2.1.5 人・AI協働と意思決定支援

（1）研究開発領域の定義

「人・AI協働」は、何らかの目的達成に向けて、人とAIが協力して取り組むことを指す。国際規格であるISO/IEC 22989：2022「Artificial intelligence concepts and terminology」において、Human-Machine Teaming（HMT）という概念が「Integration of human interaction with machine intelligence capabilities」と定義されており、これが「人・AI協働」とほぼ同義である。 Human-in-the-Loopと呼ばれる概念もHMTに含まれる。 HMTは、人（Human）とAI（Machine）の上下関係に応じて五つのパターンに整理される[1)]。人が上位となるタイプからAIが上位となるタイプの順に、Human Supervisor/User、Human Mentor、Peer、Machine Mentor、Machine Supervisorと呼ばれる（図2-1-8左）。

「意思決定」は、個人や集団がある目標を達成するために、考えられる複数の選択肢の中から一つを選択する行為である。その選択では個人の価値観がよりどころとなるが、集団の意思決定では、必ずしも関係者（メンバーやステークホルダー）全員の価値観が一致するとは限らない。関係者内で選択肢に関する意見が分かれたとき、その一致を図るためには「合意形成」も必要になる。近年、情報氾濫による可能性の見落としやフェイク生成などを用いた情報操作といった問題が顕在化し、意思決定ミスを起こすリスクが高まっている。このような問題・リスクを軽減するため、AI技術を活用した「意思決定支援」が期待されている[2), 3)]。これはHMTの五つパターンのうち、主にHuman Supervisor/UserやHuman Mentorに該当する（Humanが意思決定者）。

本領域は、「人・AI協働」のための、より良い枠組みと、そこで必要とされる技術を開発する研究開発領域である。その中でも特に「意思決定支援」のためのAI技術活用を重点的に取り上げる。

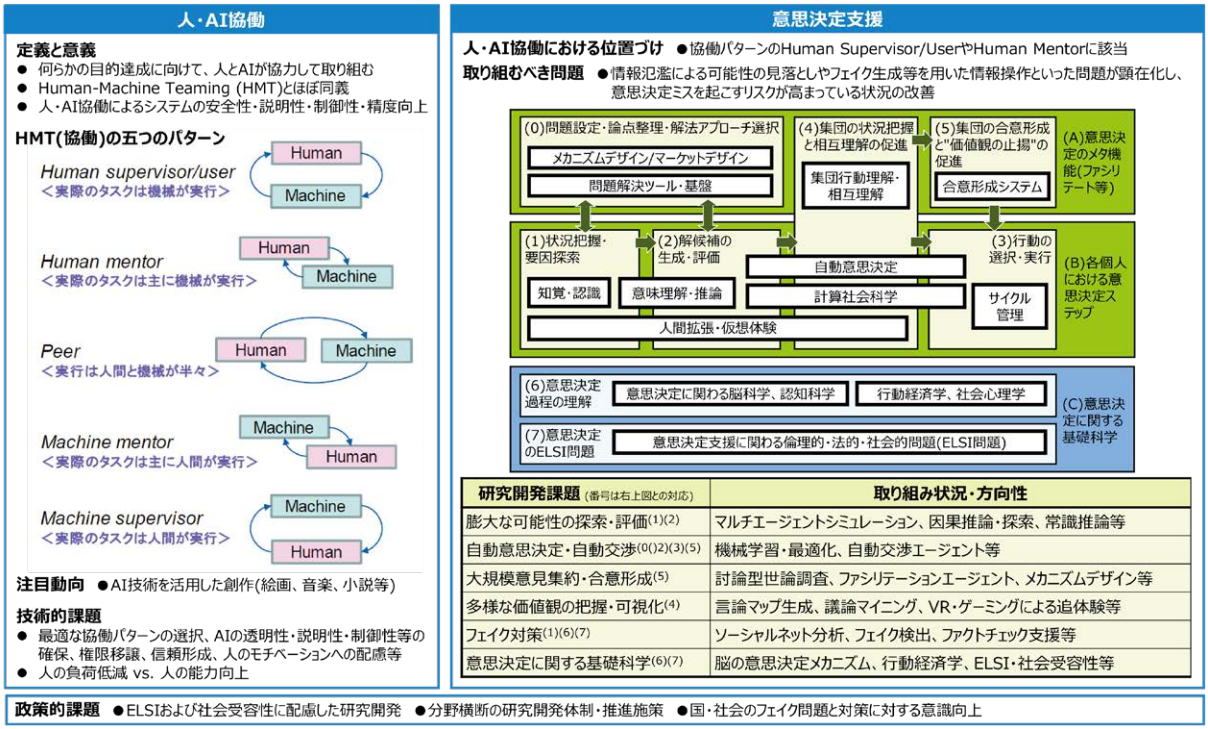

図2-1-8　　領域俯瞰：人・AI協働と意思決定支援

（2）キーワード

Human-Machine Teaming、Human-in-the-Loop、意思決定、合意形成、意見集約、フェイクニュース、フェイク動画、デジタルゲリマンダー、インフォデミック、議論マイニング、マルチエージェント、自動交渉、計算社会科学、行動経済学、処方的分析

（3）研究開発領域の概要

［本領域の意義］

われわれは日々さまざまな場面で意思決定を行っている。クリティカルな場面での意思決定ミスは個人や集団の状況を悪化させ、その存続・生存さえも危うくする。例えば、企業の経営における意思決定ミスは、企業の業績悪化・競争力低下を招き、国の政策決定・制度設計における意思決定ミスは、国の経済停滞や国民の生活悪化にもつながる。また、個人の意思決定における判断スキル・熟慮の不足は、その個人の生活におけるさまざまなリスクを誘発するだけでなく、世論形成・投票などにおける集団浅慮という形で、社会の方向性さえも左右する。

情報技術が発展し、社会に浸透した今日、情報の拡散スピードが速く、膨大な情報があふれ、影響を及ぼし合う範囲が思わぬところまで広がっている。そのような意思決定の行為自体の難しさが増していることに加えて、意思決定の際のよりどころとなる価値観の多様化[4]によって、合意形成の難しさも増している。さらには、価値観の対立から悪意・扇動意図を持った情報操作（フェイクニュース、フェイク動画など）まで行われるという問題も顕在化し[5), 6), 7), 61)]、社会問題化している。さらに、それが国家の意思決定を誤らせたり、人々を混乱・対立させたりといった目的に使われるケースも起きており、新種のサイバー攻撃ともみなされる。さらに2020年に世界を一変させたCOVID-19パンデミックでは、インフォデミック[1]による社会混乱も発生した。

このような意思決定の困難化（意思決定ミスを起こすリスクの増大）という状況に対して、AI技術を活用することで、意思決定におけるさまざまな選択肢の探索や吟味を行いやすくしたり、悪意・扇動意図を持った情報操作に惑わされにくくしたりといった対策が考えられている。これによって、問題のすべてを解決できるわけではなくとも、リスクを軽減し、状況を改善する手段になり得る。新種のサイバー攻撃やインフォデミックによる社会混乱への対策としての意義も高まりつつある。

以上では、特に人が主体となったAI技術活用による意思決定支援という面について述べたが、HMTのバリエーションとして、逆にAIによってフルに自動化されたプロセスに対して、人が参加すること（Human-in-the-Loop）の意義・効果についても述べる。 AIが十分学習できていないケースや苦手なケースについて、人が参画することで、システムの安全性が確保されたり、人（特に専門家）からのフィードバックを通してAIの精度が向上したりといった効果が期待されている。また、AIによるフル自動化では結果の説明や制御が難しいという問題が指摘されているが、人が参画することで、説明性や制御性も改善され得る。

