CRDS-FY2019-WR-02

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～

国立研究開発法人科学技術振興機構 研究開発戦略センター
Center for Research and Development Strategy, Japan Science and Technology Agency

エグゼクティブサマリー

国立研究開発法人科学技術振興機構（JST）研究開発戦略センター（CRDS）は、国の科学技術イノベーション政策に関する調査・分析・提案を中立的な立場に立って行う組織として、今後わが国が取り組むべき重要な研究領域に関する提言を行ってきている。また、その提言に向けた前段階として、広く技術分野の状況を俯瞰し、注目すべき研究動向の抽出を行ってきている。これら俯瞰や提言は、文部科学省・内閣府をはじめとする政府の各種政策立案のためのインプットとなる。

この提言活動の一環として、今回、公開ワークショップ「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～」を開催した。この開催の背景と狙いは以下の通りである。

わが国が描く社会像 Society 5.0 は「人間中心の社会」であり、社会に広がる人工知能（AI）技術に関して「人間中心の AI 社会原則」を掲げている。AI 技術が進化し、さまざまなタスクの自動化が可能になりつつあるが、自動化するタスクの定義や結果の最終判断等、上位の意思決定に責任を持つのは人間である。しかし、そのように人間の主体的意思決定が求められる一方で、情報の氾濫、社会のボーダーレス化、フェイク情報の流通等によって、人間の主体的意思決定が難しくなり、悪意・扇動意図を持った情報操作によって社会の方向性が左右される危険性さえ生じている。このような状況において、上記懸念を回避し、人間の主体的意思決定を支援するための情報科学の重要性が増している。

この問題意識のもと、JST CRDS では、この技術分野における研究開発の状況・動向の調査や、有識者へのインタビュー、技術課題を俯瞰するためのワークショップ等を実施し、2018 年 3 月には戦略プロポーザルも発行した。これまでの動向調査・提言内容の詳細は、付録 4 に示す関連資料にまとめている。

今回さらに、次の 3 点を狙いとして公開ワークショップを開催した。

1. これまで技術課題を広く俯瞰してきたが、今回、5 件の発表から技術アプローチの状況を把握・共有することを通して、取り組むべき具体的な技術開発目標の議論につなげる。
2. フェイク情報問題や分断問題等、ますます深刻化する意思決定の問題に対して、技術のみで解決できるわけではないという現実も踏まえつつ、技術開発の必要性や有効な取り組み方を明らかにする（総合討論での主要議題として扱う）。
3. これまでのヒアリング等から、ここで取り上げたような問題意識を持って取り組んでいる研究者が増えつつあるという感触を得ており、公開ワークショップの形で開催することで、研究コミュニティの形成・活性化を促進する。

5 件の発表では、まず笹原和俊氏（名古屋大学）から計算社会科学によるアプローチとして、フェイクニュース問題を中心に、人間の認知特性や SNS（Social Networking Service）における拡散傾向の分析や、それを踏まえた対策の可能性が示された。次に山岸順一氏（国立情報学研究所）からメディア解析技術によるアプローチとして、フェイク動画やフェイク音声の検出技術の現状が示された。乾健太郎氏（東北大学）からは自然言語処理技術によるアプローチとして、肯

定・否定の両面でネット上の意見を整理して見せる言論マップや、人間が行うファクトチェックを支援するシステムの事例等が紹介された。伊藤孝行氏（名古屋工業大学）からは集合知・エージェント技術によるアプローチとして、エージェントによる自動ファシリテーションを行う大規模意見集約システム D-Agree の技術と適用事例が紹介された。最後に森永聡氏（NEC）から機械学習・最適化技術によるアプローチとして、自動意思決定システム、および、そのようなシステム間の自動交渉技術への取り組みが示された。

総合討論では、まずコメンテーターの立場で小林正啓氏（花水木法律事務所）から「法はなぜフェイクニュースを直接規制しないのか？」「フェイクニュースを研究する実益は？」「集団的意思決定支援技術の実益は？」等について、法律家の視点からの見解や問題提起が示された。これを踏まえた総合討論では、課題の重要性や技術的対策の必要性が確認されるとともに、人間・社会の側の変化や受け止め方も含めて、多面的な取り組み・議論を進めていくことの必要性が再認識された。

フェイクとファクトを見分けることが人間に難しくなるということは、犯罪捜査・裁判等における証拠能力や、さまざまな取引や人間関係における信用・信頼が揺らぎ、また、社会的・政治的な主張や科学技術的な研究成果における捏造のリスクも高まることになる。それが悪意・扇動・干渉意図をもってソーシャルメディアで拡散されるならば、選挙を含めた社会の方向性の選択や国の安全保障にも影響を与えかねない。

このような問題は技術だけで解決できるものではないが、今回の公開ワークショップでは情報科学技術によるアプローチが状況改善に貢献し得る可能性が示された。人間は認知限界・認知バイアスを持っており、それらを乗り越えるために、情報科学技術をうまく活用していくことが役立つはずである。

本ワークショップには 150 名の参加者があり、開催後のアンケートではこのような取り組みの意義を支持する回答・意見が多数寄せられた（付録 2・付録 3）。引き続き、この問題に対する幅広い視点からの議論と、「意思決定のための情報科学」の研究開発の活性化を進めていきたい。

目　次

1．趣旨説明

福島俊一（科学技術振興機構 研究開発戦略センター）

1.1 開催趣旨・問題意識

ワークショップの開催趣旨や今回の問題意識について述べる。

日本が描く社会像である Society 5.0 では「人間中心の社会」を掲げている。一方で、AI（人工知能）技術が進化し、我々の生活の中にもいろいろ入ってきて、AI によってさまざまなタスクが自動化されるようになってきた。しかし、人間中心の社会ならば、我々が AI を使いこなすとか、AI が我々をサポートしてくれるといった形になるわけで、いろいろな判断・意思決定に関して、最終的には人間がしっかりしていなければならない。つまり、AI で自動化が進んでいっても、何を自動化するかの決定や、自動化するタスクの定義だったり、その結果に関わる責任だったり、上位の判断や最終的な判断は人間が負うべきである。そうすると、人間の意思決定というものがしっかりしたものでなくてはいけないということになる。

しかし、その反面、情報化社会の進展の中で、逆に我々の意思決定が難しくなってきている場面が増えているように感じる。例えば、情報が氾濫していて、それをどう使って意思決定していけばよいかは難しい。社会がボーダーレス化し、何かアクションを起こすと、その影響が及ぶ範囲が非常に広くて、思いもよらなかったような問題が起きてしまう。それから、フェイク情報の流通が引き起こす混乱。そういったことがあって意思決定が難しくなってきている。さらには、そこにつけ入るような形で、悪意や扇動意図を持って情報を操作する者、他者の思考を誘導しようとする者が出てきている。例えば選挙に影響が出るなど、社会の方向性が左右されるような危険性さえ生じている。

以上のような問題に直面している状況で、人間が主体性を持って、しっかりとした意思決定を行えるように、情報科学技術によって何かできないかと思う。ここであげた問題・懸念は必ずしも技術だけで全てを解決できるものではない。社会の制度や文化、場合によったら民主主義の考え方のようなものまで絡んでくるものかもしれない。しかし、技術だけで解決できるものではないとしても、このような状況になってきたことの背景には情報化の進展があり、情報科学技術が大きく関わっているので、状況の改善のために何らかの手段が提供できるだろうと思う。

ソーシャルメディアの影響やフェイク問題を中心に、今回のような問題が取り上げられる機会が増えている。しかし、それらは問題意識の提示や社会問題としての議論という面が強いものになっている。それに対して、今回のワークショップでは技術的な取り組みの可能性に目を向けたい。技術的な対策としては、どのようなものが取り組まれているのか、その発展の方向性や見込み等について論じられるような機会にしたい。

加えて 1 点補足的な話をする。

情報科学技術や AI 技術は社会や人間を変える技術だと言われる。その一方で、日本は「人間中心の社会」をビジョンとし、「人間中心の AI 社会原則」も掲げている。その中で、社会のビジョンや、社会における AI のあるべき姿が示されているが、人間自体はどうかという点にも目を向けるべきと思う。技術が引き起こすさまざまな変化の一方で、人間はそれになかなかついていけないというケースが起きている。今回取り上げている意思決定問題もその一つである。人間が

主体的に意思決定をすることが難しくなってきている状況にあるわけで、人間中心の社会であるためには、人間をどう支援するのがよいかを考えたいと思っている。

1.2 主体的意思決定の困難化

では、実際に、人間の主体的な意思決定が難しくなっていると感じる場面を、いくつか例示してみたい。1つ目はクリティカルな要因・影響の見落としである（図 1-1）。

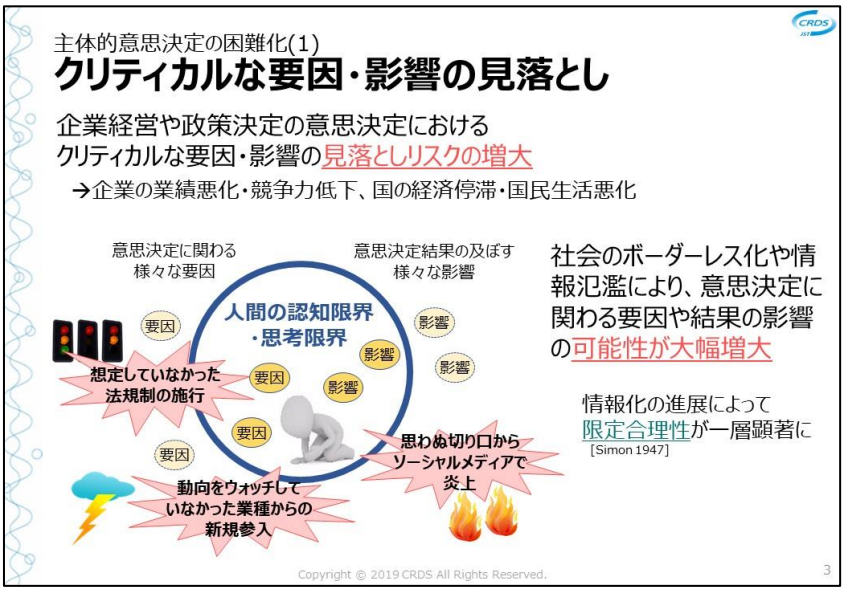


図 1-1 主体的意思決定の困難化(1)クリティカルな要因・影響の見落とし

情報化の進展に伴い、社会のボーダーレス化や情報氾濫が進んだ。これは意思決定という場面において、さまざまな可能性を考えなければいけないことになる。意思決定の際に考えねばならない要因がいろいろあるが、それが非常に膨大になってしまい、関係しそうな全ての要因やそのときの条件を考えて判断することはとても難しい。さらに、意思決定をしてアクションを起こしたときに、その結果がどんなところにまで影響してしまうか、その起こり得る可能性を全て思い描くこともまた非常に難しい。

例えば企業経営のシーンを考えると（図 1-1 に例示したように）、想定していなかった法規制が施行されたり、動向をウォッチしていなかった業種から新規参入がきたり、製品・サービスを出したときに思わぬ切り口からソーシャルメディアで炎上が起きたりといったことが、要因や影響の見落としの例である。

2つ目はソーシャルメディアによる思考誘導である（図 1-2）。最近「フェイクニュース」という言葉をよく聞くし、聞きなれない言葉ではあるが、ソーシャルメディアで政治操作をすることを「デジタルゲリマンダー」と言う。大きな話題になったのは、アメリカの大統領選挙のときにフェイクニュースやフェイク広告が多数流れて、それが選挙結果に影響を与えたのではないかという話である。また、ある実験結果によると、選挙のときにソーシャルメディア上で、友人が選挙に行ったという情報をうまく表示してやると、それを見た人が自分も選挙に行こうと行動を促

す効果があるとのことである。これを個人のプロフィール分析と組み合わせて悪用すると、特定の政党支持者だけに投票を促すようなことも起こり得るかもしれない。

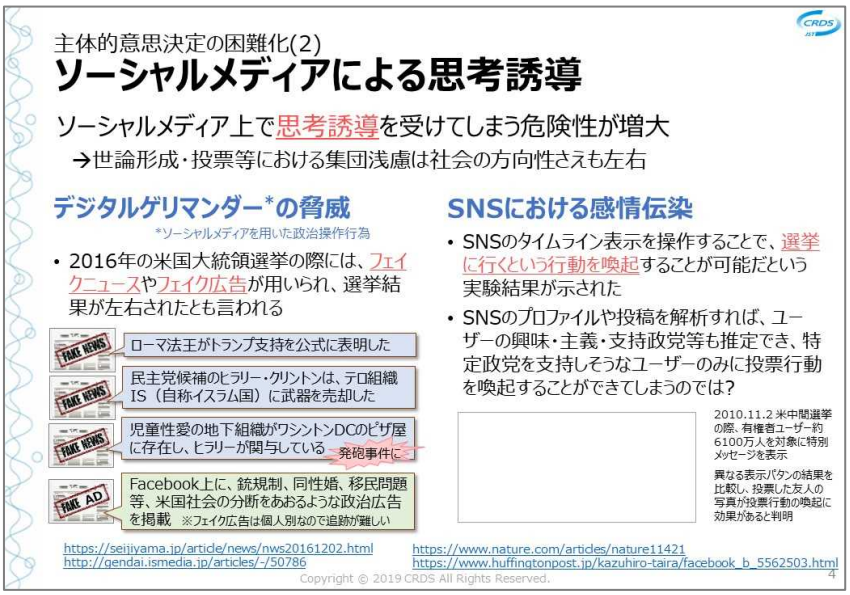


図 1-2　主体的意思決定の困難化(2)ソーシャルメディアによる思考誘導

3つ目は価値観の対立激化や社会の分極化である（図 1-3）。ソーシャルメディアでよく見られる現象として、フィルターバブルやエコーチェンバー等が知られている。そして、これが価値観の対立を激化させたり、社会の分極化を招いたりしているのではないかという分析がある。フィルターバブルというのは、自分に都合がいい、自分の意見にあったものだけが集まり、そういうものしか聞かないというような状況になっていることを指す。エコーチェンバーというのは、そういった自分に近い意見を持つ仲間内で、その意見が共鳴・増幅される現象のことである。人間はもともと、考えにそういったバイアスが入る傾向を持っているが、ソーシャルメディアでその傾向が一層強められて、対立や分極化につながると懸念される。

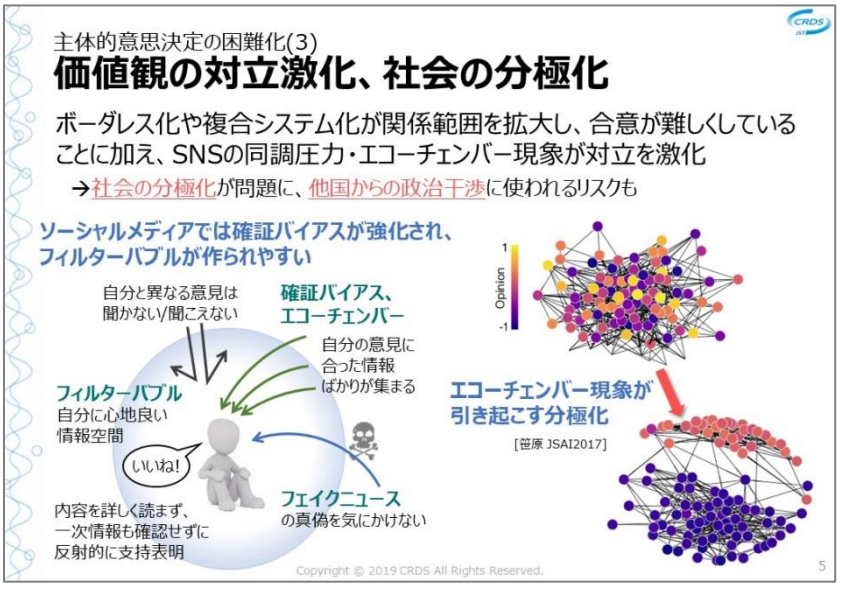


図 1-3　主体的意思決定の困難化(3)価値観の対立激化、社会の分極化

4つ目は、先ほどのフェイクニュースと近いが、「リアルなフェイク」の流通の問題である（図1-4）。「リアルなフェイク」というのは言葉としておかしいかもしれないが、あたかも本物のようなフェイクという意味で、特に動画や音声で簡単に、あたかも本物のようなフェイク動画・フェイク音声が作られるようになってしまった。そういったものを作る技術が非常に進歩した。例えば、政治家が実際には発言していないのに、あたかもその政治家が発言したかのようなフェイク動画を作ることができてしまい、多くの人々は動画で見せられると、それを信じてしまいやすい。何が真実で、何が偽物かを見分けるのが本当に難しくなり、意思決定が惑わされてしまう。これは非常に危険な事態である。

