機はどのような条件の下で生じるのか。

7.2 知能における身体性論争

この問いを考える手がかりとして、身体性をめぐるAI研究の論争を参照する。AI研究者の見解は、この問題に関してほぼ二分されている [28][29]。一方には、知能は身体があって初めて成立すると考える立場がある。他方には、身体は知能にとって本質的ではないと考える立場がある。筆者は、知能そのものに関しては、後者の立場にも一定の説得力があると考える。現在の大規模言語モデルが示す知的能力は、身体を持たないシステムにおいて実現されているからである。

7.3 興味と挑戦における身体性・有限性の役割

しかし、知能とは区別される「興味」や「挑戦」に関しては、事情が異なる。筆者は、これらには身体が深く関与している可能性があると考える。ヒマラヤ登山を例にとる。それは、自らの命を危険にさらしてまで、周囲からは合理的とは言えないことを行う行為である。当事者にとっては合理的なのであろうが、そのような行為の動機には、身体が強く関わっていると考えられる。

より一般化すれば、身体、生命、そして限りある人生といった要素が、興味や挑戦に深く関係している。人間が何に興味を持つかということ自体が、実は身体性、すなわち自分に与えられた制約によって規定されている。この制約には二つの側面がある。一つは身体という制約であり、もう一つは時間という制約である。これらの制約が、内的な興味と挑戦を生み出す。

この論点は、思考実験的な対偶によって補強できる。もし永遠に生き続けることができ、不死身の身体を持っているならば、おそらくそれほど挑戦する必要はない。挑戦という行為の動機は、有限性の自覚によって支えられている。この観点からすれば、子どもが単純な遊びを繰り返しながら自らの興味を見つけていく過程に身体知が関わっているという観察は、身体と時間という二重の制約の下で内発的動機が形成されることの一例として理解できる。

7.4 含意

この考察から導かれるのは、AIが知的作業の多くを代替しうる時代において、人間に固有のものとして残るのは、有限な身体と有限な時間を持つ存在として、何に挑むかを選ぶことである、という命題である。知能が身体を必要としないとしても、挑戦の選択は身体と有限性を必要とする。この命題は、次節で論じる教育のあり方に直接の含意を持つ。

8. AI時代の人間の役割と教育：対面、体験、クリエイティビティ

8.1 対面の価値は失われない

AIは、人と人とのコミュニケーションにおける摩擦を軽減し、言語の障壁を低減させる。筆者の研究室には、アフガニスタン、チュニジア、中国など多様な国の出身者が在籍しており、共通語は英語のみである。それぞれの母語で話すことができないかを試みたことがあるが容易ではなく、このような場面ではAIによる翻訳が極めて有効であり、その精度はすでに相当に高い。

しかし、AIによるコミュニケーションの円滑化がどれほど進んでも、対面による関係構築の価値は失われない。この点について、筆者が近年観察している現象は示唆的である。コロナ禍からしばらくの間、学生は懇親会のような対面の交流の場にあまり来なかった。ところが最近では、学生の側から教員に対して、そうした場を設けてほしいと求めてくるようになった。この現象が重要なのは、現在の学生がAIを極めて頻繁に利用している世代だからである。AIを最も使いこなす世代が、それでもなお、対面で友人になり、互いに分かり合うことに価値を見出している。筆者はこれを、人間が根本的に持つ欲求の表れであると解釈する。前節の議論と照らし合わせれば、対面の交流は身体を伴う交流であり、身体を持たないAIが代替できない領域の一つである。

8.2 体験による技術受容

一方で、地方自治体や企業、とりわけ中小企業においては、AIの利用が限定的にとどまっている実態がある。これに対して筆者が実践しているのは、説得の言葉を探すことではなく、体験の機会を提供することである。生成AIエージェントを用いると、各人が抱えている課題を解決するソフトウェアがどれほど速く構築できるかを、実際に体験してもらう講演会やワークショップを各所で行っている。体験をすれば、従来との違いがこれほど大きいのかということが直ちに理解される。論理的な説得によってではなく直接的な経験によって認識が変容するということは、ここでも身体性を伴う経験が決定的な役割を果たしていることを示している。