［研究開発の動向］

❶ 意思決定問題への取り組み

個人・集団の意思決定問題は古くから検討されてきた問題である。意思決定に関する先駆的な研究としては、1978年にノーベル経済学賞を受賞したHerbert A. Simonの取り組み[8)]がよく知られている。Simonは意思決定プロセスを、（1）情報（Intelligence）活動、（2）設計（Design）活動、（3）選択（Choice）活動というステップで構成されるとした。（1）で意思決定に必要な情報を収集し、（2）で考え

1 インフォデミック（Infodemic）は「情報の急速な伝染（Information Epidemic）」を短縮した造語で、正しい情報と不確かな情報が混じり合い、人々の不安や恐怖をあおる形で増幅・拡散され、信頼すべき情報が見つけにくくなるある種の混乱状態を意味する。

られる選択肢を挙げ、（3）で選択肢を評価し、どれを選択するか決定する。これらのステップにおいて、必要な情報をすべて集めることができ、可能性のあるすべての選択肢を挙げることができ、各選択肢を選んだときに起こり得るすべての可能性を列挙して評価することができるならば、合理的に最良の選択が可能になる。しかし、現実にはそのようなすべての可能性を考えて意思決定することはできず、人が合理的な意思決定をしようとしても限界がある。このSimonが導入した「限定合理性」（Bounded Rationality）と言う概念は、意思決定に関する研究発展の基礎となった。 Simonは、経営の本質は意思決定だと考え、限定合理性を克服するための組織論も展開した。

そのように人の判断・行動が必ずしも合理的になり得ず、心理・感情にも左右されるものであることを踏まえて、行動経済学が発展し、その中では意思決定に関わる興味深い知見が示されている。特に有名なのは、Simonの後、行動経済学の分野でノーベル経済学賞を受賞した2人、Daniel Kahneman（2002年受賞）とRichard H. Thaler（2017年受賞）の研究である。 Kahnemanは、直観的な「速い思考」のシステム1と論理的な「遅い思考」のシステム2から成るという二重過程モデル[9]や、人は利得面よりも損失面を過大に受け止めがちだといったプロスペクト理論[10]を提唱し、Thalerは、軽く押してやることで行動を促す「ナッジ」（Nudge）という考え方[11]を提唱した。

また、脳科学分野における脳の意思決定メカニズムの研究も進んでいる（詳細は「2.1.7 計算脳科学」を参照）。ドーパミン神経細胞の報酬予測誤差仮説などが見いだされ、モデルフリーシステムによる潜在的な意思決定と、モデルベースシステムによる顕在的な意思決定が協調および競合しつつ、人の意思決定が動作していることが分かってきた[12]。モデルフリーシステムは、事象と報酬との関係を直接経験に基づき確率的に結び付ける。モデルベースシステムは、事象と報酬との関係を内部モデルとして構築し、直接経験していないケースについても予測を可能にする。このような2通りのシステムは二重過程モデルとも整合しており、意思決定が合理性だけによるものではないことの裏付けにもなる。

このような人文・社会科学分野や脳科学分野における意思決定に関する研究が、主に人の側から掘り下げられてきた一方で、近年の情報技術の発展、Webやソーシャルメディアの普及は、意思決定を行う人の環境を大きく変化させた。その結果、意思決定問題は新たな様相を呈するようになり、以前とは異なる困難さが生じている。今日、意思決定問題は情報技術との関わりが大きなものになっている。

❷ 意思決定問題の新たな様相・困難さ

新たに生じている困難さを示す事象（問題）として顕著なものを四つ挙げる。

一つ目は、クリティカルな要因・影響の見落としの問題である。例えば、グローバル化したビジネス競争環境において、世界のあらゆる地域、思ってもいなかった業種から新たな競合が生まれ、想定していなかった法規制やソーシャルメディアで思わぬ切り口からの炎上も起こり得る。膨大な情報があふれ、社会がボーダーレス化した今日、意思決定に関連しそうな要因や意思決定結果の影響に膨大な可能性が生じ、人の頭でそのあらゆる可能性をあらかじめ考えるのは極めて難しい。 Simonのいう限定合理性が極度に進み、問題として深刻化している状況である。

二つ目は、ソーシャルメディアによる思考誘導の問題である。Webやソーシャルメディアを用いた情報発信・交流が広がり、それが人々の意思決定や世論形成に与える影響は無視できないものになっている[13), 14)]。2016年の米国大統領選挙はその顕著な事例であり[5), 6)]、SNS（Social Networking Service）などのソーシャルメディアを用いた政治操作は「デジタルゲリマンダー」と呼ばれ[15)]、フェイクニュースが社会問題化した。 SNSでは、価値観が自分に近い相手としかつながらず、自分の価値観に沿った情報しか見ない、いわゆる「フィルターバブル」状態[16]に陥りやすいことも、SNSが思考誘導の道具になりやすい原因になっている。

三つ目は、価値観の対立激化、社会の分極化の問題である。集団の合意形成に難航し、対立が激化する傾向が強まっている。価値観の対立は古くから起こってきた事象だが、社会のボーダーレス化に伴う関係

者範囲の広がりや、SNSでの同調圧力やエコーチェンバー現象による意見同質集団の形成強化が、対立を強め、社会の分極化（Polarization）や政治的分断と言われる事態も引き起こされている[17), 18)]。

四つ目は、まるで本物のようなフェイク動画・画像の流通の問題である[61)]。前述のフェイクニュースは言葉（SNSテキストなど）で伝達されるものが主であったが、深層生成モデル（詳細は「2.1.1 知覚・運動系のAI技術」参照）によって、まるで本物のように見えるフェイク動画やフェイク画像が簡単に作れてしまうようになった（Deepfakesなど）[19)]。特にフェイク動画は本物だと信じ込まれやすく、政治家や有名人の架空の発言・行為などを作るためにこれが悪用され、社会に流通すると、何が真実で何かフェイクか、真偽判断を見誤るリスクが増大し、さまざまな混乱が生まれると危惧される[17), 20)]。さらに2020年には、まるで人が書いたかのような自然なフェイク文章を生成することができるGPT-3[21)]というシステムも登場した。さらに2022年11月にはGPT-3.5をベースとして対話にチューニングされたChatGPTが一般に公開された。ChatGPTは、自然言語対話という分かりやすいインターフェースで、さまざまなタスクにまるで人間の専門家のような自然かつ詳細な応答をするので、大きな話題になっているが、もっともらしい応答に虚偽が含まれることが多々あることから、人々に虚偽を教えたり、悪用されたりすることが、強く懸念されている。

以上の問題に見られるように、（1）意思決定に関わる要因や意思決定結果の影響に、膨大な可能性が生じるようになってしまったこと、（2）悪意・扇動意図を持った、他者の意思決定に作用する情報操作が容易になってしまったことが、意思決定の困難化の原因として顕著である。