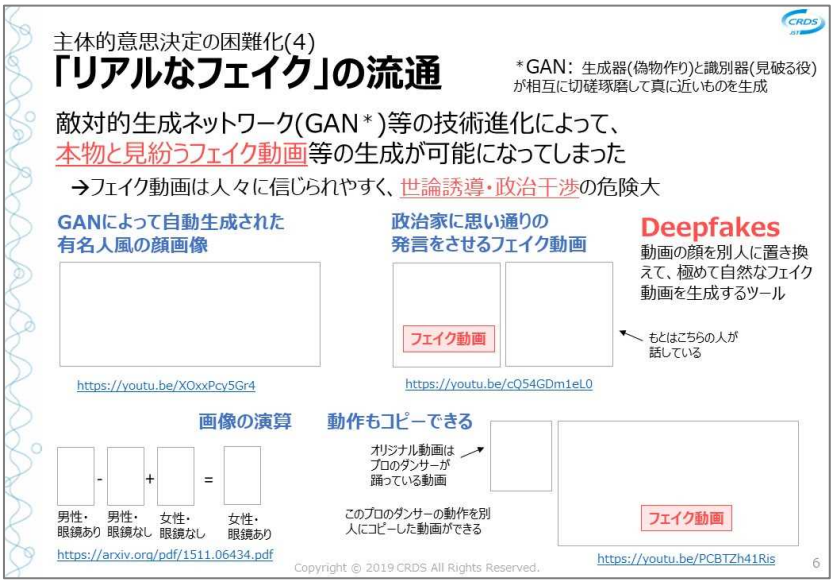


図 1-4　主体的意思決定の困難化（4）「リアルなフェイク」の流通

1.3　技術的対策の方向性

ここまで述べてきたような主体的な意思決定が難しくなってきている状況がある中で、どのような対策が考えられるか（図1-5）。この困難化の二大要因としてあげられるのは、クリティカルな要因の見落としと悪意を持った思考誘導であろう。なので、これらの二大要因を取り除くために技術をうまく活用していくという、リスク低減型の対策がまず考えられる。

これは、いわば防御の立場だが、逆に、こんな良い形に持っていきたいという促進型の対策も考えられる。どのような意思決定の形へ促すのがよいかというと、さまざまな可能性やさまざまな意見を熟慮・熟議して決めていけるような形ではないかと思う。きちんとよく深く考えた上で判断することを促すような仕組み、別な言い方をすると「集合知」を生み出すための仕組みと言ってもいい。これを情報技術で支援するというのが、もう一つの対策になる。

では、そのような対策が実現されたときに、具体的にどのような支援機能や支援システムの形態で、我々を支援してくれるのか。この図（図1-5）には3通りのイメージを示した。

第一の形態は「寄り添うエージェント」、我々が意思決定をするときに、スマートフォンやスマートスピーカーのようなデバイスでエージェントソフトウェアが動作していて、もし我々が何か考えるべき可能性を見落としていたら、我々の傍らでそれを指摘してくれるというものである。

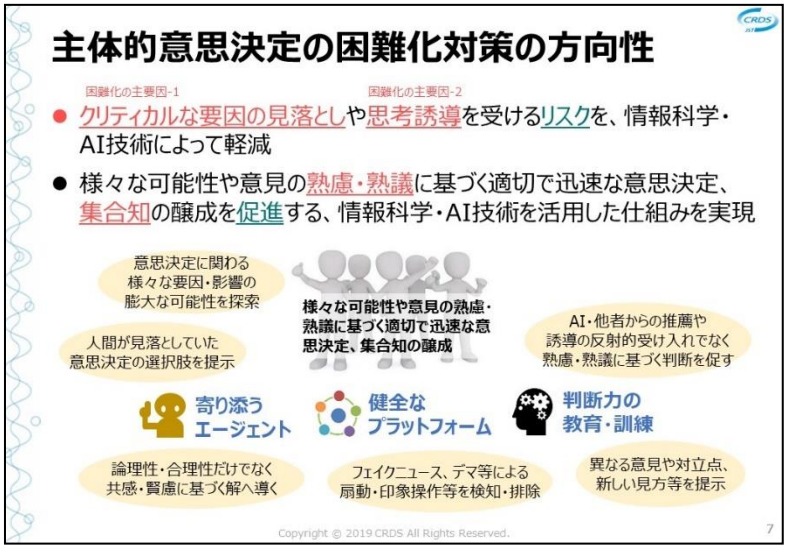


図 1-5　主体的意思決定の困難化対策の方向性

第二の形態は「健全なプラットフォーム」である。多数の利用者が参加して共同で使う、議論や意見集約のためのプラットフォームがあるとき、そこでフェイクが広まったり、炎上が起きたり、といった悪いことが起こりにくいようにモニタリングする。あるいは、声の大きい特定の人の意見だけに傾かないように、さまざまな人の意見が集まりやすい仕組みを作るといった形である。

第三の形態は「判断力の教育・訓練」である。AI のレコメンドやアドバイスに頼ってばかりいると、我々自身があまり考えなくなってしまい、人間自身の判断力・意思決定力が低下してしまうという懸念が生じる。そこが低下していると、人間中心の社会がおかしな方向へ行ってしまうかもしれない。なので、唯一の解が示されて人間がただそれに従ってしまうという形ではなく、人間自身もよく考えることを促すなど、自然に判断力の教育・訓練になるような仕組みを情報技術をうまく使って作っていくことも考えないといけないのではないかと思っている。

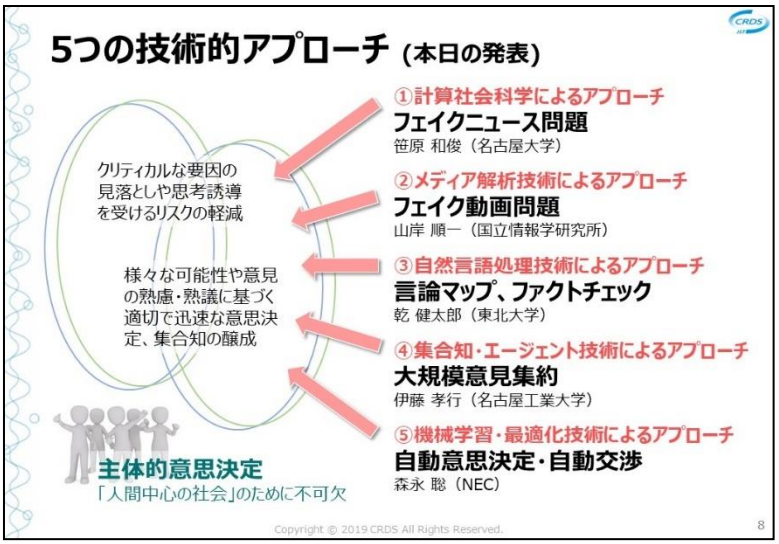


図 1-6　5 つの技術的アプローチ

このような支援機能・支援システムが、実際にどのような技術で、どのようなことまでできそうなのかを、このあと、5 人の先生方にご講演していただく（図 1-6）。笹原先生から計算社会科学によるアプローチ、山岸先生からメディア解析技術によるアプローチ、乾先生から自然言語処理技術によるアプローチ、伊藤先生から集合知・エージェント技術によるアプローチ、森永先生から機械学習・最適化技術によるアプローチが紹介される。一つの技術で全てを解決できるというものではないので、多面的なアプローチが進められており、それぞれの取り組みからいろいろな可能性が感じられるものと思う。

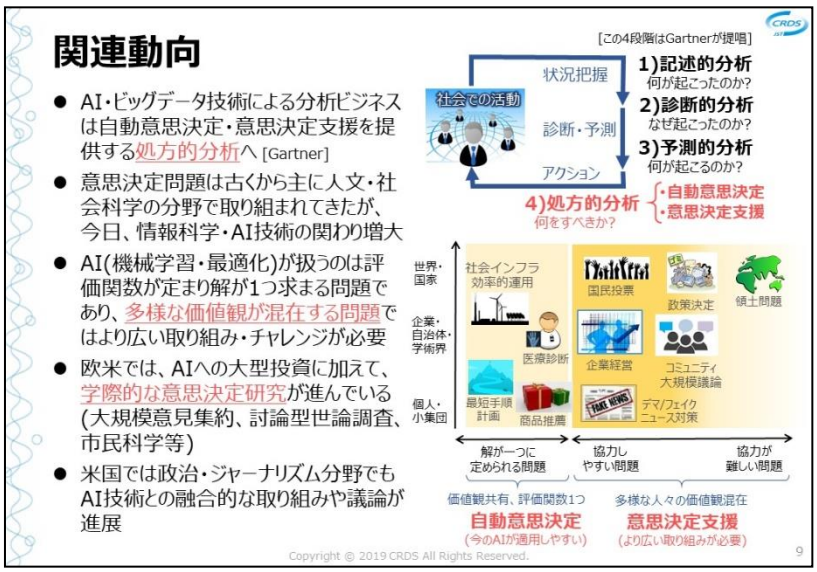


図 1-7　関連動向

1.4　関連動向

このような取り組みに関連する国内外の動向についても簡単に触れておく（図 1-7）。

1 つはビジネス的な面でも意思決定に関わる技術が重視されているということ。データサイエンスとか、データ分析ビジネスとか言われるところでは、ビッグデータを分析して見える化して、今どうなっているのかを把握し、それに対して判断・予測を行う。このようなステップを踏まえたとき、やはり最終的な価値を生み出すのは、アクションをうまく実行する、あるいは、どういうアクションを実行すればよいのかを提示する、というところまで踏み込んだときだと言われる。ガートナー社はこの図（図 1-7）にあるように、データ分析ビジネスを記述的分析、診断的分析、予測的分析、処方的分析という 4 段階に分けており、顧客から見て最も価値が高まるのは処方的分析だとしている。

それから、AI 技術、特に機械学習・最適化技術による判断の自動化が進んでいるが、これは解が一つに定められる問題、つまり、価値観がはっきり定まっていて、1 つの評価関数が定義できるタイプの問題を扱うのが基本である。こういうタイプに落とし込める問題は、いま AI 技術による判断の自動化がどんどん試みられている。しかし、それとは異なるタイプの問題がある。それは、さまざまな人の価値観が混在したり、対立したりしているような問題である。これにも AI 技術が活用できる面はあるが、それだけでは必ずしもうまくいかない。人文・社会科学系の知見も含めた、より総合的な、より広い取り組みが必要になってくると思う。欧米での先進的な取り

組みでは、学際的なアプローチが取り入れられている。

こういった視点も含めて、これからどのように考えて、技術的な取り組みを進めていけばよいか、についても本日は議論していきたい。

1.5 いくつかの論点

5 名の先生方からそれぞれの技術的アプローチを発表していただいた後、総合討論を予定している。そのときの議論の素材として、このスライド（図 1-8）にいくつかの論点をあげた。

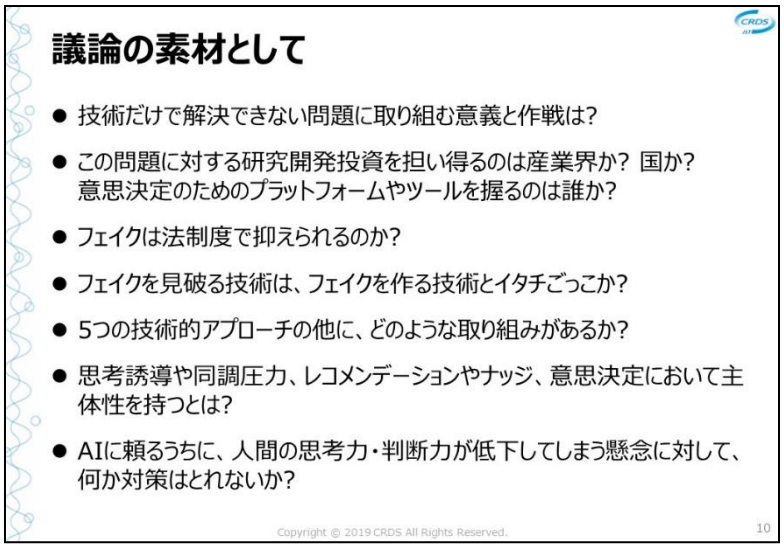


図 1-8 議論の素材として

まず、そもそもの話として、最初に述べたように、今回の問題は技術だけで全て解決できるものではない中で、技術的な取り組みを進める意義をどう考えるか、また、取り組んでいく上での作戦あるいは考え方といったあたりを共有したいと思っている。

それから、意思決定というのは人間にとって生きていく上で基本的なもので、フェイク問題はある意味、国としても深刻な問題になり得るような話だけれど、産業界がビジネスにできるネタなのかというと、必ずしもそういうイメージはぱっとは思い浮かばない。今日この後に話のあるいろいろな技術に対する研究開発投資は国が担うのがよいか、どういう形になると産業界で推進できるのかといったところも、JST のようなファンディング機関としては考えないといけない点だと思う。その際に、フェイクを見破る技術とフェイクを作る技術はイタチごっこではないかという点も、投資の考え方を左右する。

また、意思決定のためのプラットフォームになるようなものが Facebook だったり、Google だったり、海外の企業がコントロールしている場だとしたら、なんらかの懸念が出てくるかもしれない。このあたり、どういう方向に持っていけばよいだろうか。

それから、やはり気になる点として、技術だけでは解決できないので、法制度などを用いてコントロールできないかという考えが出てくる。しかし、実際そのような方法で問題がうまく抑えられるものなのかというのも知りたい。

今回、私の方で、取り組みが進んでいると思った重要なアプローチを 5 つ選んで、5 人の先生

方にその紹介をお願いした。しかし、他にもいろいろな取り組みが考えられると思う。こんなアプローチが考えられるとか、あるいは、ご自分ではこういう考えでこの問題にアプローチしているとか、総合討論ではそういったご意見もいただきたい。

以上、問題意識を中心に、今回のワークショップの趣旨を説明させていただいた。ぜひ議論させていただきたい。

2．発表

2.1　フェイクニュース問題：計算社会科学によるアプローチ

笹原和俊（名古屋大学）

私が研究している分野は、計算社会科学と言われるもので、英語では Computational Social Science である。日本ではまだまだ聞かれない研究分野だが、10 年前に Computational Social Science というタイトルで、David Lazer をはじめとする非常に著明な研究者たちが名前を連ねた論文がサイエンスに出た（図 2-1-1）。黄色のマーカを引いている「digital traces」と「computational social science」がキーワードである。

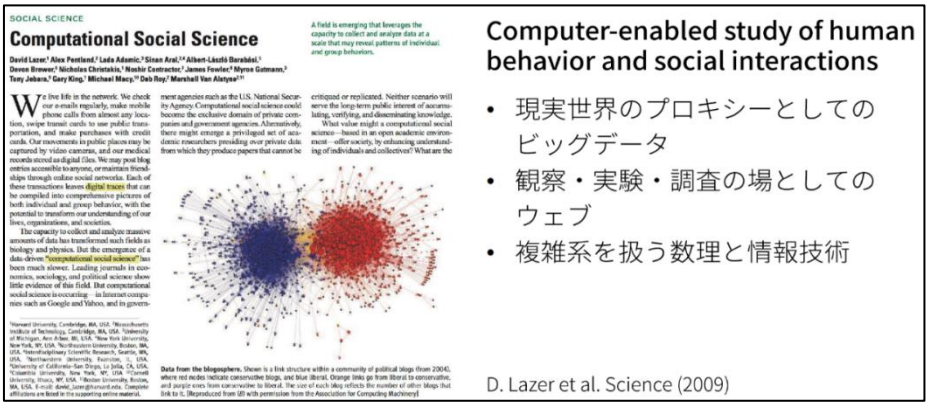


図 2-1-1　計算社会科学

当時、人間の行動というものは E メールのやりとりだったり、携帯電話の通話ログだったり、まだ SNS はなくてブログのデータ、それから安価になりつつあったいろいろなセンサーのデータとして記録された。そういうものに残っている人間の行動の電子的な痕跡、それをビッグデータという言葉で呼んだりするのだが、そういうものを分析することで、これまでの社会科学がアプローチできなかったような人間行動や社会現象を定量的に研究しようという学問が計算社会科学である。

フェイクニュースという言葉は非常に悩ましい言葉で、フェイクニュースに関して議論しようとすると、異なる分野の人が異なる概念を想起してしまうので、コミュニケーションがなかなか成立しないということがある。いろいろなタイプのフェイクニュースがあるので、それに応じた対策を考えなければいけない、あるいは、そういうものの実態を調査しなければいけないというような流れになってきている。

計算社会科学でフェイクニュースに切り込むとき、重要な切り口の一つは拡散である。私自身もフェイクニュースの拡散を止めないと、この問題の緩和にはならないだろうと思っている。これまでもデマやプロパガンダの類いはあったわけだが、SNS はたくさんの人をつなげてしまっているので、ひとたび、誰かのアテンションに情報が上って共有され始めると、あとはチェーンリアクションでどんどん拡散していく。これを止めるには、こういう拡散がどんな特徴を持っているか、あるいはどんな情報が拡散されやすいのか、そういった実態をきちんと定量的に把握しなければいけない。そういう研究が今は非常に盛んに行われている。