なお、AI利用に伴う個人情報のリスクについては、主要なものとして二つを指摘しておく。第一は、海外にサーバーを持つサービスに個人情報を入力することが、自らの手で情報を国外に送信することに他ならないという点である。第二は、入力した情報がモデルの訓練データとして利用される可能性であり、これについてはオプトアウトの設定を行うことがまず講じるべき対策である。技術の体験的普及は、このような基本的なリスク理解と並行して進められなければならない。

8.3 創造の喜びの回復

ソフトウェア開発の領域では、AIの利用によって仕事が奪われるのではないかという懸念がある。しかし筆者自身の経験は、それとは異なる。生成AIを用いてコーディングを行えば、従来三ヶ月を要した作業が一日で完成する。筆者にとってこれは、人生で初めてコンピュータに触れたときの感激が、いま再び戻ってきたような経験である。自分の作りたいものを次々と作れるという感覚が、そこにはある。技術の進展が自動的に人を幸せにするわけではなく、それを動かすのはビジョンと規範の力であるという指摘 [30] は正しい。しかし、技術を通じて創造の喜びを取り戻すことはできる。

8.4 AIのクリエイティビティの上にあるクリエイティビティ

この経験から導かれる教育的含意は次のとおりである。AIという技術を適切に使いこなした上で、その上にクリエイティビティを発揮しなければならない。ここで筆者は、クリエイティビティを二つの層に区別する。一つはAIが発揮できるクリエイティビティであり、もう一つはそのさらに上にあるクリエイティビティである。後者の領域にこそ、より楽しいものがある。

この上位のクリエイティビティの領域が、まさに教育におけるビジョンの所在である。大学教育が伝えるべき価値は、ビジョンと挑戦である。第7節で論じたように、何に挑むかを選ぶ力は人間に固有のものであり、AIが知的作業の多くを代替する時代においてこそその重要性を増す。教育の役割は、AIが到達できる水準を踏まえた上で、その先に何を目指すかを自ら選び、そこに挑む力を育てることにある。

9. 結論

本稿では、生成AIとAIエージェントの発展が提起する二つの問い、すなわち社会的意思決定におけるAIの役割とその統治、および人間に固有に残るものについて、筆者の立場を論じてきた。その要点は以下の四点に集約される。

第一に、AIを介した合意形成の正当性は、理想的な基準ではなく現実に利用可能な代替手段との比較において評価されるべきであり、さらに制度設計そのものを市民の協働に開くことによって根本的に解決されうる。アフガニスタンにおけるRCTは、AIファシリテーターがインタラクションの促進と議論の規律化を通じて偏見を低減させうることを示唆している。

第二に、生成AIの構造的限界はローカル知識にあり、合意形成におけるAIの役割は、ローカル知識を代替することではなく、人間が持ち寄るローカル知識を引き出し構造化することにある。

第三に、AIエージェントの真のリスクは単一エージェントの逸脱ではなく、善意で設計された複数エージェントの相互依存から生じる創発的帰結にある。これに対する原理的なアプローチは、外部制約としてのガードレールではなく、インセンティブの整合性を保証するメカニズムデザインであり、それを高速かつ密に結合したマルチエージェント系へと拡張することが現在の研究課題である。AIの未来は予測不可能であるが、人類は準備を整えた上で技術を活用すべきである。

第四に、AIが知的作業を代替する時代において人間に固有なものは、有限な身体と有限な時間という制約に根ざした「何に挑むかを選ぶ力」である。対面の交流や体験による認識の変容が失われない価値を持つのは、それらが身体性を伴うからである。教育の役割は、AIが到達できるクリエイティビティの水準を踏まえた上で、その上にあるクリエイティビティへと学生を導くことにある。