❸ 意思決定支援のための技術群

図2-1-8の右上部に、個人・集団の意思決定プロセス（合意形成を含む）に対応させて、関連する技術群を示した。Simonの3ステップに相当する（B）各個人における意思決定ステップを中心に、（A）意思決定のメタ機能と（C）意思決定に関する基礎科学を上下に配置した3層構造で技術群を整理した[2), 3)]。以下、これらを六つの技術群に分けて、取り組みの現状と今後の方向性について述べる[20), 22)]。

a. 膨大な可能性の探索・評価

上記❷に示した原因への対策としてまず求められるのは、意思決定に関わる要因や意思決定結果の影響における膨大な可能性を探索し、それらの組み合わせの中から目的に合うものを評価して絞り込む技術である。マルチエージェントシミュレーションによるWhat-If分析（「2.1.3 エージェント技術」参照）、統計的因果推論による選択肢評価・反事実的予測（「2.2.1 因果推論」参照）、自然言語処理による因果関係探索[23)]などの研究開発が進められている。自然言語処理による因果関係探索を用いたシステムの例としては、情報通信研究機構（National Institute of Information and Communications Technology：NICT）で開発された、「なぜ？」「どうなる？」などの因果関係に関する質問応答を扱うことができるWISDOMX[24)]が挙げられる。しかし、さまざまな分野・文脈で推論が行えるようにするには、常識を含め推論に必要な知識の獲得や、推論が成立する前提条件の精緻化など、取り組まなくてはならない技術課題がまだ多く残されている。

b. 自動意思決定・自動交渉

米国Gartner社は、データ分析の発展を記述的分析（Descriptive：何が起きたか）、診断的分析（Diagnostic：なぜそれが起きたか）、予測的分析（Predictive：これから何が起きるか）、処方的分析（Prescriptive：何をすべきか）という4段階で自動化が進むとし、4段階目の処方的分析は「意思決定支援」と「自動意思決定」という2通りがあるとしている[25)]。この段階が進むほど、データ分析の顧客価値が高く、ビジネス上の競争も処方的分析へと進みつつある。「自動意思決定」はデータ分析の結果に基づき、何をすべきかというアクションまで自動決定するものであり、「意思決定支援」はアクションの候補を人に提示し、どんなアクションを実行するかは最終的に人が決定するものである。一見すると、意思決定支援よりも自動

意思決定の方が、より発展したものであるかのように思えるが、現状、意思決定問題の性質が異なると考えるのが適切である。すなわち、コスト、精度、速度、売り上げなどのような明確な指標（いわば価値観に相当）が定められ、それを評価関数・効用関数として合理的に解が一つ定められる意思決定問題は、機械学習・最適化などのAI技術を用いて「自動意思決定」が可能になる。それに対して、さまざまな価値観が混在している状況下、あるいは、価値観が不確かな状況下での意思決定問題は、最終決定に人が関わる「意思決定支援」の形が基本になる。これに関しては、人参加型（Human-in-the-Loop）のAI・機械学習が考えられている[26)]。

自動意思決定には、強化学習や予測型意思決定最適化など、機械学習・最適化技術をベースとした方式が開発・適用されている。強化学習（Reinforcement Learning）[27)] は、学習主体が、ある状態で、ある行動を実行すると、ある報酬が得られるタイプの問題を扱う機械学習アルゴリズムである。将来的により多くの報酬が得られるように行動を選択する意思決定方策を、行動選択と報酬の受け取りを重ねながら学習していく。囲碁で世界トッププロに勝利したGoogle DeepMindのAlphaGo[28)] で使われたことがよく知られている。強化学習が適するのは、大量に試行錯誤することが可能な類いの意思決定問題である。一方、古典的なオペレーションズリサーチ（Operations Research：OR）で扱われているような類いの意思決定問題（例えば大規模システムの運用計画や小売業の商品価格設定戦略など）は、意思決定で失敗したときのダメージが大きく、大量の試行錯誤は難しい。このような類いの問題を、機械学習からの大量の予測出力（予測が当たるかは確率的）に基づくOR問題とみなした新しいアプローチが予測型意思決定最適化[29)] である。

さらに、集団の意思決定を、異なる価値観（効用関数）を持ったエージェント間の交渉として定式化した自動交渉技術の研究も注目される。自動交渉は、それぞれの効用関数（いわばそれぞれの価値観）を持った複数の知的エージェント（AIシステム）が相対する状況において、一定の交渉プロトコルに従ってうまく合意案を見つける技術である。ある問題について、複数のステークホルダーの間で対立したり、協調しようとしたりするとき、交渉プロトコルや効用関数を定めてエージェントに代行させて自動交渉を行うならば、人同士が交渉するよりも、その条件で考えられる最適な合意点に高速に到達できると期待される。2010年からは毎年、国際自動交渉エージェント競技会ANAC（Automated Negotiating Agents Competition）が開催されており[30)]、これを共通の場として技術発展が進んできた。マルチエージェントシステムの考え方がベースにあり、応用事例を含めて「2.1.3 エージェント技術」で取り上げている。

c. 大規模意見集約・合意形成

上述の自動交渉は異なる価値観を持つ者の間の勝負という面があり、集団の意見集約・合意形成を目的とするならば、建設的な議論の進め方や相手への共感による価値観の変化といった面、および、そこでのファシリテーターの役割[31)] が重要なものになる。

集団の意見集約・合意形成のために情報技術を活用するシステムは、古くはグループウェアやCSCW（Computer Supported Cooperative Work）の研究分野での取り組みが見られる。例えば、Issue（課題、論点）をベースに木構造の表現でまとめるファシリテーション技法であるIBIS（Issue Based Information Systems）法をグラフィカルに実現したgIBISという意思決定支援ツール[32)] がよく知られている。一方、政治学の分野では、あるテーマについて回答を得る前に回答者にグループ討論をしてもらう討論型世論調査（Deliberative Poll）[33)] が、熟議に基づく民主主義の方法論として有効だと認識されるようになった。

近年は集合知の収集・活用の学際的研究が進んでおり、米国マサチューセッツ工科大学（MIT）に2006年に設立されたMIT Center for Collective Intelligence（集合知研究センター：CCI）が注目される。インターネットを使った大規模な議論を、その論理構造の可視化によって支援するシステムDeliberatorium[34)] や、地球温暖化問題を取り上げて、解決プランを協議するシステムThe

ClimateCoLabなどのプロジェクトを進めている。さらに、CCIのトップであるThomas W. Maloneは、2018年の著書[35]で、人の集合知にAIとの協働を含めたSuperMindsの方向性を示した。

そのような方向に沿って伊藤孝行研究室[2]では、議論構造の可視化に加えて、エージェント技術によるファシリテーター機能を導入した大規模合意形成支援システムD-Agree[36]を開発し、大学発ベンチャーAgreeBit社も起業している。D-Agreeは、国内で名古屋市のタウンミーティングなどで社会実験適用の実績があるが（D-Agreeの前身Collagree[37]もその実績あり）、さらに海外（アフガニスタンのカブール市など）にも展開されている。

d. 多様な価値観の把握・可視化

多様な価値観が混在する状況下での意思決定・合意形成に向けては、その状況や価値観の違いを可視化する技術が有効である。賛成・反対の各立場から意見と根拠を対比する言論マップ生成[38]、主張・事実などへの言明とその間の関係（根拠・支持、反論・批判など）を推定する議論マイニング（Argumentation Mining）[39]、議題に対して賛成・反対の立場でディベートを展開するシステム（IBMの「Project Debater」、日立の「ディベートAI」[40]など）が研究開発されている。より応用をフォーカスし、論理構築・推論を深める研究として法学AI[41]もある。これらは自然言語処理技術を用いた手法だが、集団の相互理解促進のためにはVR（Virtual Reality）技術やゲーミング手法を用いて相手の立場を追体験させるアプローチも効果がある。