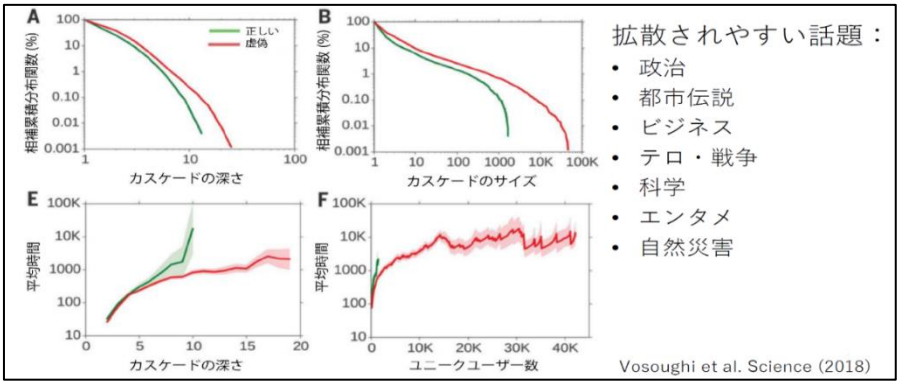


図 2-1-2　正しいニュースと偽ニュースの届き方の比較

情報拡散の研究で一番有名なのは図 2-1-2 に示したものであろう。MIT の研究者たちがツイッター創業の年から 2017 年までの英語の全てのツイートを対象として、ファクトチェック機関がこれは正しい、これはフェイク、とラベリングし、分類できた情報に関してのみ、どういうふうにそれが拡散するかというのを調べた非常に大規模な研究である。この研究で分かったことだが、ニュースがどのくらい深くまで到達するのかを正しいニュース（緑）とフェイクニュース（赤）で比較した結果が A である。深さはフェイクニュースの方がはるかに深くまで到達するということが分かる。一方、B はカスケードのサイズ、つまり、トータルで何人ぐらいの人に到達したかというのを正しいニュースとフェイクニュースで比較したものである。横軸は対数なので、桁違いにたくさんの人にフェイクニュースが届いていることが分かる。

さらに、E はカスケードがある深さに達するまでにどのくらいの時間がかかったかという、伝達の速さを調べたものである。フェイクニュースの方がはるかに速く到達していることが分かる。また、拡散されやすい話題というのも主に政治に関することであることが分かってきている。

図 2-1-3 に示したのは、誰がフェイクニュースを流しているのかを表すグラフである。左のグラフはどれだけのツイッターユーザーが関与したかを累積分布で示したものである。例えば 8 割ぐらいのフェイクニュースというのは、全体の約 0.1%の人で共有されている、ということが分かる。中央のグラフと右のグラフからは、若者よりは高齢者の方が、あるいはリベラルよりはコンサバティブの方がフェイクニュースを見たり共有したりする確率が高いことが分かる。

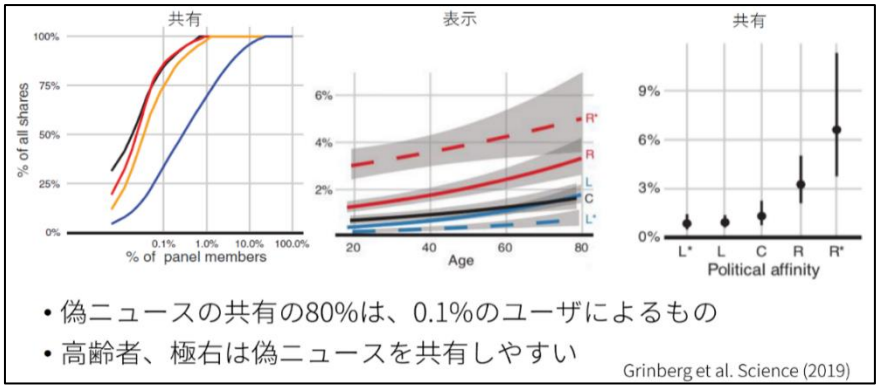


図 2-1-3　偽ニュースを拡散する一部のユーザー

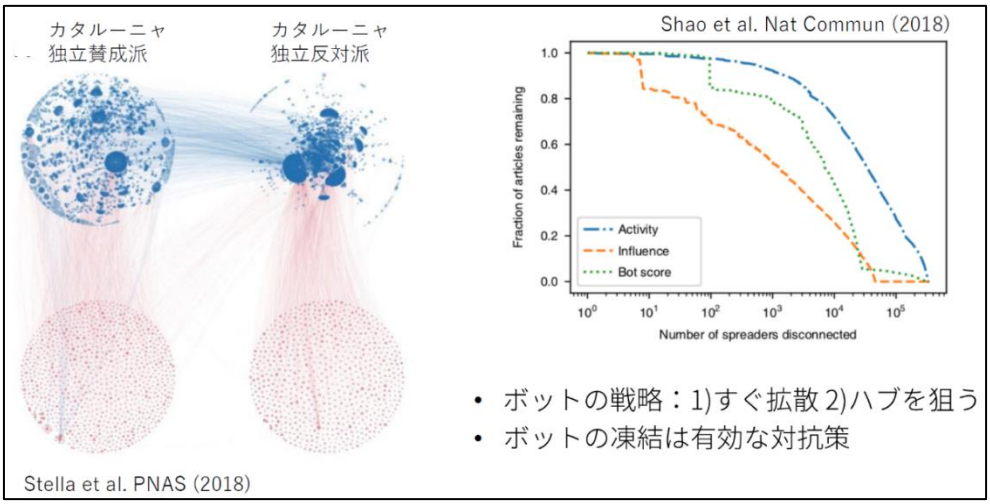


図 2-1-4　偽ニュースを増幅するボット

また、フェイクを拡散するのは人間だけではない。ボットと呼ばれる存在がある。ボットとは計算機による自動発言システムである。図 2-1-4 はカタルーニャの独立に関する選挙期間中のツイートを分析して、リツイートがどのように流れたかを可視化したものである。青が人間による発言、赤がボットによる発言と、アルゴリズムによって判定されたものである。丸の大きさはフォロワーの数を表しているので、独立賛成派にしろ、反対派にしろ、フォロワーがたくさんいる大きい丸、いわゆるインフルエンサーとかハブと呼ばれる人に向けて、ボットが何か情報を発信している、つまり、インフルエンサーを使って何か情報を拡散しようとしている、ということが見てとれる。賛成派のボットはハブに向かって、より独立に関する情報を拡散させようとし、反対派のボットは反対派の人間のハブに向かって、反対だという声を大きくしようとしている、そんなことが見てとれる。さらに、ボットを取り除くとフェイクニュースの拡散を減少させるのに非常に有効だということを示した研究もある。

次に、フェイクニュースになぜだまされてしまうのかという、私たち人間の要因について簡単に説明する。一つは認知バイアスと言われるものがある。ウィキペディアに掲載されているものを数えただけでも 180 個以上ある。情報が多過ぎるとか、意味が足りないとか、時間が足りないとか、自分のメモリーが足りない、といった認知的な理由によって合理的な判断をせずに、つい認知的なショートカットを使ってしまっている。例えば直感だったりとか、経験だったりとか、思い込みといったショートカットである。そういうものは大抵うまく働くのだが、フェイクニュースに関しては直感とかが悪さをして、ついついだまされてしまう。

認知バイアスの代表的なものの一つに、バックファイアというものがある。これは非常に衝撃的な実験で、例えばアメリカのツイッターユーザーに、保守の人に民主党の議員のツイッターアカウントをフォローしてもらって、1 カ月間、そのツイートを見るというような実験を行う。その後、自分の政治的な考え方が変わったかというのを調べてみると、より自分の考え方に強固に固執するようになる、そういう傾向が見えたということが報告されている。全ての話題にこういうバックファイアが起こるわけではないが、間違っているよと指摘すると、かえって意固地になってしまうことがある、というのがバックファイア効果のややこしいところである。

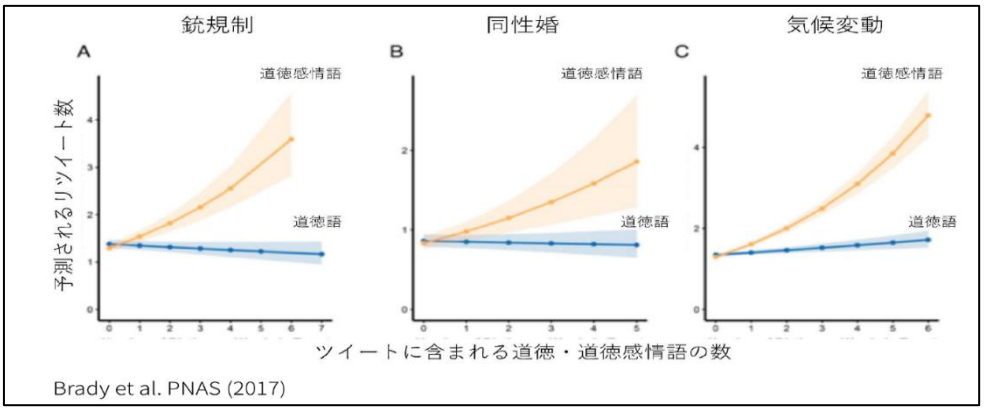


図 2-1-5　道徳感情の伝染

また、感情の問題がある。私たちは感情的な生き物なので、扇情的な書き込みとかを見るとついつい反応してしまう。これもアメリカの例だが、ツイートの中から非常に議論の余地がある話題に関するツイート、具体的には銃規制に関するツイート、同性婚に関するツイート、気候変動に関するツイートを集めてきて、ツイートの中に道徳に関する語が 1 個、2 個、3 個、4 個と含まれていた場合、さらに道徳かつ感情的な語が 1 個、2 個、3 個、4 個と含まれていた場合に、リツイートされる確率がどういうふうに上がるかというのを縦軸に示したものが図 2-1-5 である。いずれの話題でも道徳的かつ感情的な語の数が増えれば増えるほど、リツイートされる確率が上がる、つまり、感情は伝染しやすいということを示している。実際のリアルな社会では、情動伝染という現象が知られているが、オンラインでも感情が伝播するということをこのデータ分析は示している。

次に、私のメインの研究領域である環境要因や構造要因について取り上げる。認知的な脆弱性を持った私たちがプラットフォームと相互作用すると、一体、どんなことが起こるかという研究である。そうなると、見たいものしか見えないような情報環境がおのずとできてしまう。その典型的な例がエコーチェンバーと言われるものである。簡単に言うと、エコーチェンバーとは似た者同士だけでつながった、閉じた情報環境ができる、図 2-1-6 に描かれたような状況である。これは、人は似た者に囲まれて、見たいものだけを見て、つながりたい人とだけつながっている状態である。そうなると自分と違う意見を持つ相手が見えず、自分は仲間しか見えず、自分が多数派だと思い込むようになる。そして、フェイクニュースの温床になるということである。

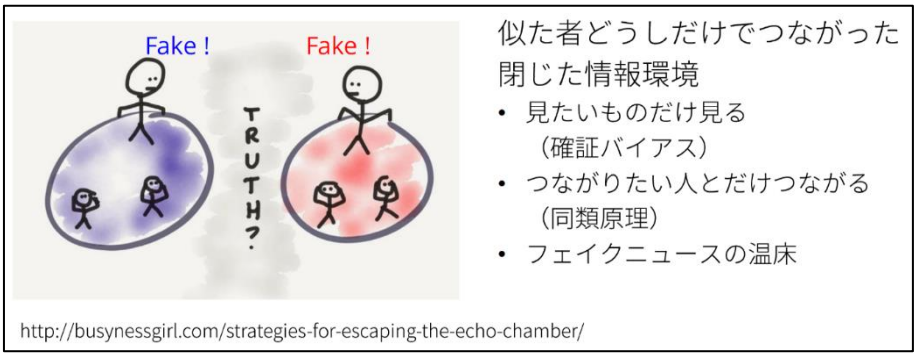


図 2-1-6　集合知を阻害するエコーチェンバー

これはあくまでもコンセプチュアルな説明だが、本当にこんなことが起こっているのかということを分析したのが図 2-1-7 に示した私たちの分析である。アメリカ大統領選 2016 年のツイートを集めてきて、どのようにリツイートされたかということを分析した。青で示されたのがリベラル系に関するハッシュタグが多数含まれているリベラル系のユーザー、赤はコンサバティブ系、つまりトランプ支持派である。リベラル系はリベラル系だけと専らコミュニケーションをとり、逆もまたしかりである。そして、クロスイデオロギーなやりとりというのは、それに比べれば少ない、そういう典型的なエコーチェンバーの構造をしていたということを示している。

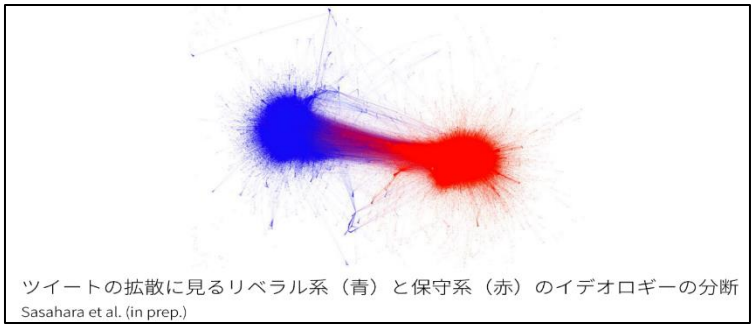


図 2-1-7　2017 年の米大統領選のエコーチェンバー

いったんこういう構造ができてしまうと、フェイクニュースの温床になりやすく、非常に問題である。では、どういうメカニズムでこういう状態に至るのかということを知りたいと思い、簡単な計算モデルを作ってシミュレーションを行った。https://osome.iuni.iu.edu/demos/echo/で公開しているので、もし興味があれば見てほしい。色の違いで意見の違いを表した多数のユーザーがつながっている状態を初期状態としている。ユーザーは自分と近しい意見からの影響を受けてちょっとだけ意見を変えるが、自分とは全く相容れない意見を言うユーザーはアンフォローして、別の人をフォローするという、我々がオンラインでやっていることをコンピューター上の仮想人工社会の中でやってみて何が起こるかを見た。その結果、自分とつながった友人の意見の平均多様性が時間とともに単調減少し、二度と増加することがないことが分かった。先ほど述べたようなダイナミクスを回していくと、やがては自分と自分の周りが似ていくということが起こる。このようなダイナミクスを繰り返していくと、リベラルな人たちはよりリベラルな意見を持つようになり、保守的な意見の人たちはより保守的な意見を持つようになり、意見が極性化するだけではなくて社会的なつながり、ネットワークもだんだん分断されていって、青は青でまとまり、赤は赤でまとまり、最後は完全に分断されてしまう。

もちろん、現実世界はこんな簡単ではないが、現在オンラインで私たちがやっているような行動のエッセンスだけを持ってきてモデル化することで、このモデルの仮定においてはという限定付きではあるが、私たちの社会的なオンライン状態のやりとりの中に、極性化する、あるいは分断するということが最初から埋め込まれているということが言える。

さらに社会的影響の強さと、つながりを切り替える頻度の 2 つでどのくらいエコーチェンバーが加速されるかというのをシミュレーションで調べた。社会的影響が強いほど、またつながりを切り替える頻度が高いほどエコーチェンバー化の速度が速いことが分かった。ソーシャルメディアを利用していると、社会的影響の強さも、社会的なつなぎかえであるフォロー／アンフォロー

の頻度も上がっていくので、結果的にエコーチェンバー化が加速される方向にいる、ということを示している。

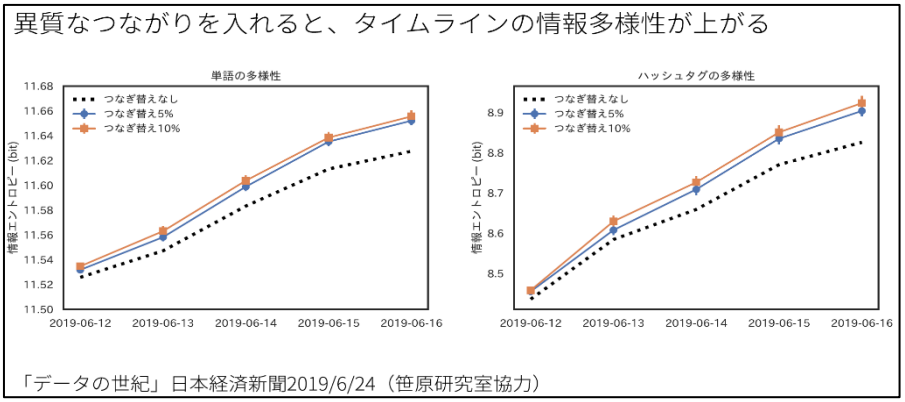


図 2-1-8　エコーチェンバーを緩和する

では、どうしたらエコーチェンバーを緩和できるのかという一つのヒントは、次のような実験から得られた（図 2-1-8）。これは日経新聞の記者がアメリカ民主党の政治家 300 人のツイッターアカウントをフォローして、タイムライン見るときに、少しずつ共和党議員のツイッターアカウントにつながりを切りかえていくという実験である。仮にその切りかえを行わなかった場合の情報の多様性、単語の多様性というのは図 2-1-8 の点線で示されるように、日々、新しい言葉に出合うので、情報量は増えていく。仮にそのつながりの中に 5%とか 10%異質なつながり、つまり、逆のイデオロギーの人を入れると、つながりを変えないベースラインよりは情報の多様性が上がるということを示している。