これら四点を貫くのは、AIの能力と限界を現実的に評価した上で、人間の役割を「制度を設計すること」「ローカルな知識を持ち寄ること」「何に挑むかを選ぶこと」に再定義するという姿勢である。この姿勢は、技術楽観論でも技術悲観論でもなく、技術を前提とした上で人間が担うべき役割を積極的に再定義しようとするものである。マルチエージェントシステムとメカニズムデザインの研究は、この再定義を支える理論的基盤を提供するものとして、今後いっそう重要になる。

参考文献

筆者らによる関連研究

[1] T. Ito, R. Hadfi, S. Suzuki, "An Agent that Facilitates Crowd Discussion: A Crowd Discussion Support System based on an Automated Facilitation Agent," *Group Decision and Negotiation*, 2021. https://doi.org/10.1007/s10726-021-09765-8 （GDN Journal Best Paper Award）

[2] T. Ito, "Discussion and Negotiation Support for Crowd-Scale Consensus," in D. M. Kilgour and C. Eden (Eds.), *Handbook of Group Decision and Negotiation*, Springer, 2021.

[3] R. Hadfi, J. Haqbeen, S. Sahab, T. Ito, "Argumentative Conversational Agents for Online Discussions," *Journal of Systems Science and Systems Engineering*, 2021. https://doi.org/10.1007/s11518-021-5497-1

[4] S. Sahab, J. Haqbeen, R. Hadfi, T. Ito, R. Imade, S. Ohnuma, T. Hasegawa, "E-contact Facilitated by Conversational Agents Reduces Interethnic Prejudice and Anxiety in Afghanistan," *Communications Psychology*, Vol. 2, Article 22, 2024. https://doi.org/10.1038/s44271-024-00070-z

[5] R. Hadfi, S. Okuhara, J. Haqbeen, S. Sahab, S. Ohnuma, T. Ito, "Conversational Agents Enhance Women's Contribution in Online Debates," *Scientific Reports*, 2023. https://doi.org/10.1038/s41598-023-41703-3

[6] J. Haqbeen, S. Sahab, T. Ito, P. Rizzi, "Using Decision Support System to Enable Crowd Identify Neighborhood Issues and Its Solutions for Policy Makers: An Online Experiment at Kabul Municipal Level," *Sustainability*, Vol. 13, 5453, 2021. https://doi.org/10.3390/su13105453

[7] Y. Dong, S. Ding, T. Ito, "An Automated Multi-phase Facilitation Agent based on LLM," *IEICE Transactions on Information and Systems*, 2023. https://doi.org/10.1587/transinf.2023ihp0011

[8] R. Hadfi, T. Ito, "Structural Complexity Predicts Consensus Readability in Online Discussions," *Social Network Analysis and Mining*, Springer, 2024. https://doi.org/10.1007/s13278-024-01212-1

[9] Y. Ikeda, R. Hadfi, T. Ito, A. Fujihara, "Anomaly Detection and Facilitation AI to Empower Decentralized Autonomous Organizations for Secure Crypto-Asset Transactions," *AI & Society*, 2025. https://doi.org/10.1007/s00146-024-02166-w

[10] R. Hadfi, T. Ito, S. Ohnuma, S. Shiramatsu, "Guest Editorial: Special Issue/Collection on AI and Democracy: Transformative Impacts, Challenges, and Solutions," 2024.

[11] T. Ito, D. C. Parkes, "Instantiating the Contingent Bids Model of Truthful Interdependent Value Auctions," *Proc. 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006)*, pp. 1151–1158, 2006. https://doi.org/10.1145/1160633.1160842 （Best Paper Award）

[12] T. Ito, M. Klein, H. Hattori, "Multi-issue Negotiation Protocol for Agents: Exploring Nonlinear Utility Spaces," *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pp. 1347–1352, 2007.

[13] M. Yokoo, T. Ito, T. Matsuo, M. Zhang, J. Lee (Eds.), *Electronic Commerce: Mechanism Design and Applications*, Springer, 2008.