Project Debaterは、2018年6月に米国サンフランシスコで開催されたイベントWatson Westにて、イスラエルの2016年度ディベートチャンピオンとライブ対戦[3]し、「政府支援の宇宙探査を実施すべきか否か」という議題で勝利して話題となった。ニュース記事や学術論文を3億件収集・構造化して用いており、2011年に米国のクイズ番組Jeopardy!で人のチャンピオンに勝利したIBM Watson[42]の自然言語処理に加えて、ナレッジグラフや議論マイニングなどの技術が組み合わせて実現されたものと考えられる。

e. フェイク対策

ソーシャルメディア上での情報伝播の傾向や、そこで起きている炎上、フェイクニュース、エコーチェンバー、二極化などの現象を把握・分析すること[5), 7), 14), 18), 43), 44)]は、フェイク対策のための基礎的研究となる。フェイクニュースへの対抗としては、発信された情報が客観的事実に基づくものなのかを調査し、その情報の正確さを評価・公表するファクトチェックという取り組みが立ち上がっている[5), 45)]。ファクトチェックを行う団体として比較的早期に立ち上がった米国のSnopes（1994年～）やPolitiFact（2007年～）がよく知られている。この動きは世界的に広がっており、日本では2017年にFactCheck Initiative Japan（FIJ）、2022年にJapan Fact-check Center（JFC）が発足した。また、NPO法人アイ・アジアが2019年にファクトチェック部門を開設し、2020年からNPO法人インファクトとして活動している。2015年には国際的に認証されたファクトチェック団体から成るInternational Fact-Checking Network（IFCN）が発足し、日本は対応が遅れていたが、2023年5月にインファクトとIFCNが加盟した。

ただし、大量に発信される情報を迅速にチェックするには人手では限界がある。そこで、コンピューター処理によってフェイクニュースの検出を効率化する試みが進められている（FIJでの取り組み[46]や2016年から始まった競技会Fake News Challengeなど）。また、フェイク動画・フェイク画像・フェイク音声などの判定については、オリジナルの動画・画像・音声から改ざんされていないか、当事者が実際に発話や行

2 2020年9月まで名古屋工業大学、10月から京都大学。

3 ライブ対戦の進行は、対戦する両者が議題の肯定派と否定派に分かれ、まず4分間ずつ主張を述べ、次に4分間ずつ相手の主張に対する反論を述べ、最後に2分間ずつまとめを述べるという形で行われ、その勝敗は聴衆の支持数で決まる。ディベートの議題は直前に与えられ、その場で相手の主張も踏まえつつ、自分の主張を組み立てることになる。

動をしていない虚偽の動画・画像・音声ではないか、といったことを動画・画像・音声の特徴分析によって判定することも行われている。フェイク検出技術の詳細は、［新展開・技術トピックス］で述べる。ただし、フェイクの検出技術とフェイクの作成法は往々にしていたちごっこになるため、技術開発だけでなく、メディアリテラシーの教育・訓練や、表現・言論の自由を損なわないように配慮しつつ法律・ルールの整備による対策も進めることが必要である[5), 14), 47)]。

f. 意思決定に関する基礎科学

情報技術によって人の意思決定を支援するにあたって、そもそも人の意思決定とはどういうものか、どうあるべきかを理解しておくことは重要である。既に言及した通り、行動経済学や脳科学の分野（「2.1.7 計算脳科学」を参照）で意思決定プロセスのモデルやメカニズムが研究されてきた。また、社会心理学・認知科学などの分野で研究されている確証バイアスを含む認知バイアス[48)]も意思決定に大きく関わる。加えて、人の意思決定・合意形成を支援する機能がELSI（Ethical, Legal and Social Issues：倫理的・法的・社会的課題）の視点から適切であるかについても常に考えておかねばならない。

❹ 人とAIの協働

人がある目的を達成するために、一部のタスクをコンピューターで実行するというのは、コンピューターが発明された頃から行われていたことだが、それはコンピューターでできることがごく限られた処理だけだったため、それ以外は人が対応するしかなかったということである。しかし、AIに代表されるように、今日コンピューターでできることは飛躍的に拡大し、人が行うよりも高速・高精度にさまざまなタスクを実行できるようになった。そこで、目的に応じて、人とAIシステム（あるいはAI技術が組み込まれたロボット）とでどのような役割分担を行うのが最適であるか、人とAI（Artificial Intelligence、Machine Intelligence）の最適協働の在り方としてHMT（Human-Machine Teaming）が考えられるようになった。本節の冒頭でHMTには五つのパターンがあることを示したが、［新展開・技術トピックス］❷では、それらの各パターンの内容や状況について述べる。

［論文や特許の動向］

研究開発領域別の論文・特許調査の結果において、本領域は、論文数も特許数も大きく増加している。2021年の論文数・論文数シェアとも国別では、米国が1位、中国が2位、英国が3位、ドイツとインドがほぼ同程度で続く。欧州は米国を若干上回っている。論文数の世界上位10機関中には、米国から4機関、英国から2機関が入っているほか、フランス、イラン、カナダ、中国などが入っていて、他領域と比べて多様性が見られる。本領域のホットトピックとしてフェイク問題が含まれており、国による社会的問題・関心の高まりの差が表れている可能性がある。

（4）注目動向

［新展開・技術トピックス］

❶ フェイク検出技術

フェイクニュース検出は次の四つの面から試みられている[44)]。一つ目は知識ベース検出方式で、従来の人手によるファクトチェックを強化するように、クラウドソーシング的な仕組みを使って専門家集団に検証してもらったり、あらかじめ蓄積された知識ベースと自動照合したりする取り組みがある。二つ目はスタイルベース検出方式で、誤解を生みやすい見出し表現や欺くことを意図したような言葉使いなどに着目する。三つ目は伝播ベース検出方式で、情報拡散のパターン（情報伝播のグラフ構造やスピードなど）に着目する。例えば、フェイクニュースは通常ニュースよりも速く遠くまで伝わる傾向があることが知られている。四つ目は情報源ベース検出方式で、ニュースの出典・情報源の信頼性やその拡散者の関係などから判断する。社

会環境・文脈などによって真偽の捉え方が変わるし、科学的発見によって真理の理解が変わることもあり、真偽が定められない言説も多いため、最終的には人による判断が不可欠だが、上に示したような技術は怪しいニュース・情報を迅速に絞り込むのに有効である。また、1件のニュース単独で真偽判定するよりも、複数の情報の間の関係比較や整合性判断、および、複数の視点からの複合的なチェックを行う方がより確かな判断が可能になる[49)]。

また、動画・画像・音声などがオリジナルから改ざんされていないか、当事者が実際に発話や行動をしていない虚偽の動画・画像・音声ではないか、といったことを動画・画像・音声の特徴分析、ニューラルネットワーク・機械学習を用いて判定したり、改ざんの箇所や方法を特定したりといった技術が開発されている[20), 50), 51), 61)]。例えば、不自然なまばたきの仕方、不自然な頭部の動きや目の色、映像から読み取れる人の脈拍数、映像のピクセル強度のわずかな変化、照明や影などの物理的特性の不自然さ、日付・時刻・場所と天気の整合などが手掛かりになる。フェイクの検出技術とフェイクの作成法はいたちごっこだとも言えるが、人の目・耳では見分けがつかないレベルのフェイク動画・画像・音声が作れてしまう事態において、コンピューターによる分析は不可欠である。さらに、動画・画像・音声の内容解析とは別に、ブロックチェーンを使って履歴を管理することで、改ざんが入り込むことを防ぐという方法もある。