したがって、普段の SNS を使っている文脈に置きかえると、嫌なやつを自分のつながりの中に混ぜておくと意外な気づきが得られる、エコーチェンバー化を抑制する効果があるかもしれないということである。ただし、嫌なやつを混ぜるというのは、実は結構、精神的なストレスもあるので簡単にできることではないのだが、何とかそこを技術でうまくサポートできたら、エコーチェンバー化は緩和できるのではないかと思う。

2 つ目はフィルターバブルの問題である。（図 2-1-9）インターネットにはほぼ無限に情報があるが、アルゴリズムがユーザーの過去の行為、過去の履歴に基づいて、どの情報がユーザーに関係あるかということを最適化して見せてしまう。最適化の仕方はユーザーごとに違うので、各個人がどんどん異なる世界を見るようになり、孤立していくというのがフィルターバブルのメタファが意味していることである。このようにパーソナライゼーションには負の側面がある。

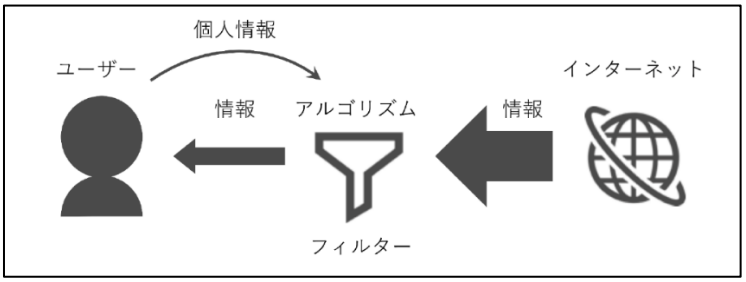


図 2-1-9　フィルターバブル

アルゴリズムはどのぐらい予測できるのかということを調べた研究がある。Facebook のユーザーはいろいろな記事に「いいね！」をするわけだが、どの記事に「いいね！」をしたかということを大量にデータを集めてきて、機械学習させたモデルを使うと、例えば白人かアフリカ系かとか、あるいは性別がどうか、ゲイかどうかとか、あるいは場合によっては薬物を使用したことがあるかどうか、そんなことまで、どの記事に「いいね！」をしたかということでかなり正確に当てられるようになる。

同じようなことを私たちもツイートでやってみたが、当てられないものもあるが、当てられるものもかなりある（図 2-1-10）。例えば、男女かどうかは当てられる。それから、デジタルネイティブかどうか、つまり 1980 年より前に生まれた人か、後に生まれた人か、既婚かどうか、子どもがいるかどうか、あるいは Facebook 上に 150 人以上友達がいるかは、意外と当たる。当てられないのは、例えばイヌ派かネコ派かとか、これは意外と当てられない。

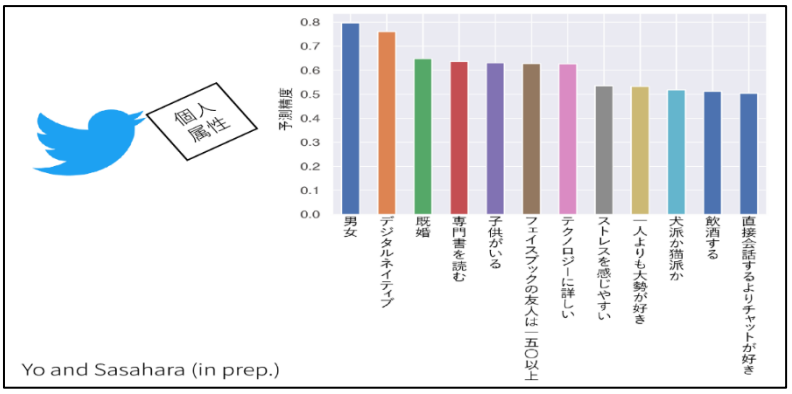


図 2-1-10　ツイートが運ぶ個人属性

フィルターバブルは、確かに潜在的には非常に問題だが、本当に悪さをしているのかというと、実はそこまで悪さはしていないのではないかということを示唆する研究もある。図 2-1-11 は Facebook のデータで研究したものである。例えば保守派の人は、もしも何もしていない状態であれば 40%の確率では自分とは異なるイデオロギーの情報を見る確率がある、あるいはリベラル派の人だったら、もっと高い確率で自分のイデオロギーとは逆の情報を見る可能性がある。

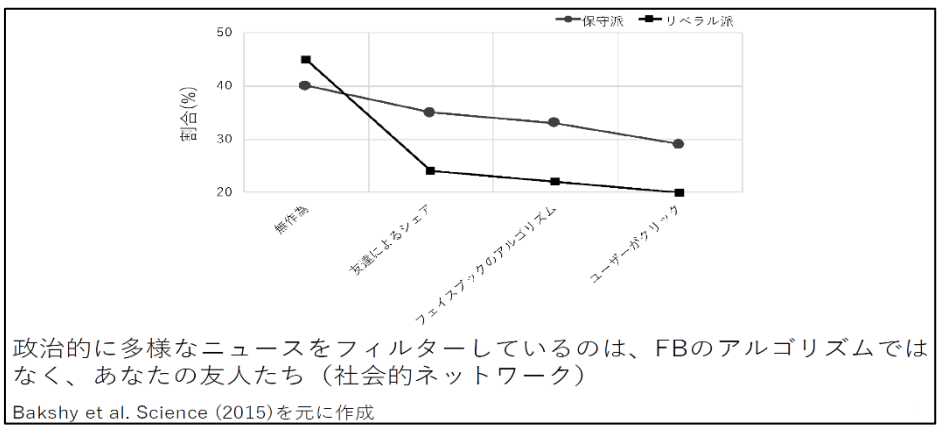


図 2-1-11　アルゴリズムは悪さしない？

ところが、友人がどの情報をシェアしたかという、情報を使うと一気に情報の多様性が減って、その後で Facebook のアルゴリズムがどの情報をどういう順番で見せるかということをやって、さらに多様性が減る、という順番で減っていく。したがって、自分とは異なる意見を見せなくしているのは、Facebook のアルゴリズムではなくて、あなたが誰とつながっているかということが情報の多様性を失わせる最大の原因になっている、ということをこのグラフは示唆している。しかし、これは Facebook の研究者が Facebook のデータを使ってやったものなので、私たちはこれを再現できない。そういう問題もある。しかし、これが仮に正しいとするならば、まだ、現時点ではそこまでフィルターバブルの問題は深刻化していないという可能性がある。

話をまとめると、こうなる。

・虚偽情報に対する要素レベルの脆弱性
  - 認知バイアス、社会的影響
・要素レベルの脆弱性をシステムレベルに増幅
  - エコーチェンバー、フィルターバブル
・虚偽情報拡散の分脈
  - 情報過多、フェイクの自動化・巧妙化（ボット、ディープフェイク）

フェイクニュースをコンテンツの問題と考えると矮小化し過ぎで、情報生態系全体の問題と捉えなければいけないというのが私の意見である。そもそも、情報生態系の要素である我々人間が虚偽情報、フェイクニュースに対する脆弱性を持っている。それは認知バイアスだったり、社会的影響だったり、である。そういう脆弱性を持つ人間がプラットフォームでつながり合うことで、要素レベルの脆弱性がシステムレベルに増幅されてしまっている。その結果生じている副作用がエコーチェンバーだったり、フィルターバブルだったり、そういった問題である。これらは基本的には同じ問題の異なる側面を指しているメタファなのだと思う。

虚偽情報が拡散しやすい文脈というものもある。例えば、情報が多過ぎるので認知バイアスが働いたりとか、あるいはアルゴリズムでフィルタリングしなければいけなかったりという問題がある。情報が多過ぎる問題というのは深刻である。それから、フェイクの自動化とか、ボットなどによる巧妙化とか、この後話があると思うが、ディープフェイクの問題とか、こういった問題は、この後のアメリカ大統領選に向けてより一層深刻化する懸念がある。

では、ポスト真実の時代を乗り越えるために、科学者として何ができるかということを考えなければいけない。私は科学者なので、なぜ人はフェイクニュースにだまされやすいのかとか、エコーチェンバーといった環境の問題などの科学的な知識を情報リテラシーとして教えたり、そういった知識を身につけてもらったり、ということが必要だと考えている。それに加え、社会基盤としてファクトチェックという行為、あるいはシステムを根づかせるということが大事だし、プラットフォームレベルでできることはやるべきである。例えばパターン化されたフェイクというのはどんどん取り除くべきだし、多様なつながりを促進する何かマイルドな工夫というのが必要だと思っている。今日は時間がなくて余り話すことはできないけれども、私たちの研究室では学生たちと一緒に Polyphony というシステムを作っている（図 2-1-12）。

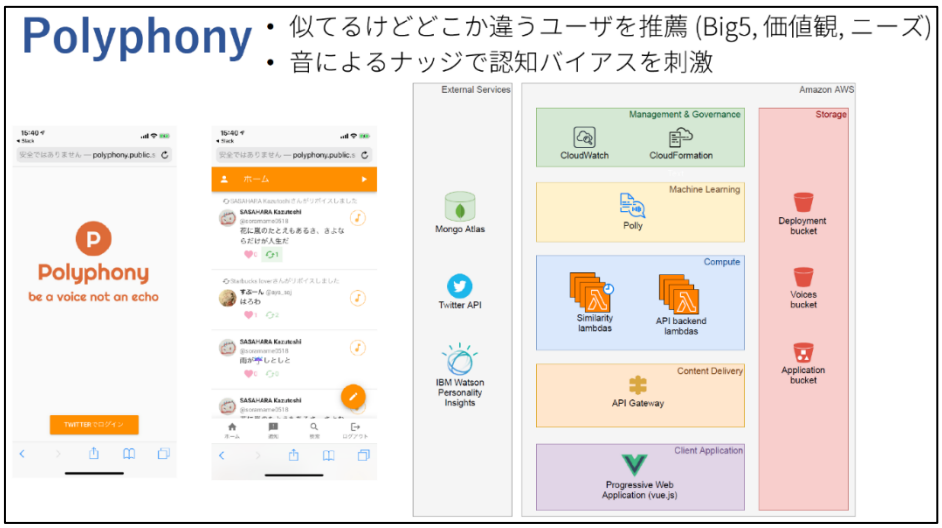


図 2-1-12　Polyphony

これはまだ準備段階だが、放っておくと似た人とつながってしまい、結果として同質化のタコつぼができるというのがエコーチェンバーなので、その逆をやろうというものである。似ているけれども、どこか違うというユーザーのつながり、本来ならばつながるはずのないつながりを促進するような工夫を入れよう、しかも、それを意識的にやるのではなく、音によるナッジで認知バイアスを刺激して、自ら進んでそういうことを選ぶような仕組みがシステムの中に埋め込まれているようなものを作ろうと、今、目指しているところである。

質問応答

Q：道徳・道徳感情語の例を教えていただきたい。

A：道徳語というのは Jonathan Haidt の道徳基盤理論（Moral Foundations Theory）というのがあり、その研究グループが辞書を作っている。単語は、5 つの道徳基盤のどこに属するということが定義されている。それとは別に、LIWC という心理学で使われている単語の心理学的な属性を調べるソフトウエアがあり、そこにはある単語がポジティブかネガティブかということが記載されているので、これらを合わせると道徳的でかつ感情的というのが分かる。具体的な例はすぐには思い浮かばない。

Q：SNS でのニュースの拡散、例えばボットがインフルエンサーを狙う、といったことは可能か。

A：図 2-1-4 が、ボットがインフルエンサーを狙うことができる、ということを示した図である。アルゴリズムでやっているのか、人がパペットのように操っているのかは分からない。

Q：SNS 上のニュースの拡散、ツイートの拡散はどういうアルゴリズムでやられていて、透明性、公平性、そういうものがあるのかどうか、あるいはそれは SNS 運営会社のアルゴリズム任せになってしまっているのか、そういったツイートの拡散の仕方は現状どうなっているのか。

A：ツイッターの場合だと、情報の拡散には必ずしもフォロー・フォロワー関係は必要ない。したがって、基本にはフォロワーが多いユーザーが起点になる、あるいはそのユーザーを介すことがカスケードを増幅させるのに寄与していることが多いが、話題や内容と、ハブユーザーを間に介すかどうかの掛け算で決まるのではないかと思う。

Q：対策を示していただいたが、異質なつながりを入れてエコーチェンバー化を防ぐと言いなが

ら、実はそうすることによってバックファイアが起こり、自分が余計、自分の意見に固執してしまったり、虚偽情報を抑制するところ自体にバイアスが入ってしまったりしないのか。

A：図 2-1-8 は、本当にそういうことに意味があるのかということを意図的に調べたわけだが、そこにバイアスができるだけ入らないようにというのが私たちのシステムのデザインである。しかしながら、後ろでは、ビッグファイブとか、価値観とかニーズというのをその人の投稿から測るところで、たぶんこの人とこの人はつながる可能性があるといったことをアルゴリズムで測っている。したがって、そこにバイアスがないかと言われれば、なかなか、ないとは言い切れない。しかし、少なくともこのシステムでは、推薦を受け入れるかどうかの最後の意思決定は本人がするので、別にシステムがそういう邪魔をしても、本人が嫌だったら無視できるという意味では、いいのではないかと思っている。

2.2 フェイク動画問題：メディア解析技術によるアプローチ

山岸順一（国立情報学研究所）

ディープフェイクについて、どういう対策ができるか！？どのように検出できるか！？私たちの取り組みを紹介する。

まず最初に、本発表におけるフェイクが何を意味するのかを定義しておく。本発表では、機械学習もしくはコンピューターによって生成されたメディア（デジタル・クローン・メディア）をフェイクと呼ぶ。つまり、動画が表す中身ではなくて、映像そのもの、音声そのもの、テキストそのものがフェイクかどうかを対象とする。

関連分野としては、生体認証（バイオメトリック認証）がある。フェイク映像やフェイク音声は、実は以前から生体認証分野には存在しており、例えば、顔のマスクを作って顔認証システムをだましたり、指紋をシリコンジェルでとって指紋認証システムにかけるような成り済まし攻撃（Presentation Attack）と言われるものがあった。この攻撃は、最近では機械学習によってさらに高度化し、あたかも本人そっくりのような顔映像や音声が簡単に作れるようになってきたことから、生体認証システムだけではなくて、人間に対しても攻撃できる状況になってきた。

本発表では、コンピューターが生成した人間のような音声と人間の音声をどう識別できるか！？という話をした後、顔のフェイク映像について話をする。そして、最後に、Amazon 等のショッピングサイトでのレビューをコンピューターで作ってフェイク情報を数多く載せることは可能か！？、もしフェイクレビューがあったとしたら、どのように見分けられるのか！？という話をする。

まず最初に、フェイク音声の話をする。これは、音声合成という分野が関係している。音声合成というのは、与えられた文章から音声波形を生成して提示する技術であり、2017 年 12 月にチューリングテストをパスしている。つまり、人間の耳では、どちらが人間の声か合成音声かは分からない状況にある。この分野にはさまざまなプレイヤーがいて、GAFA（Google、Amazon、Facebook、Apple）を初め、BAT（Baidu、Alibaba、Tencent）等、物凄い勢いで研究が進展している。ある声優の声だけでなく、さまざまな人の声を合成できるようになったり、アニメ『名探偵コナン』の蝶ネクタイ型ボイスチェンジャーのように、誰かの声をリアルな他人の声に変換することも可能になりつつある。このことにより何が困るかというと、合成音声で人間をだますことに使われるだけでなく、その他、例えばスマートデバイスに向けて合成音声を提示して個人

認証を突破することにも使われることだ。もちろん、これを防がなければいけないというニーズが今高まっており、重要な研究課題となっている。防御側の技術は PAD（Presentation Attack Detection）と呼ばれ、ISO の国際標準化が済んでいる。

もう一度、今の状況を図 2-2-1 に分かりやすく説明する。皆さんがお使いの iPhone に、皆さんが「Hey Siri、今日の天気は何？」と聞くと、最初の「Hey Siri」という短いフレーズで個人認証をかけている iPhone が、登録ユーザーだと判定したときだけ音声認識を実行し、皆さんがいる地域の天気をチェックして、「今日の天気はいいです」とか「暑いです」という文書を生成して、それを読み上げることをやっている。