[14] K. Hara, T. Ito, "A Scoring Rule-based Truthful Demand Response Mechanism," *International Journal of Networked and Distributed Computing*, Vol. 4, No. 3, pp. 182–192, 2016.

[15] T. Baarslag, K. Fujita, E. Gerding, K. Hindriks, T. Ito, N. R. Jennings, C. Jonker, S. Kraus, R. Lin, V. Robu, C. Williams, "Evaluating Practical Negotiating Agents: Results and Analysis of the 2011 International Competition," *Artificial Intelligence*, Vol. 198, pp. 73–103, 2013. https://doi.org/10.1016/j.artint.2012.09.004

[16] R. Hadfi, R. Aydoğan, T. Ito, R. Arisaka (Eds.), *Recent Advances in Agent-Based Negotiation: Applications and Competition Challenges*, Studies in Computational Intelligence, Springer, 2023. https://doi.org/10.1007/978-981-99-0561-4

[17] T. Ito, Y. Dong, J. Haqbeen, T. Matsuo, S. Sahab, "Hyperdemocracy: Towards Creative Consensus Building between Humans and AI," *Proc. International Conference on Agentic AI (ICA 2025)*, 2025. https://doi.org/10.1109/ica67499.2025.00034

[18] T. Ito, Y. Dong, J. Haqbeen, S. Sahab, T. Matsuo, "Multiple AI Agents that Support Crowd Discussion," *Proc. ACM Collective Intelligence Conference (ACM CI 2024)*, 2024.

[19] S. Ding, T. Ito, "A Deep Reinforcement Learning Based Facilitation Agent for Consensus Building among Multi-Round Discussions," *Proc. 20th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2023)*, 2023. https://doi.org/10.1007/978-981-99-7025-4_23

[20] S. Sahab, J. Haqbeen, T. Ito, "AI Facilitation and Affective Dynamics of Online Deliberation," *Discover Psychology*, 2026. https://doi.org/10.1007/s44202-026-00976-1

[21] Z. Wu, T. Ito, "The Hidden Strength of Disagreement: Unraveling the Consensus-Diversity Tradeoff in Adaptive Multi-Agent Systems," *Proc. 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, 2025. https://doi.org/10.18653/v1/2025.emnlp-main.772

[22] Y. Dong, T. Ito, "From Consensus Theory to LLM Agents: Practical Consensus-Building for Multi-Issue Negotiation," *Expert Systems with Applications*, 2026（採録）. https://doi.org/10.1016/j.eswa.2026.132250

[23] T. Ito, M. Yokoo, S. Matsubara, A. Iwasaki, "A New Strategyproof Greedy-Allocation Combinatorial Auction Protocol and its Extension to Open Ascending Auction Protocol," *Proc. 20th National Conference on Artificial Intelligence (AAAI 2005)*, pp. 261–266, 2005.

一般文献

[24] L. Hurwicz, "The Design of Mechanisms for Resource Allocation," *American Economic Review*, Vol. 63, No. 2, pp. 1–30, 1973.

[25] R. B. Myerson, "Optimal Auction Design," *Mathematics of Operations Research*, Vol. 6, No. 1, pp. 58–73, 1981.

[26] N. Nisan, T. Roughgarden, É. Tardos, V. V. Vazirani (Eds.), *Algorithmic Game Theory*, Cambridge University Press, 2007.

[27] N. Bostrom, "Ethical Issues in Advanced Artificial Intelligence," in I. Smit et al. (Eds.), *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2, pp. 12–17, 2003.

[28] R. A. Brooks, "Intelligence without Representation," *Artificial Intelligence*, Vol. 47, pp. 139–159, 1991.

[29] R. Pfeifer, J. Bongard, *How the Body Shapes the Way We Think: A New View of Intelligence*, MIT Press, 2007.

[30] D. Acemoglu, S. Johnson, *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*, PublicAffairs, 2023.

[1]〜[23] の書誌情報は https://takayuki-ito.com/publications.html に基づく（[16] のみ CV に基づく）。