フェイクは新種のサイバー攻撃として用いられ、社会混乱も引き起こすことから、安全保障上も対策が求められる。これに対して、米国は早い時期から研究投資を行っており、特に米国国防高等研究計画局（DARPA）は、画像・動画の改ざん検知を行うMedia Forensics（MediFor）プロジェクトや、その後継で、画像・動画の意味的不整合やフェイクの検知を行うSemantic Forensics（SemaFor）プロジェクトを推進している[4]。日本では、2020年度にJST CREST「信頼されるAIシステム」に採択された「インフォデミックを克服するソーシャル情報基盤技術」（CREST FakeMedia、研究代表者：越前功）が、フェイク問題に本格的に取り組むプロジェクトとして注目される[61)]。この活動を推進する拠点として、国立情報学研究所にシンセティックメディア国際研究センター（SynMedia Center、センター長：越前功）が設立された。

❷ HMT（Human-Machine Teaming）の五つのパターン

本節の冒頭で述べたHMTの五つのパターンの概要・状況を簡単に述べる[1)]。なお、Human（人）、Machine（AI）ともに単独のケースも複数のケースも考えられ、また、実際の問題では、複数のパターンが組み合わせられることもある。各パターンにおいて、人とAIが協働する中で、人・AIそれぞれの能力が高まって、関係性・パターンが変化していくこともある。

Human Supervisor/Userは、人がAIの上位に位置するパターンで、人が上司となるケース（Human Supervisor）と人が単なるユーザーのケース（Human User）がある。実際のタスクはAIによって実行される。医療画像診断などがHuman Supervisorのケースであり、人が責任を持って監督・介入するHuman Oversightが重要な課題である。その前提としてAIの透明性・説明性・制御性などが求められる。Human Userのケースでは、人がそこまで深く関与せず、一般にHuman Machine Interfaceが重要である。

Human Mentorは、人がAIのやや上位に位置するパターンで、AIが実際のタスクを主に実行し、人はそのタスクを実行しようと思えば実行できるものの、主としてMentorとして機械を指導する。問い合わせ

4 米国DARPAでは、悪意を持ったオンラインやオフラインでの誘導・干渉によって人々の思考や行動に影響を与える問題に対処するための取り組みをCognitive Securityと総称し、MediFor、SemaFor以外にも、個人情報・プライバシーが目的外に使われないように管理するBrandies、状況を理解し、アクションするため、さまざまな情報源の間の矛盾・整合を踏まえながら仮説を生成するAIDA（Active Interpretation of Disparate Alternatives）、人の心理的な隙や行動のミスにつけ込んで個人が持つ秘密情報を入手する攻撃（ソーシャルエンジニアリング）を検知・防御するASED（Active Social Engineering Defense）、悪意のあるボットネットワークや大規模マルウェアに対抗する自律ソフトウェアエージェントHACCS（Harnessing Autonomy for Counering Cyberadversary Systems）などのプロジェクトを実施している。

や検査などについて、AIで可能な範囲は自動処理して、難しいケースのみ人に対応させるというのが、その一例である。AIへの権限移譲や人へのエスカレーションの仕方が重要課題である。そのために、人・AI双方が相手のモデルを持つこと（Mutual Model）が必要だと言われる。

Peerは、人とAIが同格で、どちらもタスクを実行する能力を有している。ただし、条件によってどちらのパフォーマンスが優れているかが変わってくるため、状況に応じてどちらがタスクを実行するかを決める必要がある。人とAIがタスクを分担して並列に実行することもあり得る。自動運転のレベル3はPeerに該当する。権限移譲の管理やMutual Modelが重要である。特に、人がAIの能力を適切に把握していることが望ましく、過信や不信を避けるように信頼較正[52)]という手法が考えられている。

Machine Mentorは、人が主としてタスクを実行するものの、一部のタスクに関してはAIが実行する。自動運転のレベル1や2はこれに該当する。 AIが人の作業・行動をモニタリングしていて、危ない状況や不適切な状況が検知されたら、注意や助言を行うケースもこのパターンの一例である。人のモチベーションへの配慮が求められる。

Machine Supervisorは、人が専ら実際のタスクを実行し、上司の立場にあるAIはタスクの実行には携わらない。ライドシェアサービスUberが代表例である。また、クラウドソーシング[26)]をAIで最適管理するようなケースも該当する。人のモチベーションを考慮したタスクアサインやフィードバックが課題として挙げられる。

❸ AI技術を用いた創作

深層生成モデルの発展によって、芸術作品（絵画・音楽など）や文学作品（小説・俳句など）の創作へのAI技術の活用が広がった[53), 54), 55), 56)]。

まず絵画作品の生成では、2015年に、深層ニューラルネットワーク構造中の情報を操作することで、通常の画像を夢に出てくるような神秘的な画像に変換するDeepDreamが開発された。ゴッホやレンブラントらの絵画作品から画風を学習し、入力画像をそのような画風に変換するシステムなども開発され、深層生成モデルを学習したり、その内部情報を操作したりすることで、絵画的表現を生成したり変換したりする手法やツールが種々開発されるようになった。さらに、2021年に発表されたDALL-Eでは、簡単な文からそれを表現した画像が生成されるText-to-Imageが実現された。これに続いて、DALL-E2、Imagen、Perti、Museなど、より高品質な画像を生成できる技術が開発されるとともに、MidjourneyやStable Diffusionなど、文を入力するだけで簡単に使えて、まるでプロが書いたようなテイストの画像が生成できるシステムが、インターネット上で使える形で公開され、Text-to-Imageは2022年に大きな話題になった（深層生成モデルと画像生成AIの技術面については「2.1.1 知覚・運動系のAI技術」に記載している）。

音楽作品の生成（作曲）は、メロディー、リズム、コード進行などに関する音楽理論の蓄積があり、音楽理論・ルールをベースとした生成手法が古くから取り組まれていたことから、絵画作品の生成とは異なる発展を示している。深層学習を用いてバッハ風やビートルズ風といった新曲を作る取り組み、歌詞を入力として音楽理論やルールに基づいて曲を付ける取り組み、既存曲を入力として遺伝的アルゴリズム（Genetic Algorithm：GA）などの進化計算によって新たな曲を作る取り組みなど、さまざまなアプローチがなされている[57)]。また、深層学習でもGANだけでなく、時系列データを扱いやすいLSTM（Long Short-Term Memory）ネットワークもよく用いられている。さらに最新の話題として、Text-to-Imageの技術を用いた音楽生成システムRiffusionが2022年12月に公開された。このシステムでは、Stable Diffusionに音楽のスペクトラム画像を追加学習させることで、テキストからスペクトラム画像を生成し、それを音楽として再生する。

文学作品の生成では、2012年にスタートした「きまぐれ人工知能プロジェクト：作家ですのよ」がよく知られている[53)]。星新一のショートショート小説作品をコンピューターで解析して生成した作品を星新一賞に応募する試みを行っている。ここでは、登場人物の設定や話の筋、文章の部品に相当するものは人間が

用意しておき、それらを用いた文章生成という部分にAI技術を適用したというものである。さらに現在大きな話題になっているのが、前述のChatGPTである。簡単な例示や指示を与えて文章を生成することができる。人間が書いたような自然な文章だと言われており、文学的な文章の生成事例も報告されている（技術的面については「2.1.2 言語・知識系のAI技術」に記載している）。