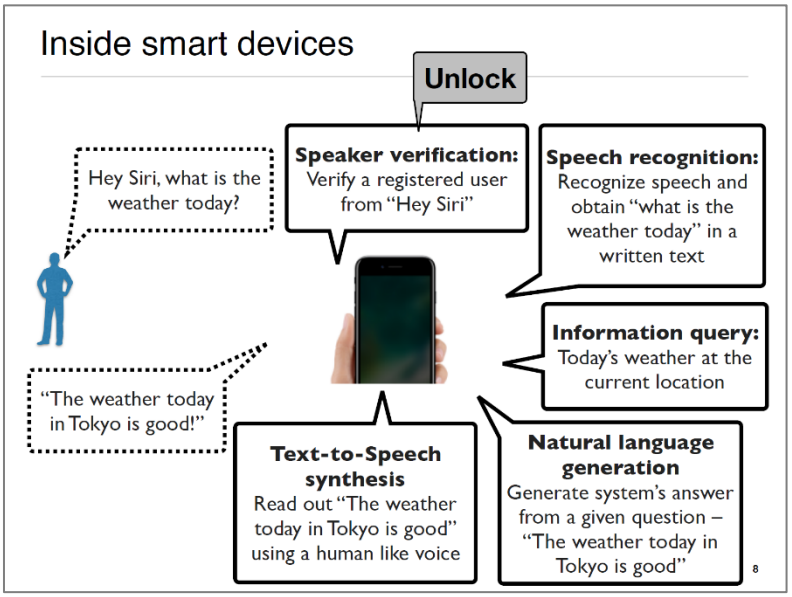


図 2-2-1　スマートデバイスの内部

このようなスマホに対する攻撃としては、音声合成や声質変換を使って、悪意のある人が皆さんのスマホに向かって、皆さんがいないところでフェイク音声をこっそり再生するような攻撃が考えられる。例えば、悪意のある人が皆さんの声を少しだけ入手しておけば、「Hey Siri、110 番通報をして」という声を合成できるので、iPhone の個人認証ロックが解除されてしまい、皆さんの知らないうちに 110 番通報をしているということが起こり得る。そして、この後、「放火しました」と言われたら、大変なことになってしまう。

その他、どのような悪用ができるかというと、非常に怖いことに、「Hey Siri、お母さんに電話して」と言って実際の息子の iPhone から電話をかけて息子に成り済ますことも考えられる。つまり、他人のスマホではなく自分の息子の iPhone から電話がかかってくるものの、実際に電話をかけているのは別人で、息子のフェイク音声を使っているという、非常に高度な成り済ましが起こり得る。あるいは、「Hey Siri、virus.com を開いて」といった攻撃もあるかもしれない。この virus.com というのにはウイルスが仕掛けられていて、ブラウザーを開くと何かのウイルスをダウンロードしてしまうのだ。また、「Hey Siri、ツイートをして」というフェイク音声で、本当は自分がツイートしていないのに、勝手にツイートされてしまうようなことも起こり得るかもしれない。

私たちは、これを防ぐことが非常に大事であると考え、この問題に 2013 年頃から取り組んでいる。指紋認証に対する PAD は以前から研究されていたが、音声に対する PAD の研究分野がなかったので、分野自体を作るようにした。2015 年から、まず標準データベースを作り、セキュリティ分野の研究者に配付して、みんなで認証性能を競い合うコンペティション（ASVspoof Challenge）をずっとやっている（図 2-2-2）。ここで、ASVspoof Challenge とは、Automatic Speaker Verification Spoofing and Countermeasures（ASVspoof）に関するコンペティションであり、2015 年には 16 組織、2017 年には 49 組織（そして 2019 年には 63 組織）が世界中から参加している。

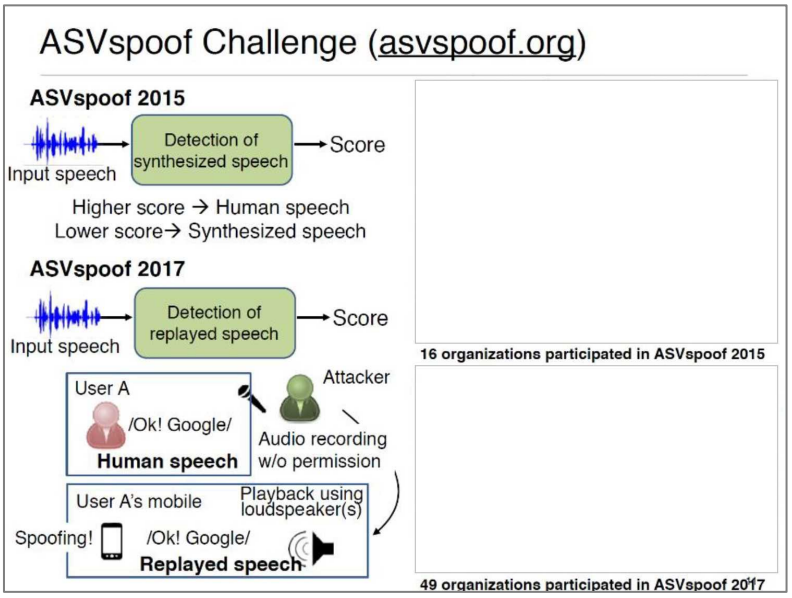


図 2-2-2　音声認証性能コンペティション

私たちが提唱している PAD は、敵対的生成ネットワーク（Generative Adversarial Network: GAN）と何が違うのか！？を、図 2-2-3 に説明しておく。GAN というのは教師なし学習モデルの一つで、（訓練データと近いデータを生成する）ジェネレーター（Generator: G）を学習するときに（生成データの正否を判定する）ディスクリミネイター（Discriminator: D）の識別手法を使うという枠組みである。G と D がどちらもお互いの内部を知り合っているような状況であり、D がリアルかフェイクかという判断をする。

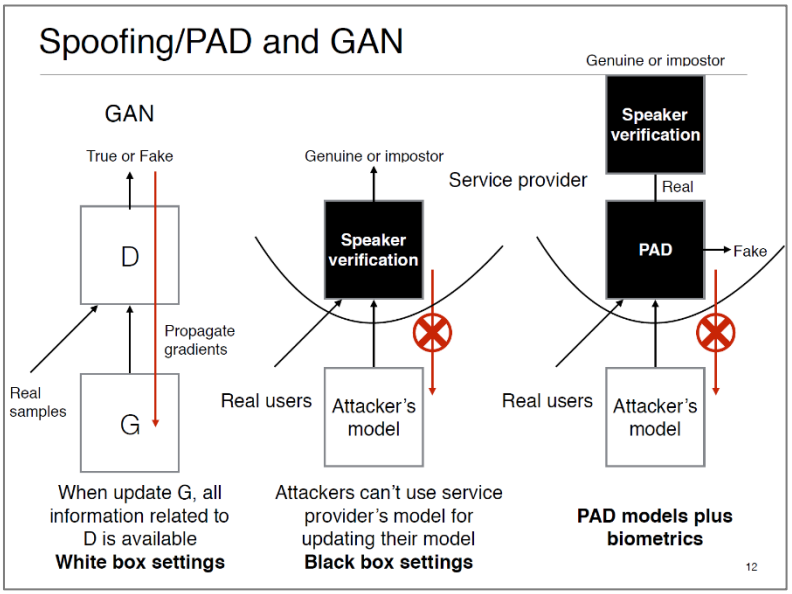


図 2-2-3　スプーフィング/PAD と敵対的生成ネットワーク（GAN）

PAD とスプーフィング（成り済まし）とは非常に似ているが、お互い中身を知り合っているというホワイトボックス・セッティングになっている。一方、私が今まで説明してきた成り済ましと、その成り済ましの検知は、状況が違う。まず、防御側が完全にブラックボックスになっていて、攻撃者からは中身が分からない状況にある。さらに PAD と個人認証の 2 つのモジュールがある状況になっている。皆さんの iPhone も、実はこのような状況にある。

私たちが考えているシナリオはさらにもっと複雑で、下記のような状況を想定している。攻撃者は一人ではなく、さまざまな攻撃があると考える。さらに新しい攻撃が出てくるかもしれないので、未知の攻撃も考える。そして、防御側も複数の防御モデルアルゴリズムを使う。この防御側と攻撃者側は、ともに中身は知らないという条件で ASVspoof Challenge をやり、どれだけフェイクを検出できるかというタスクを競い合う。具体的には、まず、企業（Google、NTT、iFlytek 等）の協力を得て、彼らが持っている合成音声を提供してもらい、それと自然音声を組み合わせた大規模データベースを作成する。そして、その大規模データベースを約 150 の世界中の組織に配布して、与えられた音声から、人間の音声かコンピューターが生成した音声かを識別してもらって、識別性能を私たちがランキングする。

繰り返し言うが、音声合成技術はチューリングテストをパスしているので、合成音声か人間の音声かは、人間が聞いても分からない。そこで、2 つのエラーとして、人間の音声を機械と誤ってしまう際のエラー（FRR: False Rejection Rate）と、機械の音声を人間と誤ってしまう際のエラー（FAR: False Acceptance Rate）を考え、この 2 つのエラーが等しくなる（FAR=FRR）点として等価エラー率（EER: Equal Error Rate）という基準を定義して、評価指標として私たちは使っている。そのためには、図 2-2-4 に示すように、まず、人間の生体音声のスコア分布と機械音声のスコア分布を、ある閾値で切らなければならない。そして、この 2 つのエラーを閾値を変えながらプロットすると、図のような DET（Detection Error Tradeoff）カーブが描ける。しかしながら、人間の音声か機械の音声かを識別しただけでは個人認証を突破できなくて、後段にもう一つ機械があって個人認証をしているので、そのエラーも考慮する必要がある。

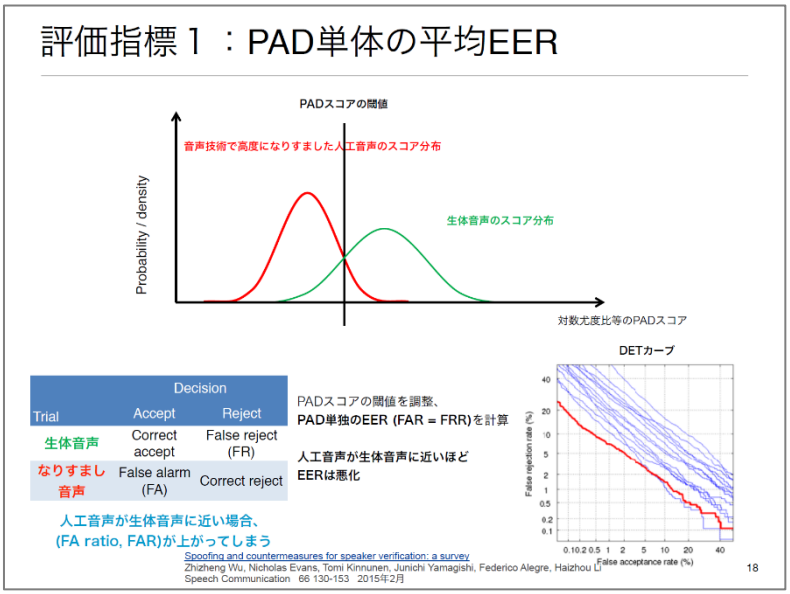


図 2-2-4　評価指標 1：PAD 単体の等価エラー率（EER：Equal Error Rate）

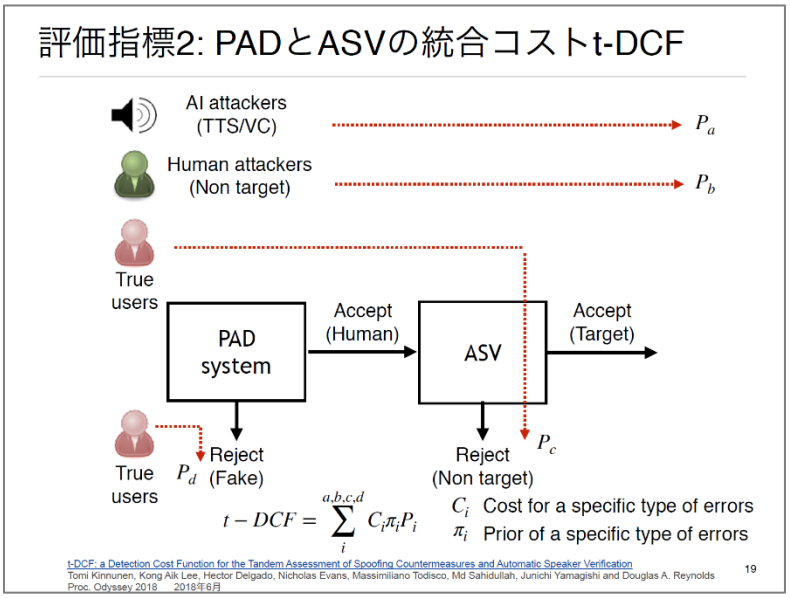


図 2-2-5　評価指標 2：PAD と ASV の統合コスト t-DCF
（tandem Detection Cost Function）

実際には、合計 4 種類のエラーとして、(1)人間が PAD で間違ってリジェクトされてしまうケース、(2)真のユーザーが個人認証でリジェクトされてしまうケース、(3)AI ではなく、他の人間が間違って受け入れられてしまうケース、(4)AI の声が受け入れられてしまうケースがあり、それぞれのエラーの発生確率 P_i は全然違う。例えば、AI によるやり方は結構コストがかかり（他に比べて）発生確率は低いと考えられるので、全ケースについて、π_i（Prior of a specific type of errors）で表して考慮する。また、それぞれに対する起きたときの被害額（コスト）も違うので、それぞれに対するコスト C_i（Cost for a specific type of errors）も別に定義しておく。そして、最終的には、図 2-2-5 に示すように、それらを掛け合わせた（重みづけした）コストの総和を、

統合コスト t-DCF（tandem Detection Cost Function）として、評価指標に使う。

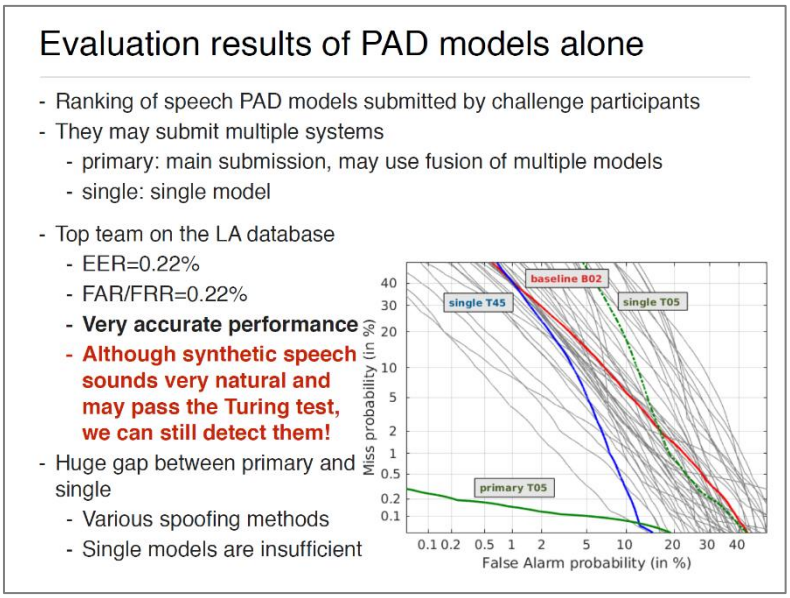


図 2-2-6　PAD 単体の評価結果

コンペの結果を図 2-2-6 に示す。合成音声か人間の音声かは人間が聞いても分からないが、PAD 手法により EER=0.22%という高精度で識別できている。つまり、言い換えれば、人間が聞いているポイントと機械が聞いているポイントは全く違っていることが分かる。ただ、1 位のシステムと他のシステムの性能には結構ギャップがあって、しっかりした技術を使わないと破られてしまうこともある。また、強い攻撃手法も存在するので、それに対する早急な対応も必要である。以上が、ASVspoof Challenge で明らかになった有用な情報である。

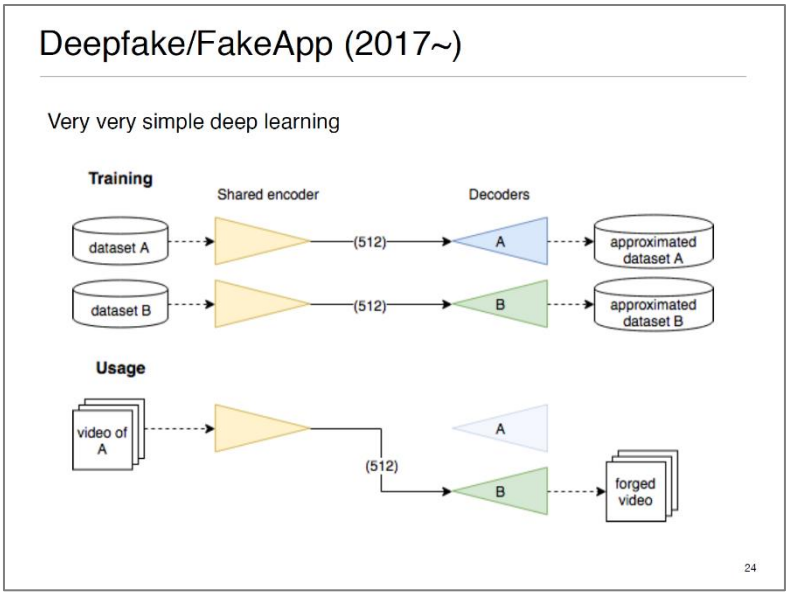


図 2-2-7　ディープフェイク（Deep Fake）動画の作成

次に、ディープフェイク（Deep Fake）の話をする。最近、アメリカのニュースには、コンピューターが生成したあたかもリアルなようなフェイク動画が、選挙に影響を及ぼすのではないか！？という記事が数多く上がっている。