人間の持つ創作欲求（美意識や自己表現など）そのものをAIが持つことは当分難しいとしても、表現の種や見本となるものがあるときに、そのスタイルをまねたり、新たな連想を生み出したりといったことにAI技術が活用できるようになってきた。実際、絵画・音楽の領域では、AI技術を用いて人間の創作活動を支援・活性化するさまざまなソフトウェアツール（ラフスケッチの写実画への変換、着色、作曲、画風・曲調の変換、前衛的・不気味な絵の生成など）が実用化されている。

「AI美空ひばり」「AI手塚治虫」といったプロジェクト[58]も、そのようなツールを活用したものだが、AI技術を用いて故人の名を冠した新作を創作するという行為、あるいは、AI技術を用いて故人を仮想的によみがえらせるというコンセプトに対して、倫理的な視点や社会受容といった面から議論が起きた。両プロジェクトとも、故人の遺族の了承・意向を踏まえつつ取り組まれたものだが、ELSI面からは引き続き議論を深めていくことが必要である。

また、画像生成や文章生成の技術が高品質化し、プロ並みのものが簡単に生成できるようになったため、人間による創作物かAIで自動生成したものかの区別が難しくなってきた。そのため、創作物の権利や偽作に関する問題や、プロのアーティストからの反発なども生じつつある。

［注目すべき国内外のプロジェクト］

❶ NEDOプロジェクト「人と共に進化する人工知能に関する技術開発事業」

国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）のもとで実施されている5年間のプロジェクトである（実施期間：2020年度～2024年度、プロジェクトリーダー：辻井潤一）。人と共に進化するAIシステムの基盤技術開発、実世界で信頼できるAIの評価・管理手法の確立、容易に構築・導入できるAI技術の開発という三つの研究開発項目のもと、16の研究テーマが推進されている。

❷ ソニーのGT Sophy（Gran Turismo Sophy）

GT Sophyは、PlayStationのドライビングシミュレーター「グランツーリスモSPORT」において、世界最高峰のプレーヤーをしのぐドライビングスキルを学習したAIエージェントである[59]。深層強化学習によって訓練された自動意思決定AIだが、実在のレーシングカーやコースの見た目だけではなく、車体の重量バランスや剛性、空気抵抗やタイヤの摩擦などの物理現象に至るまで、現実のレーシング環境が限りなくリアルに再現された仮想環境で学習や自動制御が行われる。さらに、高速走行中の相手やコースのダイナミックな変化に応じたリアルタイム制御、スリップストリームやクロスラインからのオーバーテイクのような高度なレーシングスキル、レーシングエチケットなども獲得している。2022年2月10日発行のNature誌の表紙を飾った。

（5）科学技術的課題

❶ 意思決定支援AIの技術課題

［研究開発の動向］❸に挙げた、膨大な可能性の探索・評価（マルチエージェントシミュレーション、統計的因果推論など）、自動意思決定・自動交渉（強化学習、最適化など）、大規模意見集約・合意形成、多様な価値観の把握・可視化（言論マップ生成、議論マイニングなど）、フェイク対策、意思決定に関する基礎科学（意思決定プロセスのモデル、ELSIなど）のそれぞれは、さらなる研究開発が必要である。

それらの技術を用いての意思決定を支援する機能の提供形態としては、人（個人や集団）に寄り添うAIエージェントという形や、意見集約・合意形成のためのプラットフォームという形が考えられる。AIエージェ

ントのデザインでは、HAI（Human-Agent Interaction、詳細は「2.1.3 エージェント技術」を参照）の設計方法論や、人とAIエージェントの間のトラスト形成（「2.4.7 社会におけるトラスト」を参照）という観点も踏まえる必要がある。また、意見集約・合意形成のためのプラットフォームでは、その健全性・公平性を確保するため、フェイク対策の取り込み、一次情報や意見根拠の追跡・確認、声の大きい意見だけでない意見集約の公平性確保なども課題となる。

❷ HMTの技術課題

［新展開・技術トピックス］❷では、HMTの五つのパターンを示し、そのような人・AI協働の形に応じて、AIの透明性・説明性・制御性などの確保や、人からAIへの権限移譲の管理、AIに対する信頼形成、人のモチベーションへの配慮などに課題があることを述べた。また、各パターンでの協働を通して、人やAIそれぞれの能力が高まり、関係性・パターンが変化する可能性があることにも触れた。どのような問題に対して、どの協働パターンが適していて、協働することでどれほど全体パフォーマンスが高まるか、あるいは、人・AIそれぞれの能力がどう高まり得るか、といった分析や方法論も今後さらに探究されていくものと期待される。

関連して、上記の意思決定支援AIでは、それに人が依存し過ぎると、人の判断能力自体が低下するという懸念がある。 HMTでも同様の懸念は生じ得る。意思決定支援AIやHMTにおいて、人にとって負荷が減って楽になることと、人の能力が高まることの両面からどのようなバランスが望ましいかを考えていくことも必要であろう。

（6）その他の課題

❶ ELSIおよび社会受容性に配慮した研究開発

意思決定支援AIは、倫理的・法的・社会的な視点（ELSI面）から適切であるかを常に考えておかねばならない。例えば、人の支援機能を意図したものが、思考誘導や検閲（表現・言論の自由の制限）と受け止められてしまう可能性もある。逆に、フェイク問題のようなケースに対しては法的規制をかけてしまえばよいのではないかという意見を聞くことがあるが、絶対的な真偽が定まらない言説は非常に多く、法的規制が強く働くと表現・言論の自由が制限されるリスクが高まることに注意を要する[20)]。この点を踏まえて、法的規制の検討は、極めて慎重に行う必要がある。その一方、人のメディアリテラシーを高める教育が重要である。

HMTはより広い人・AI協働の概念であり、同様に配慮が必要であろう。このような面に対して、利用者から見た透明性を確保し、社会受容性に配慮した技術開発が求められる。そのためには、実社会の具体的な問題に適用して社会からのフィードバックを受けるプロセスを、短いサイクルで回しながら判断・改良していくのがよいと考える。

❷ 分野横断の研究開発体制・推進施策

本研究開発領域は、AI技術などの情報技術だけでなく、計算社会科学、脳科学、認知科学、心理学、経済学、政治学、社会学、法学、倫理学などが重なる学際的な領域であり、分野横断の研究開発体制・推進施策が必要である。そのためには、初期段階から分野横断で研究者を共通の問題意識・ビジョンのもとに束ねる研究開発マネジメントが望ましい。現状、本研究開発領域の個々の技術課題・要素技術に関わる研究者は多いものの、研究者それぞれの取り組みは全体の問題意識に対してまだ断片的なものにとどまっている感が強く、分野横断の連携・統合による骨太化が求められる。その際、情報技術側で扱いやすい形の問題にしてしまうとか、人文・社会科学側から結果に対して駄目出しするとかではなく、具体的な問題に対する定式化において双方がコミットすべきである。実社会への適用において発生するさまざまな制約事項を、アルゴリズム・原理のレベルで扱うのか、運用上の制約（法規制など）の形で扱うのかによって、技

術的なアプローチは変わってくる。

❸ 国・社会のフェイク問題と対策に対する意識向上

米国が国家安全保障の観点から重要な研究開発領域と位置付けて投資しているのに対して、日本ではその意識がまだ弱い。日本は米国の事例ほど、フェイク問題や社会分断が深刻化していないため、国・社会の危機感が薄いように思われるが、民主主義を揺るがし得る、社会の方向性を左右し得る、国・組織・個人に対する新しいサイバー攻撃になり得る、といった国・社会にとっての大きなリスクが生じることに備えておくべきである。フェイク問題への対策、フェイクによる攻撃への防御技術を育てておくことや、人々のメディアリテラシーを高めるための教育や啓発施策などを進めることを通して、健全な社会的意思決定・集合知を育てる意識・環境が、安全で信頼できる社会を発展させていくために極めて重要である。

❹ 経済安全保障面の課題

フェイク生成は新種のサイバー攻撃手段となっており、外国からの政治干渉や世論誘導に使われ得ることが日本社会にとって脅威となる。そのための防御技術は安全保障上、自国で育成・確保しておく必要がある。また、より広く意思決定の適切化・迅速化は、社会・生活のさまざまな場面で安全性・健全性および競争力を確保するために不可欠である。そのためのプラットフォームとして、SNSや検索エンジンは重要な役割を担っているが、これに関して海外プラットフォーマーに依存している部分が大きいことは懸念材料と考えられる。

（7）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	↗	マルチエージェントシステムの分野で、オークション・マッチングの理論研究やインセンティブメカニズムの研究が多い。HMTについては、SIP「ビッグデータ・AIを活用したサイバー空間基盤技術」で関連基礎研究を推進しており、日本発のHAI（Human-Agent Interaction）も強みとなる。
	応用研究・開発	○	↗	大規模合意形成支援システムなどで先端的な取り組みや、AI間の交渉・協調・連携に関するCOCNの取り組みが進展している。
米国	基礎研究	◎	↗	MIT CCIのDeliberatoriumやThe ClimateCoLab、Stanford Universityの討論型世論調査をはじめ、学際的な基礎研究が根付いている。AI・マルチエージェントシステムの分野で、メカニズムデザイン、オークション・マッチングの理論研究が広く行われている。HMTの基礎研究も実績がある[35),60)]。
	応用研究・開発	◎	↗	上記基礎研究がそのまま応用研究やベンチャーによる産業化につながる傾向が強い。国および企業によるAI分野への大型投資が行われている（Metaの自動交渉エージェントなど）。
欧州	基礎研究	◎	↗	Imperial College London、Oxford University、Delft University of Technologyなど、自動交渉の基礎研究が強く、論理的なアプローチによる自動交渉の研究も行われている。
	応用研究・開発	◎	↗	市民からの意見集約や合意形成のためのシステム・応用に盛んに取り組まれている。自動交渉の応用ソフトウェア（電力売買など）への取り組みも見られる。
中国	基礎研究	◎	↗	Hong Kong Baptist Universityのメカニズムデザインや自動交渉の基礎理論研究をはじめ、取り組みが活発になってきている。
	応用研究・開発	△	→	顕著な活動は見当たらない。

韓国	基礎研究	△	→	顕著な活動は見当たらない。
	応用研究・開発	△	→	顕著な活動は見当たらない。

（註1）フェーズ

基礎研究：大学・国研などでの基礎研究の範囲

応用研究・開発：技術開発（プロトタイプの開発含む）の範囲

（註2）現状　※日本の現状を基準にした評価ではなく、CRDS の調査・見解による評価

◎：特に顕著な活動・成果が見えている　　○：顕著な活動・成果が見えている

△：顕著な活動・成果が見えていない　　×：特筆すべき活動・成果が見えていない

（註3）トレンド　※ここ1～2年の研究開発水準の変化

↗：上昇傾向、→：現状維持、↘：下降傾向

参考文献

1）丸山文宏,「人とAIの関係性を考える」,『AIデジタル研究』第7号（2023年2月）.

2）科学技術振興機構 研究開発戦略センター,「戦略プロポーザル：複雑社会における意思決定・合意形成を支える情報科学技術」, CRDS-FY2017-SP-03（2018年12月）.

3）福島俊一,「複雑社会における意思決定・合意形成支援の技術開発動向」,『人工知能』(人工知能学会誌) 34巻2号（2019年3月）, pp. 131-138.

4）Edmond Awad, et al., “The Moral Machine experiment”, *Nature* Vol. 563 (24 October 2018), pp. 59-64. DOI: 10.1038/s41586-018-0637-6

5）笹原和俊,『フェイクニュースを科学する―拡散するデマ, 陰謀論, プロパガンダのしくみ―』(化学同人, 2018年）.

6）湯淺墾道,「米大統領選におけるソーシャルメディア干渉疑惑」,『情報処理』(情報処理学会誌）58巻12号（2017年12月）, pp. 1066-1067.

7）藤代裕之,『フェイクニュースの生態系』(青弓社, 2021年）.

8）Herbert A. Simon, *Administrative behavior: a study of decision-making processes in administrative organization* (Macmillan, 1947). (邦訳：二村敏子・桑田耕太郎・高尾義明・西脇暢子・高柳美香訳,『新版 経営行動：経営組織における意思決定過程の研究』, ダイヤモンド社, 2009）.

9）Daniel Kahneman, *Thinking, Fast and Slow*, (Farrar, Straus and Giroux, 2011). (邦訳：村井章子訳,『ファスト＆スロー：あなたの意思はどのように決まるか?』, 早川書房, 2014年）

10）Daniel Kahneman and Amos Tversky, “Prospect Theory: An Analysis of Decision under Risk”, *Econometrica* Vol. 47, No. 2 (March 1979), pp. 263-291.

11）Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Yale University Press, 2008). (邦訳:遠藤真美訳,『実践 行動経済学』, 日経BP社, 2009年）

12）坂上雅道・山本愛実,「意思決定の脳メカニズム―顕在的判断と潜在的判断―」,『科学哲学』(日本科学哲学会誌）42-2号（2009年）.

13）遠藤薫,『ソーシャルメディアと〈世論〉形成』(東京電機大学出版局, 2016年）.

14）山口真一,『ソーシャルメディア解体全書：フェイクニュース・ネット炎上・情報の偏りネット炎上の研究』(勁草書房, 2022年）.

15）金子格・須川賢洋（編）,「小特集 ディジタルゲリマンダとは何か―選挙区割政策からフェイクニュースまで」,『情報処理』(情報処理学会誌）58巻12号（2017年12月）, pp. 1068-1088.

16）Eli Pariser, *The Filter Bubble: What the Internet is Hiding from You* (Elyse Cheney Literary Associates, 2011). (邦訳：井口耕二訳,「閉じこもるインターネット」, 早川書房, 2012年）