このディープフェイク動画は、実は非常に簡単な機械学習で作られることが多い（図 2-2-7）。顔を置きかえたい人（Bさん）、顔を置きかえる元の人（Aさん）、それぞれの顔画像を数千枚ほど集めてきて、オートエンコーダーというモデルを作る。最初のエンコーダーではAさんの顔のパーツを抽出して、非常に低次元の空間に射影したり、顔を再構築するというようなことをやる。それをBさんについても行い、オートエンコーダーを 2 つ作る。顔を置きかえるときは、まず、元となる人の顔画像をこちらのエンコーダーに入力して顔のパーツを抽出、この特徴点（潜在変数）を今度はBさんのデコーダーの方に入力することでBさんの顔を合成する。

問題なのは、このオートエンコーダーモデルが簡単なアプリとして出回っていて、その結果、ディープフェイク動画が普通の人でも簡単に作れてしまうことである。有名な例としては、トランプ大統領の顔がディープフェイクで顔が置きかえられ、トランプ大統領の意に反することを話すフェイク動画があり、Trump | Deepfakes Replacement というタイトルの動画が YouTube でも見ることができる（https://www.youtube.com/watch?v=hoc2RISoLWU）。

そのほか、よく問題に上がるのが、（話している人物の表情をリアルタイムで再現し、別の人物の顔にマッピングすることができる）Face2Face という技術であり、2016 年 6 月に開催された IEEE Conference on Computer Vision and Pattern Recognition（CVPR 2016）で発表された。研究とともに発表された動画には、Face2Face 技術を使って、ニュース放送の中に映っている第 43 代アメリカ大統領 George Walker Bush の表情に別人（である研究者）の表情をリアルタイムでコピーする研究映像が紹介されている。恐ろしいことに、別人が動かしている表情が Web カメラでキャプチャーされて、テレビニュースの中の人の顔の映像にリアルタイムで埋め込まれている。なお、この動画は、YouTube に、Face2Face: Real-time Face Capture and Reenactment of RGB Videos (CVPR 2016 Oral), https://www.youtube.com/watch?v=ohmajJTcpNk として、アップロードされている。

一方、私たちは、先ほど紹介した音声の PAD 技術を使ってフェイク動画を見破る研究を行っている。まず、顔の部分を抽出して VGG19 という畳み込みニューラルネットワークで特徴を抽出、その後、カプセルネットモデルで識別を行って、最終的にリアルかフェイクかという判定をしている。判定精度は、EER=1.42%なので、98～99%程度の高い精度で識別できている。アメリカの有名な歌手 Taylor Swift の顔が Face2Face 技術によって勝手に加工されたフェイク動画も存在する。口元をよく見るとフェイク動画の口の動きが不自然であるが、これも簡単に PAD 技術を使ってフェイク動画を見破ることができる。動画の圧縮の有無で性能は違ってくるが、圧縮がなければ約 99%の精度で判別することができる。今のところ、軽く圧縮がかかっていると 2%程度までエラーが増え、かなり圧縮が強い条件ではエラーは約 17%に増えてしまうのが現状である。

最近、行っているのは、リアルかフェイクかを検出するだけではなく、どこが加工されているかを見破る研究開発を行っている。フェイク動画を見破るとともに、同時にどこが加工されているのかを推定してマスクで表す技術である（H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, The 10th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2019)）。この技術の利点は、動画を見ている人に加工された怪しい部位を形で見せること

ができることであり、その部位がマスクの形をしているのならおそらく Face2Face 技術を使って加工されたことが専門家には分かる状況にある。このように言うと、フェイク動画は簡単に見破れると思うかもしれないが、実はそうではなくて、攻撃手法が既知だから見破れるのである。先ほどのモデルを未知の攻撃手法に対して使えるかどうかという実験を行ったところ、残念ながら上手く検出はできないので、まだ、フェイク動画問題を完全に解決できたわけではない。もちろん、未知の攻撃データに対して、少しでもデータがあれば、モデルのファインチューニングが可能となり、モデルを新しい攻撃に特化させることでフェイク動画を上手く識別できるようになる。しかしながら、未知に対する攻撃が課題として残っている。

最後にフェイクレビューの話をする。これは先ほどのメディアを生成するという話と一緒だが、個人認証をターゲットにしているのではなく、Amazon 等のショッピングサイトや、レーティングサイトがターゲットになっている。私たちは、実際に人間が書いたレビューを元に似たような文章を大量に生成して、Amazon 等の Web サイトにアップロードすることで、Web サイトの評判を上げたり、下げたりすることができるのではないか！？ということを検証している。そして、そのような攻撃に対する防御方法も探求している。

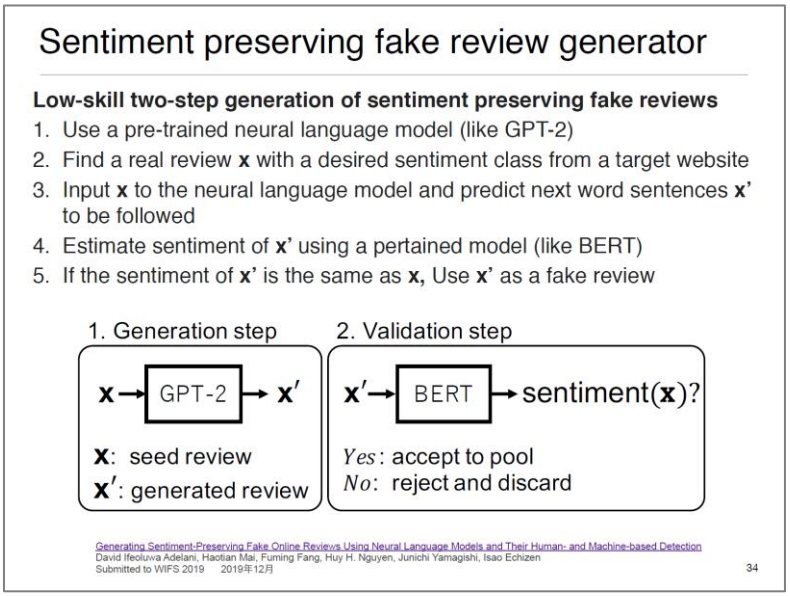


図 2-2-8 フェイクレビュー・ジェネレーターを使った研究

私たちは、図 2-2-8 に示したように、非常に単純な攻撃をベースとして GPT-2（Generative Pre-trained Transformer-2）という学習済みニューラル自然言語モデルを使ってフェイクレビューを数多く作って研究している。この GPT-2 は非常によくできていて、同じレビューを入力しても毎回違う結果を出力し、どれもあたかも人間が書いたかのように見える。次のステップでは、BERT（Bidirectional Encoder Representations from Transformers）という自然言語処理モデルを使って、生成されたテキストのポジティブ・ネガティブ具合の自動判定を行っている。BERT に生成された文章を入力すると、ポジティブ・ネガティブ具合が判定されて、所望のレベルのポジティブ・ネガティブ具合を伴った文章を選ぼうとする。それを攻撃に使えるのではないかと考えている。

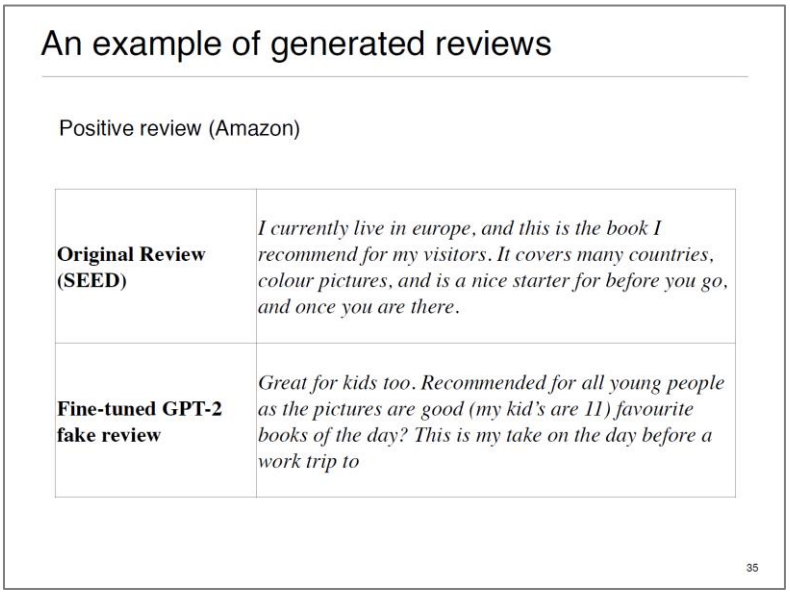


図 2-2-9　Amazon のデータベースを元に行ったフェイクレビューの作成

Amazon のデータベースを元に行ったケース（図 2-2-9）では、人間が書いたオリジナルレビュー（シード）を GPT-2 に入力すると、このような（英語としても破綻していない）文章が自動生成された。YELP というレストランレビューサイトのレビューを元に行ったケースでも、非常に人間が読んでも不思議さがない文章が GPT-2 で生成された。実際、人間である被験者 80 人に対して、四択式（チャンスレベル 25%）の実験をしてみたところ、ネイティブな被験者、ノンネイティブな被験者、それぞれ正解率は約 25%であり、コンピューターが生成したレビューなのか！？それとも人間が書いたレビューなのか！？を見分けられないことを示唆する実験結果が得られた（図 2-2-10）。

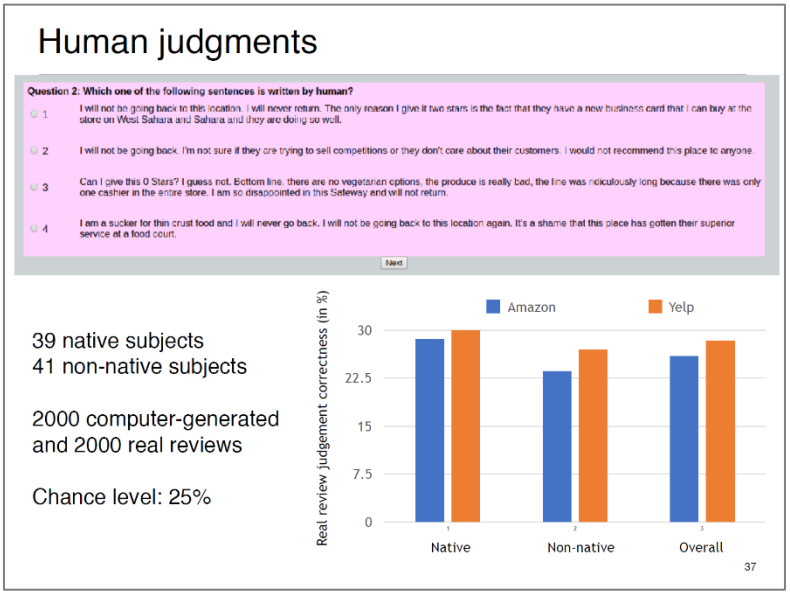


図 2-2-10　人間によるフェイクレビューの判定

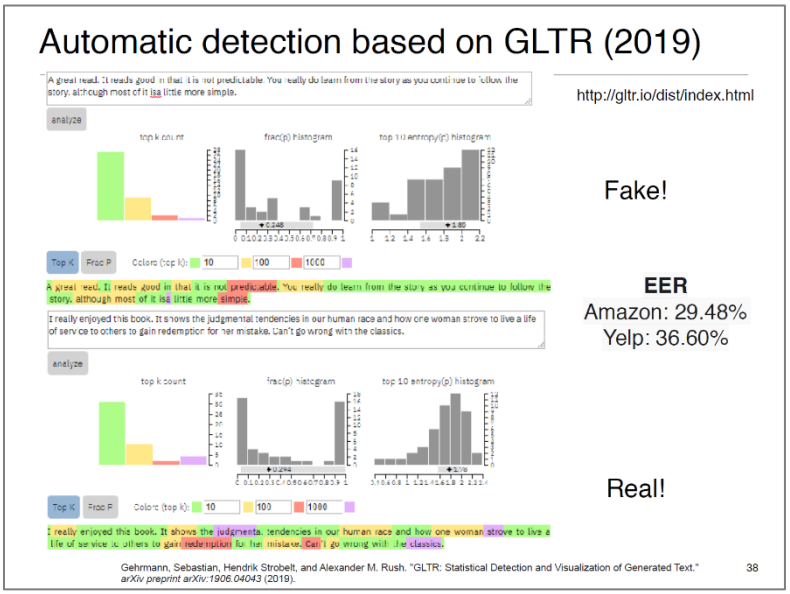


図 2-2-11　GLTR をベースとするフェイクニュースの自動検出

このようなフェイクレビューの自動検出の可能性について、今回、2 つの手法を使って、検証を行った。まず、（Web サイト上から約 4500 万もの文章を読み学習している）GLTR（Giant Language model Test Room）というシステムを使って、検証した。文章を GTLR に入力すると、文章中の単語ごとに予測しやすいかどうかを、緑色（最も予測しやすい単語）、黄色（中程度に予測しやすい単語）、赤色（予測が難しい単語）、紫色（最も予測が難しい単語）の 4 色別に表示してくれる（図 2-2-11）。言語モデルから生成される文章は非常に人間が書いたかのような流暢な文章であるが、よくある単語のつながりからなっているので予測しやすく、緑色で表示される。一方、人間が書いた文章には、前の単語系列からは予測できないような単語が含まれているケースがあるので予測が難しくなり、緑色、黄色などさまざまな色が混在して表示される。紫色で示される単語は、言語モデルからはあり得ない単語系列であると判定をして、これを根拠にリアルか、フェイクかを見破ることが可能である。ただ、実際のところ、この手法の性能はあまり良くなくて、70%程度の正解率（エラー率 30%）である。

もう一つ、言語モデルが生成したフェイクニュースを自動検出する GROVER（Generating aRticles by Only Viewing mEtadata Records）というアルゴリズムがある。フェイクニュースには、人が書いたフェイクニュースもあれば、コンピューターが自動生成したフェイクニュースもある。GROVER はコンピューターが生成したフェイクニュースを見破る技術ではあるが、フェイクレビューの自動検出に向けた研究を試行している。まだ、モデルが公開されていないので、EER を計算することはできなかったが、私たちの印象としては、ある程度は検出できたと感じている。

以上要するに、フェイク音声、フェイク動画、フェイクレビューに関して、生成側の研究は山ほどあるが、反対の防御側の研究はほとんどない。このような状況を鑑み、この分野、特に防御側の研究を広げ、発展させていく必要がある、と私たちは日々思っている。

質問応答

Q：フェイク映像はリアルタイムでやっていたが、音声のスプーフィングもリアルタイムで可能

か？

A：手法によるが、リアルタイムでできるものもある。

Q：圧縮したフェイク画像だと判定が多少難しくなるという話があったが、解像度が粗くなると多少難しいということはあるか？

A：解像度に関する評価をしていないので、正確に答えることはできないが、検出精度は悪くなると予想する。

Q：GAN でフェイク画像を生成するときのディスクリミネーターに PAD を使うと、もっとすごい技術が出てくるのではないか？

A：実は、それが私が説明したかったことだ。GAN と PAD は組み合わせられるものではない。GAN はジェネレーターとディスクリミネーターは、ともにお互いを知り合っていないとできない。一方、私たちの設定では PAD は完全に攻撃者からはブラックボックスになっている。したがって、ジェネレーターと PAD を組み合わせるというようなことは、通常は考えられない。内部に犯行者がいるような状況でない限りは、起こり得ない。

2.3 ネットの誤情報に対する自然言語処理からのアプローチ

乾健太郎（東北大学）

東北大学で自然言語処理の研究をしている。本日は、自然言語処理の立場から、最近、私の周りで実施していること、支援していることを紹介する。

私自身は、ネット情報の分析に自然言語処理の立場から携わっている。SNS のような誰でもがネットに情報を配信できるような時代になっているわけだが、だいぶ前から既に、ネット情報の信頼性が大きな問題になっていた。そこに、自然言語処理として、何かできることがないかという研究をいくつか実施してきた。例えば、コラーゲンが本当に肌にいいかということをネットに聞いたら、ネット上にさまざまなことが書かれている。それを、賛否両論を併記するような形で収集、整理するという研究を 10 年以上前にやっていた（図 2-3-1）。