17）「Politics and Technology テクノロジーは民主主義の敵か?」,『MITテクノロジーレビューSpecial Issue』Vol. 11（2018年）.
18）田中辰雄・浜屋敏,「ネットは社会を分断するのか－パネルデータからの考察－」,『富士通総研 経済研究所 研究レポート』No. 462（2018年）.
19）Ruben Tolosana, et al., “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection”, arXiv: 2001.00179 (2020).
20）科学技術振興機構 研究開発戦略センター,「公開ワークショップ報告書：意思決定のための情報科学～情報氾濫・フェイク・分断に立ち向かうことは可能か～」, CRDS-FY2019-WR-02（2020年2月）.
21）Tom Brown, et al., “Language Models are Few-Shot Learners”, *Proceedings of the 34th Conference on Neural Information Processing Systems*（NeurIPS 2020; December 6-12, 2020).
22）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報告書　複雑社会における意思決定・合意形成を支える情報科学技術」, CRDS-FY2017-WR-05（2017年10月）.
23）井之上直也,「言語データからの知識獲得と言語処理への応用」,『人工知能』(人工知能学会誌）33巻3号（2018年5月）, pp.337-344.
24）水野淳太・他,「大規模情報分析システム WISDOM X, DISAANA, D-SUMM」,『言語処理学会第23回年次大会発表論文集』(2017年）, pp. 1077-1080.
25）Yannick de Jong, “Levels of Data Analytics”, *IThappens.nu*（20 March 2019). http://www.ithappens.nu/levels-of-data-analytics/ (accessed 2020-11-08）
26）鹿島久嗣・小山聡・馬場雪乃,『ヒューマンコンピュテーションとクラウドソーシング』(講談社, 2016年）.
27）牧野貴樹・澁谷長史・白川真一（編著）,他19名共著,『これからの強化学習』(森北出版, 2016年）.
28）大槻知史(著)・三宅陽一郎(監修),『最強囲碁AI アルファ碁 解体新書(増補改訂版)』(翔泳社, 2018年).
29）藤巻遼平・他,「予測から意思決定へ ～予測型意思決定最適化～」,『NEC技報』69巻1号（2016年）, pp. 64-67.
30）藤田桂英・森顕之・伊藤孝行,「ANAC：Automated Negotiating Agents Competition（国際自動交渉エージェント競技会）」,『人工知能』(人工知能学会誌）31巻2号（2016年3月）, pp. 237-247.
31）桑子敏雄,『社会的合意形成のプロジェクトマネジメント』(コロナ社, 2016年）.
32）Jeff Conklin and Michael L. Begeman, “gIBIS: a hypertext tool for exploratory policy discussion”, *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*（CSCW ‘88: Portland, USA, 26-28 September 1988), pp. 140-152. DOI: 10.1145/62266.62278
33）James S. Fishkin, *When the People Speak: Deliberative Democracy and Public Consultation*（Oxford University Press, 2011).（邦訳：岩木貴子訳, 曽根泰教監修,『人々の声が響き合うとき：熟議空間と民主主義』, 早川書房, 2011年）
34）Mark Klein, “Enabling Large-Scale Deliberation Using Attention-Mediation Metrics”, *Computer Supported Cooperative Work* Vol. 21, No. 4-5 (2012), pp. 449-473. DOI: 10.2139/ssrn.1837707
35）Thomas W. Malone, *Superminds: The Surprising Power of People and Computers Thinking Together*（Little, Brown and Company, 2018).
36）Takayuki Ito, et al., “D-Agree: Crowd Discussion Support System Based on Automated Facilitation Agent”, Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20, New York, 7-12 February 2020), pp. 13614-13615. DOI: 10.1609/aaai.v34i09.7094
37）伊藤孝行・他,「エージェント技術に基づく大規模合意形成支援システムの創成　―自動ファシリテーションエージェントの実現に向けて―」,『人工知能』(人工知能学会誌）32巻5号（2017年9月）, pp. 739-747.

38）水野淳太・他,「言論マップ生成技術の現状と課題」,『言語処理学会第17回年次大会発表論文集』(2011年）, pp. 49-52.
39）岡崎直観,「自然言語処理による議論マイニング」,『人工知能学会全国大会（第32回）』(2018年）1D2-OS-28a-a（OS-28招待講演）. https://www.slideshare.net/naoakiokazaki/ss-100603788（accessed 2023-02-01）
40）柳井孝介・他,「AIの基礎研究：ディベート人工知能」,『日立評論』98巻4号（2016年4月）,pp. 61-64.
41）佐藤健,「論理に基づく人工知能の法学への応用」,『コンピュータソフトウェア』(日本ソフトウェア科学会誌）27巻3号（2010年7月）, pp. 36-44. DOI: 10.11309/jsssst.27.3_36
42）“Special Issue: This is Watson”, *IBM Journal of Research and Development* Vol. 56 issue 3-4 (May-June 2012).
43）Robert M. Bond, et al., “A 61-million-person experiment in social influence and political mobilization”, *Nature* Vol. 489 (12 September 2012), pp. 295-298. DOI: 10.1038/nature11421
44）Xinyi Zhou and Reza Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities”, *ACM Computing Surveys* Vol. 53, No. 5 (September 2020), Article No. 109. DOI：10.1145/3395046
45）立岩陽一郎・楊井人文,『ファクトチェックとは何か』(岩波書店, 2018年）.
46）Tsubasa Tagami, et al., “Suspicious News Detection Using Micro Blog Text”, *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*（PACLIC 32, Hong Kong, 1-3 December 2018).
47）鳥海不二夫・山本龍彦,『デジタル空間とどう向き合うか：情報的健康の実現をめざして』(日経BP, 2022年）.
48）鈴木宏昭,『認知バイアス：心に潜むふしぎな働き』(講談社, 2020年）.
49）科学技術振興機構 研究開発戦略センター,「戦略プロポーザル：デジタル社会における新たなトラスト形成」, CRDS-FY2022-SP-03（2022年9月）.
50）Darius Afchar, et al., “MesoNet: a Compact Facial Video Forgery Detection Network”, *Proceedings of IEEE International Workshop on Information Forensics and Security*（WIFS 2018, Hong Kong, 11-13 December 2018). DOI: 10.1109/WIFS.2018.8630761
51）Huy H. Nguyen, Junichi Yamagishi and Isao Echizen, “Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos”, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*（ICASSP 2019, Brighton, 12-17 May 2019). DOI: 10.1109/ICASSP.2019.8682602
52）Kazuo Okamura and Seiji Yamada, “Adaptive Trust Calibration for Human-AI Collaboration,” *PLOS ONE* Vol. 15, No. 2 (February 2020), e0229132. DOI: 10.1371/journal.pone.0229132
53）竹永康彦・他（編）,「小特集：創造性・芸術性におけるAIの可能性」,『電子情報通信学会誌』102巻3号（2019年3月）, pp. 207-264.
54）David Foster, *Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play*（O'reilly Media Inc., 2019).（邦題：松田晃一・小沼千絵訳,『生成Deep Learning：絵を描き、物語や音楽を作り、ゲームをプレイする』, オライリージャパン, 2020年）.
55）徳井直生,『創るためのAI：機械と創造性のはてしない物語』(ビー・エヌ・エヌ, 2021年）.
56）David Cope, *Computer Models of Musical Creativity*（The MIT Press, 2005).（邦訳：平田圭二監修, 今井慎太郎・大村英史・東条敏訳,『人工知能が音楽を創る』, 音楽之友社, 2019年）.
57）Jean-Pierre Briot, Gaëtan Hadjeres and François Pachet, “Deep Learning Techniques for

Music Generation - A Survey", arXiv：1709.01620 (2017).
58）折原良平（編）,「特集：AIでよみがえる手塚治虫」,『人工知能』(人工知能学会誌）35巻3号, 2020年5月, pp. 390-429）　　.
59）Peter R. Wurman, et al., "Outracing champion Gran Turismo drivers with deep reinforcement learning", *Nature* No. 602 (2022), pp. 223-228. DOI: 10.1038/s41586-021-04357-7
60）National Academies of Sciences, Engineering, and Medicine, *Human-AI Teaming: State-of-the-Art and Research Needs* (The National Academies Press, 2022). DOI：10.17226/26355
61）笹原和俊,『ディープフェイクの衝撃：AI技術がもたらす破壊と創造』(PHP研究所, 2023年).