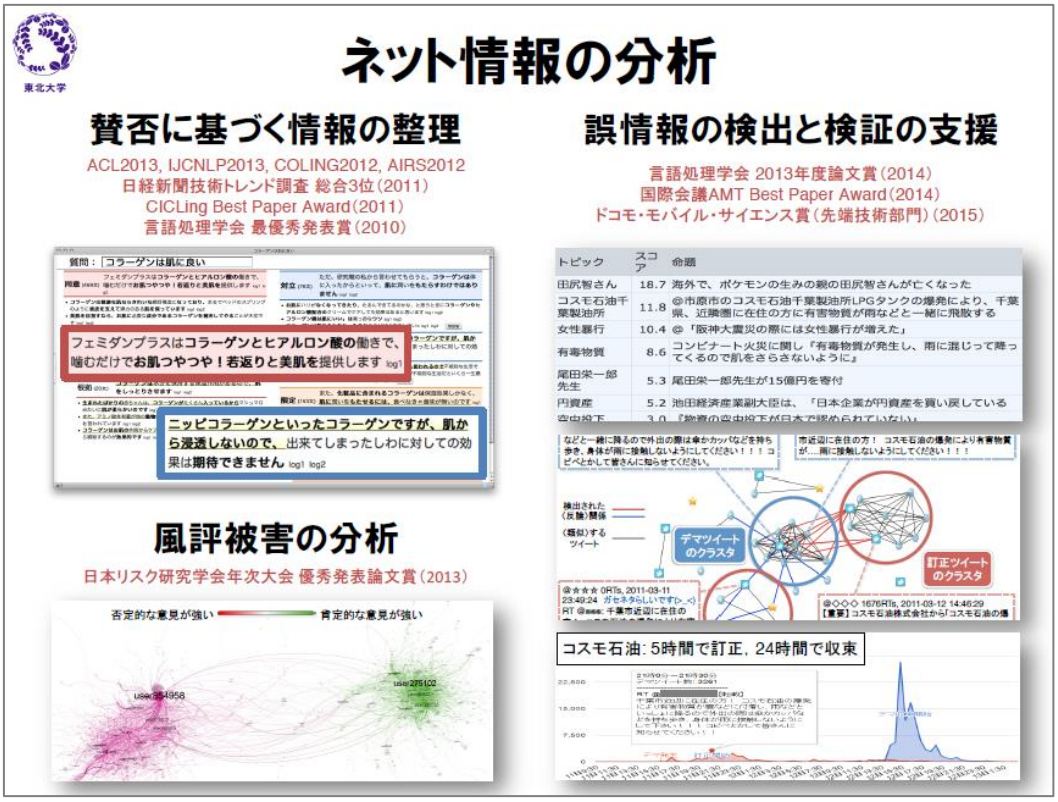


図 2-3-1　ネット情報の分析

そうした中で東日本大震災が起こり、その直後にたくさんのデマがツイッター上に流れた。そうした現象に対して自然言語処理でいくつかのアプローチを実施した。例えば、賛否両論を検索するということについては、デマに対して肯定的なツイートや否定的なツイートを集めることができる。よく反論されているような命題を集めることによって、その時点で出回っているデマをほぼ自動的に集めてくることができる。

また、先ほど笹原先生が拡散と収束の話をされていたが、まさに震災直後のツイッター上で同じことが起こっていた。あるデマがある時点で急激に拡散していくと、それに対して、いや、そんなことはない、今のはデマだとかの発言が、後からだんだん盛り上がってきて、そして、そのデマが収束していくということも確認できた。自然言語処理の技術を使って識別すると拡散・収束のさまざまなパターンが見えてくる。震災後、1 カ月の間に拡散した 60 個程度のデマについて大規模な調査をした。重要なことは、よく広がったデマというのは、必ずどこかで、それはデマだよ、それはおかしいよと、訂正するようなツイートが発信されて、それが拡散していってデマの収束につながっていくということである。

必ずしもデマとは限らないが、時間切れの情報、ある時点から先は誤情報になってしまうという場合も多くある。その場合も、人々が、時間切れとか、デマだと言い始めて収束に向かっていく。SNS を使っている人そのものがある種の社会のセンサーになって、誤情報、デマを認知して、それを拡散することが起こっている。笹原先生の話にあった通りである。

こういうことを調べていく中で、フェイクニュースの問題に対して、ネット上のファクトチェックをする必要があるということで、ファクトチェック団体を支援するファクトチェックイニシ

アティブ（FIJ）という NPO が立ち上がった（図 2-3-2）。私自身もその設立に関わった。その話をする。

ファクトチェックというのは、ネット上に、あるいはネット上でなくてもいいが、単なる意見ではなくて、事実の言説として述べられているものに対して、それが本当に事実なのかどうかということを、より広い根拠情報に基づいて、より事実である可能性が高い、あるいは事実である可能性が低い、疑わしいということを調べて、それを根拠情報とともに記事にしてネット上にまた広めていくというような活動である。

図 2-3-2　ファクトチェックへの技術支援

インターネット上の情報から記事化までのイメージを図 2-3-3 に示す。ネット上にさまざまな情報があるわけだが、そこから、どうもこの記事は怪しい、この言説は怪しいというようなものを集めて、それをエスカレーションしていって、本当に怪しいもの、本当に調査する価値があるものを予備調査、本調査し、最後に記事にしていく。

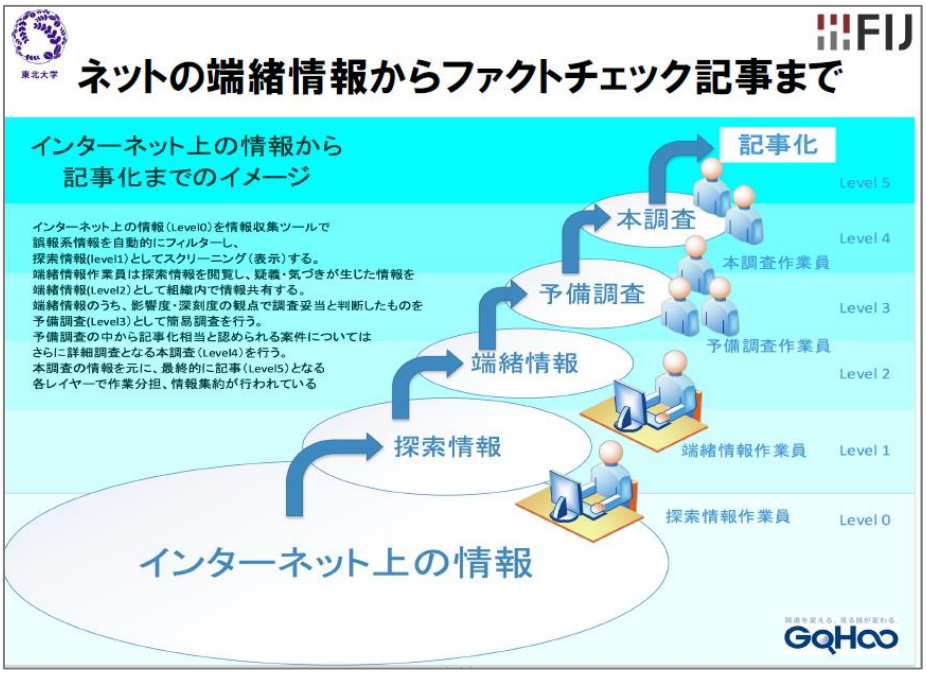


図 2-3-3　ネットの端緒記事からファクトチェック記事まで

怪しそうな記事の集め方について説明する。理屈は、先ほどのデマを訂正するツイートがたくさん流れるということと同じであり、ニュース記事に対してさまざまな人たちが SNS 上で怪しいと言う。ところが、怪しいと言っているツイートなり、メッセージを集めてくるのはなかなか大変である。例えば、「デマ」とか「誤報」というキーワードで検索しても、ほとんどはファクトチェックの役に立たないツイートばかりで、実際にファクトチェックに有用な情報は、わら山の中から針を見つけ出すような骨の折れる仕事である（図 2-3-4）。

図 2-3-4　ところが、ツイートのほとんどはファクトチェックに役立たない

現在は、ファクトチェックイニシアティブと、東北大学の乾・鈴木研の技術チームと、SmartNews とが組んで、ファクトチェックのプロセスを技術的に支援する仕組みを作る活動をしているところである。

実際に今できていることはまだまだ素朴なレベルである。いろいろなニュースに対して、さまざまなコメントがツイッター上、あるいは別の SNS 上で流れている。これらを自然言語処理で解析して、多くの人が疑っているような記事をランクの上位に上げていき、逆にそうではなさそうなものをランクの下の方に下げ、日々、入ってくる大量のニュース記事に対してランキングする。上位にランキングされた記事を人間が見て、その中からファクトチェックするべき記事をピックアップして、それを人間が判断する（図 2-3-5）。

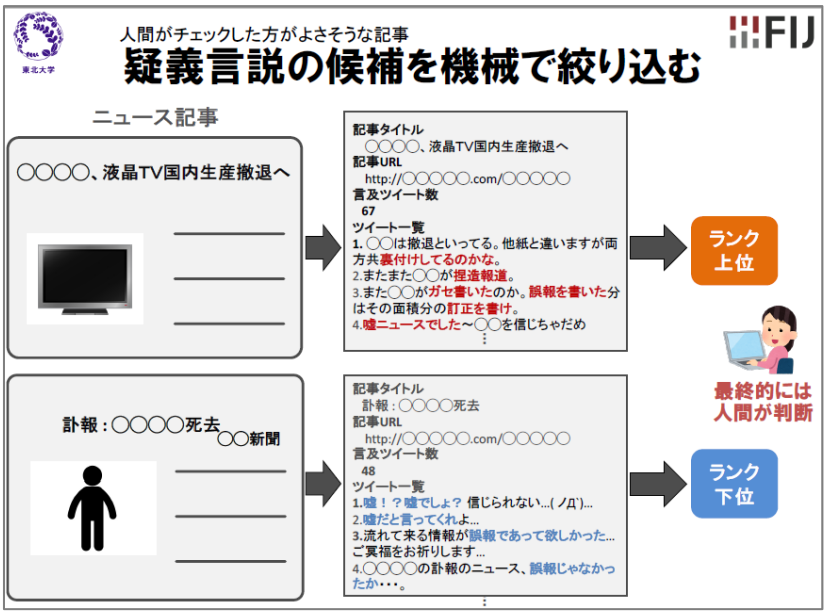


図 2-3-5　疑義言説の候補を機械で絞り込む

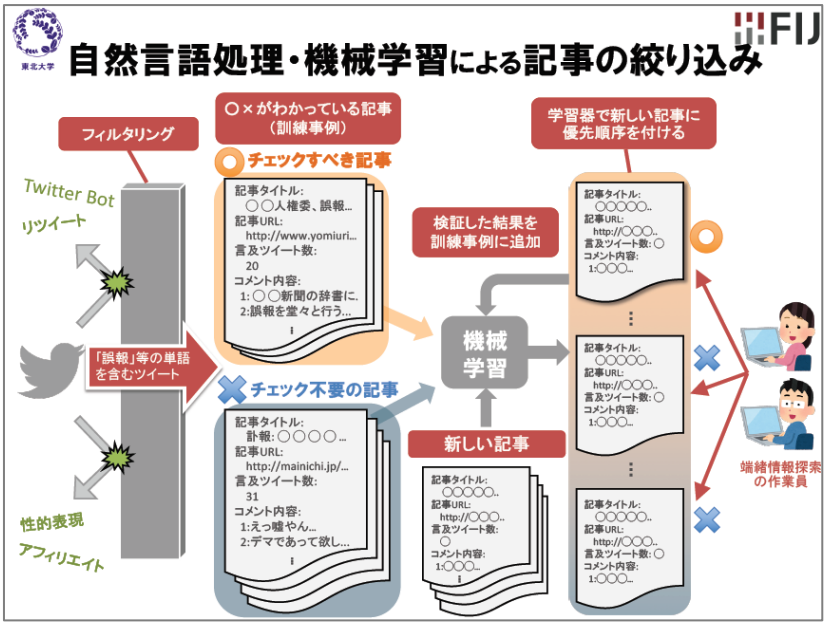


図 2-3-6　自然言語処理・機械学習による記事の絞り込み

技術的には簡単なことである。人間がファクトチェックすべき記事か、チェック不要の記事かの○×が分かっている記事を訓練データとして、そこにあるさまざまなコメントを使って機械学習する。機械学習でそれなりにいい手がかりを見つけて、よりチェックすべき可能性の高い記事がうまく取れるようになると、新しい記事を入れたときにいいランキングができる（図 2-3-6）。人間は、記事に対するツイートを見て、○（疑いあり）か、×（問題なし）を判断する。

実際にチェックした方がいいという記事は、別のシステム上に乗せられる。ファクトチェックにもさまざまな段階がある。メディアの方々に提供すると、今度はその人たちが本当に自分たちでファクトチェックをしたいものを選んで、取材や事実調査の上、ファクトチェック記事にする。このようなプロセスを回している（図 2-3-7）。

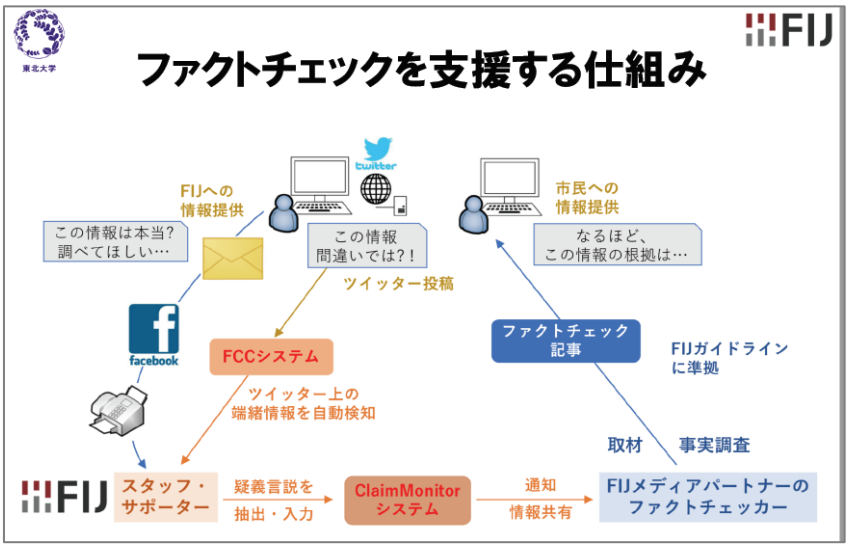


図 2-3-7　ファクトチェックを支援する仕組み

去年の沖縄県知事選でこのシステムを実際に使ってみた。選挙ということで、7 団体に参加いただき、FIJ の人たちと一緒に組んでファクトチェックのプロジェクトを実施した。FIJ は、疑義がかかっている言説を集めて 7 団体に提供する。7 団体は、疑義言説を特定し、実際に取材調査・検証した上で、記事化する。さらにそのサマリー記事を FIJ が配信する。実際にこのシステムを使って、5286 件の疑義言説をチェックして、69 件の怪しそうな言説を集めた。そのほかに外部からの情報提供もあり、沖縄県知事選の期間中に約 100 件程度の怪しい言説を捕捉した（図 2-3-8）。

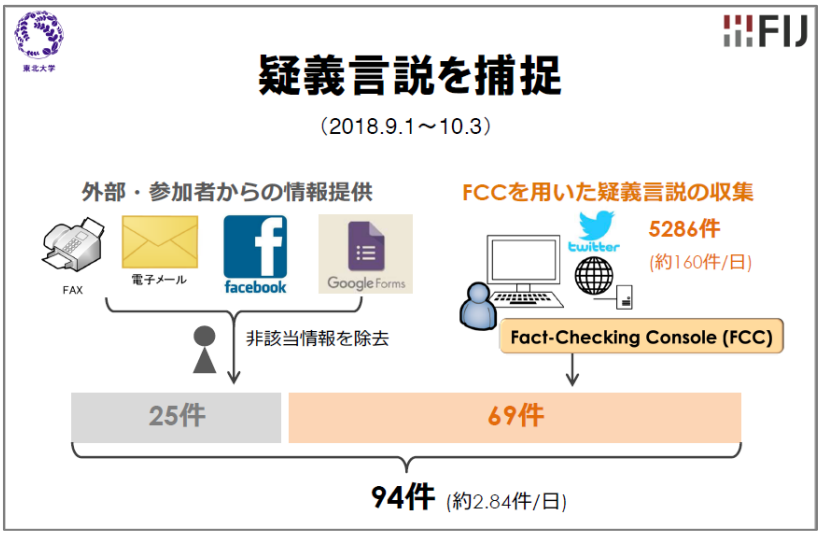


図 2-3-8　疑義言説を捕捉

図 2-3-9　ファクトチェックのプロセス

疑義言説は Facebook 上のグループを使って関係者に提供される。それぞれの団体はそこから選んで記事化していくわけだが、記事化に当たっては、FIJ のファクトチェックガイドラインに則った記事を FIJ からサマリー記事として配信するようにした（図 2-3-9）。

実際、さまざまな疑義言説が集まった。ツイートから怪しいと思われることが分かってファクトチェック記事が作られる。あるいは、逆側の候補の立場の方からも怪しいということで、ファクトチェック記事になっていく。SNS 上で流れる記事、ネットメディア、新聞の記事など、さまざまなものが捕捉できている。

重要なことは、怪しい疑義言説だということでファクトチェックしてみると、本当に誤りである可能性が非常に高いというのは、実際には半分ぐらいしかない（図 2-3-10）。残りの半分は大丈夫ではないかということである。人間が実際にファクトチェックをしないと分からないことがたくさんある。

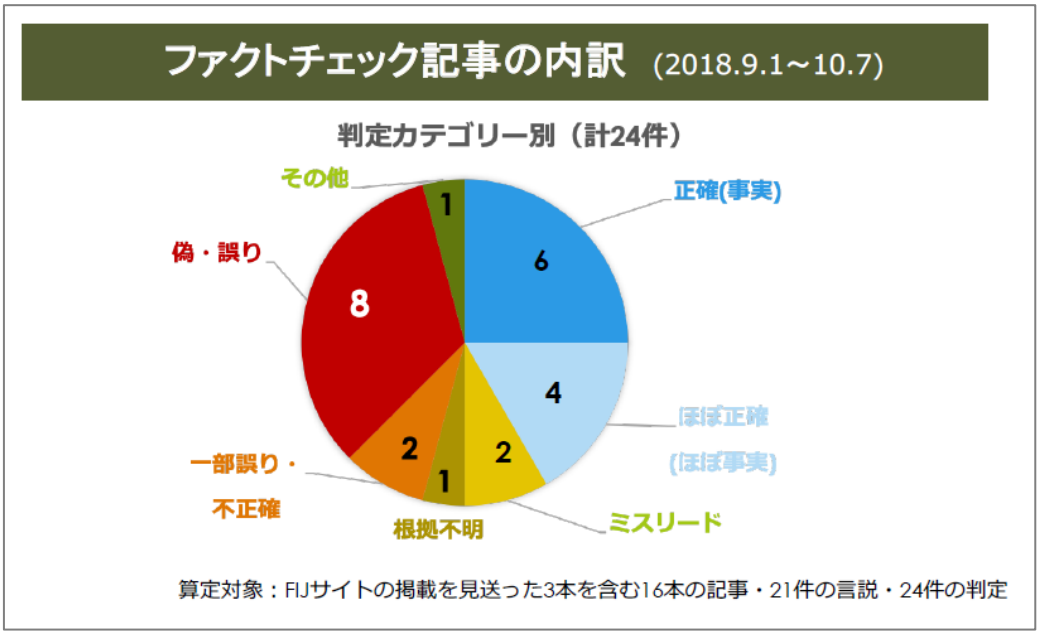


図 2-3-10　ファクトチェック記事の内訳（2018.9.1～10.7）

そういうことを実施するとデータがどんどん蓄積する。蓄積したデータをさらに機械学習にかけて、モデルを改良するということを既に半年ぐらい積み重ねてきた。モデルもよくなってきて、それまで 1 日 12 時間ぐらいかけて大量のツイートデータを見ていた人たちが今は 1 時間から 2 時間で、効率的に効果的に疑義言説を集めることができるようになった。先週の参院選（参院選 2019）についても、同様のことを実施した。統計はまだとれていないが、ファクトチェック記事がそこから多く生まれることが確認できた。

技術的には非常に素朴なことしかやっていないが、実際のファクトチェックの営みの中に技術を入れて回しているというのは、世界的に見てもかなり先進的な取り組みだと思っている。現在でも、ご紹介したシステムを毎日運用していて、1 日 100 件程度の疑義言説候補から 3～4 個程度の疑義言説をピックアップして、Facebook 上の FIJ の会員サロンで会員に共有するということを継続している。

疑義言説のスコアづけモデルは毎日、再学習している。モデル自身、システム自身も、FIJ のメディアパートナーというファクトチェックに関心を持っているさまざまなメディアの会員に無償提供している。

今は疑義を呈されている言説を集めてくるところしか支援できていないわけだが、将来は、ファクトチェックの本調査自体の支援もすべきと考えている。しかし、ファクトチェックの営みは、最終的には人間でないとできないと私自身は思っている。重要なことは、技術のための技術ではなくて、人間がやる営みをどう技術が支援できるか、こういうことを考えることである。現状は、東北大学と、それから、FIJ とスマートニュースのこのトライアングルでやっている。けれども、もっとたくさんの人に入ってきていただいて、こうした活動を広げていくということ、産学官、あるいは全然違う異業種との連携を実際に本当にやっていくことがフェイクニュース、あるいはファクトチェックにとって非常に重要なのではないかと考えている。

質問応答

Q： 内容にまで踏み込むと正誤の判定ができないものもあるのではないか？例えば、推定に基づく断定をしているのか？また、悪意を持った意図的なものは区別がつくのか？

A： 技術的には区別できない。技術でできることは本当に限られている。一方、人間は、できる場合もあるし難しい場合もある。ファクトチェックの対象は、意見や推測の真偽ではなくて、ファクトとして書かれている言説の真偽である。ファクトとして書かれている言説に対して、世の中のさまざまなソースに当たって本当に正しいかどうかということ、まさに現実に当たってグラウンディングしながら、その真偽を調べていく。それが分かる場合もあれば、真偽の証拠がまだ得られていないというところでとまる場合ももちろんある。しかし、それがファクトチェックという営みで、それを繰り返していくことが必要であると私は思っている。

Q： 最終的にはファクトチェックは人間にもできないのではないか？

A： 証拠をたくさん集めていく過程で、どちらか白黒がつくものも少なくないことがこれまでの営みで分かってきている。ただ、分からないこともたくさんある。分からないことは証拠不十分で、どちらともいえないというのが我々の態度である。

Q： ファクトチェックシステム自体の中立性、正当性はどのように担保するのか？

A： 最終的に判断するのは人間なので、システム自身はバイアスがなるべくないように作っているつもりである。ただ、機械学習に基づいている以上、もともとのデータにバイアスがあると、そのバイアスを引きずってしまう可能性があるので、設計あるいはデータの集め方自身に細心の工夫・注意が必要である。それはやり続けるしかない。

Q： いろいろな言説をいうときに、ファクトそのものは合っているが、使う場面や使い方、あるいはファクト自体が与える印象が大きいときに、印象操作ができてしまう場合がある。そういう場合にはファクトとしては正しいが、厄介な言説が出回ることになる。いろいろなパターンがありそうだが、何かできることはないか？

A： 自然言語処理の技術はまだたぶん、そこまではいけていないと思う。今言われたことは、ミスリーディング、つまり、ファクトとしては合っているが、それをこのコンテクストで使うと別の解釈をされる可能性が高いというような場合で、かなり高度な判断が必要となり今のところは人間がやるしかないと思う。

Q： 端緒情報がない言説は正しいという枠組みか？

A： 難しい。今のところ、人間がこれだけみんなネットで暮らしており、世界中の人たちがある種のセンサーだと、そのセンサーに引っかからないものについては、そこはまだ触れない領域、気づけない領域ではないか。一方で、先ほどの山岸先生のお話のように、フェイクのパターンみたいなものがある可能性はあるので、コンテンツそのもの、あるいは書き方そのものの特徴がうまく使える。相補的な営み、アプローチだと思う。

Q： 沖縄の選挙のときと今回の参議院選挙の類似性はあったか？

A： 参院選の解析は終わっていないのでよく分からないが、ネットでもリアルでも拡散するところは拡散しているし、一般の人たちからも政治家関係あるいは団体からも、それぞれ怪しい情報も出ていた。たぶん、今回も同じだったのではないかと思う。

Q： 対象の具体名を抽象名詞にすると、こういうパターンは危ないと分かるのではないか？それがオープンになると、一種の啓蒙活動に使えないか？

A： 非常におもしろい視点だと思う。

Q： このシステムのモデルが何を学習しているかに興味がある。今は URL を含むツイートが対象なので、学習しているのはその URL の前後にどういう単語が分布するかとか、そういうことか？

A： 基本的には、人がある記事に対して疑っていると表明している表現を集めている。それには色々な言い方があり、単語のいわゆる bag-of-words というか、単語だけを見ていてもよく分からない。皮肉的に言うこともあるし、かなりインダイレクトに言うようなこともあるので、学習が必要であると思っている。また、この URL から発信されたものは怪しいということも一部にはある。

Q： ファクトチェックをするときに、調査員の調査が正しいというのはどうやってダブルチェックするのか？知識が本当に正しいか、知識が足りているか、の検証はどうしているのか？

A： ファクトチェック記事も間違えるという前提で運営している。ファクトチェックは、特に欧米で先行してかなり広がってきているが日本ではまだまだ遅れている。ファクトチェックをするということ自身がファクトチェックの対象にも常になり得るということである。

Q： なるほど。白黒をつけるというよりは、信頼度を上げるという意味で行われているということか？

A： そこのレーティングもサマリー情報としては重要だが、より重要なのは、どういうエビデンスに基づいてそのレーティングに至っているかである。エビデンス自身がファクトチェックの記事ということになる。したがって、記事というのは、こういうエビデンスがあると、これらに従って考えると、これはどうも灰色なり、黒に近いと判断できる、というようなロジックである。それ自身をさらにメタに確認するというようなことは当然あり得る。

Q： 調査は実際にはどうしているのか？ネットで調査するのか、それとも、物理的にどこかに行って資料を生で見て確認するのか？

A： 私自身はそのプロセスには全く関わっていないので、本当には分からない。ただ、例えば、政治家の言説に対してミスリーディングであるという判断をするには、最終的には記事を書いた人に、その意図で書いたのかというようなことを確認するというところまで実際にやるということがガイドラインの中には入っている。最終的には誰も本当には分からないのだから、プロセスの中でなるべく悪いことが起こらないように、みんなで留意しつつ実施していると思う。

2.4 大規模意見集約：集合知・エージェント技術によるアプローチ ～自動ファシリテーションエージェントを用いた大規模社会実験～

伊藤孝行（名古屋工業大学）

「大規模意見集約：集合知・エージェント技術によるアプローチ」ということで今回のテーマに迫りたい。Collective Intelligence（集合知性）は、アリとかハチとかバッタなど、個体はそれほど賢くないが、全体としての集合的な行動に知性があるように見えるものである。これらの種が生き残ってきたのは、集団としての行動が進化的に優位であったということから、集団として動くということに関して何らかの新しいことを考えていかなければいけないと考えている。

特に人間の Collective Intelligence を考えたとき、50 年前、100 年前と違い、現在はスマホとかインターネットなどにより、コミュニケーションは完全に変わっているため、我々が、群れと

して、組織として動くときに、何らか違うものがあり得るだろう。MIT の Thomas W.Malone 教授は、よりスマートなスーパーマインズと言っているが、私もそれが大事だと思って研究している。

今回、システムを開発し事業化した。まずは、その PR ビデオをご覧いただきたい。
（D-Agree の紹介ビデオ上映　https://youtu.be/iZ5hyX7XCnI）

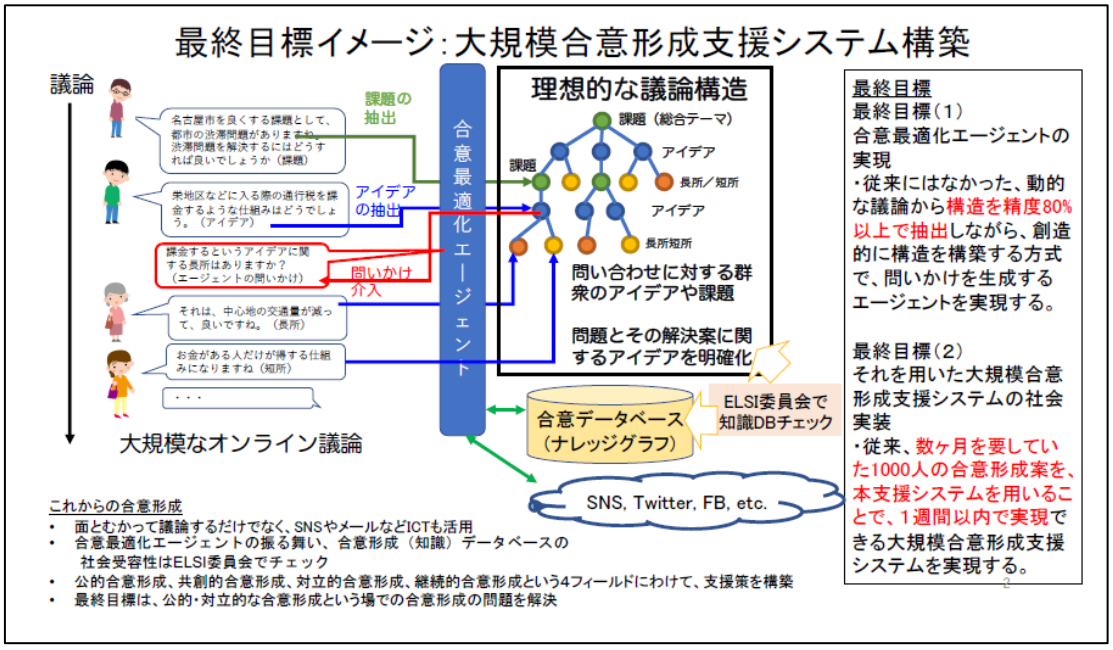


図 2-4-1　最終目標イメージ：大規模合意形成支援システム構築

具体的に何をやっているかを説明する（図 2-4-1）。このシステムには掲示板がある。掲示板の参加者が発言したものを、エージェントと呼ぶプログラムが読んで、議論の構造を見るために構造化する。そのあと、エージェントは、例えばこの辺が足りないなとか、この辺はもっと深掘りするべきだというのがあれば、参加者に質問する。

この例では、最初の人は、「名古屋市をよくする課題として都市の渋滞問題がありますが、どうすればよいでしょうか」と言っている。プログラムは、これが課題だと分かれば、「課題」としてデータベースに入れる。次の人が、「例えば栄地区などに入るときに、通行税を課すようなアイデアはどうでしょうか」と言ったら、プログラムは、これはアイデアであり、それもこの課題に対するアイデアの可能性が高いというのが分かれば、構造としてこの上の人の課題のアイデアだというふうに積み重ねていく。

このように構造を作った上で、例えば、まだアイデアの長所、短所が述べられていないので、プログラムは、「ほかに長所はありますか」とか、「アイデアはありますか」とか、問いかけることで議論を深めていく。議論の構造化の方法として IBIS（Issue-Based Information System）を採用している。これはアメリカ等ではよく使われるファシリテーションの議論を活性化させるための仕組みで、課題（イシュー）をベースに観点をたくさん出すことで課題に対する見方を深めていくという方法である。

エージェント・プログラムは、議論が発散したり、違った方向にいってしまう場合に、なるべく理想的な議論にしていくことが一つの目標になっている。IBIS の詳細は割愛するが、図 2-4-2 のような仕組みになっている。エージェントは、議論のテキストに深層学習を使って抽出し、データベースを使って問いかける。なお、暴力的な発言などにはフィルタリングをかけている。

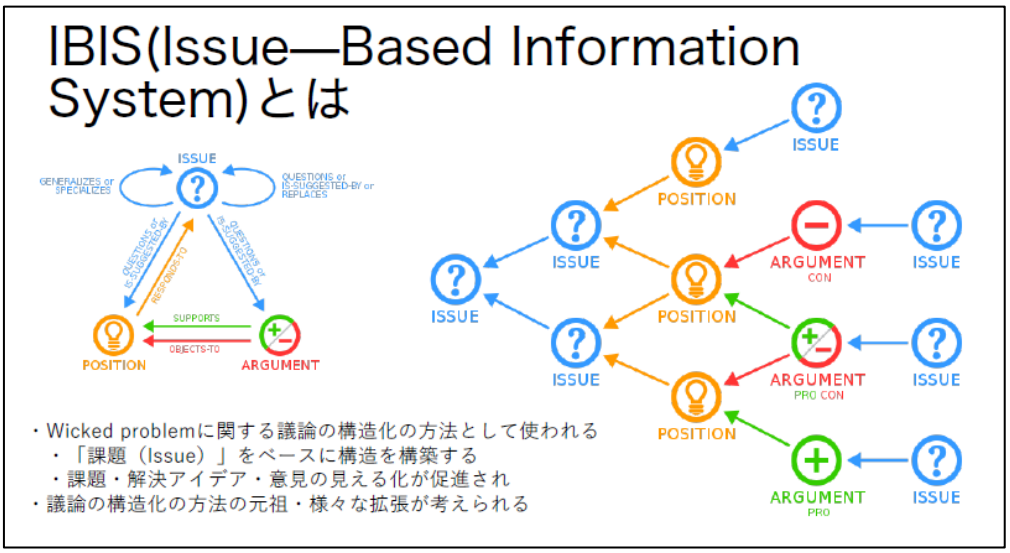


図 2-4-2　IBIS（Issue-Based Information System）とは

構造のノードとのリンクには、アーギュメンテーションマイニングの技術的なアイデアも使って抽出する。モデルを訓練するために、こういう形に沿った議論のデータというのを大量に積み重ねている。

議論の構造が取り出せたら、ある部分に対して、どういう発言を、ファシリテーターがするか、プログラムがするかというのは、テンプレートを用意している（図 2-4-3）。ジェネレーティブに文を生成することも考えているが、現状は、テンプレートで、ルールベースでやっている。

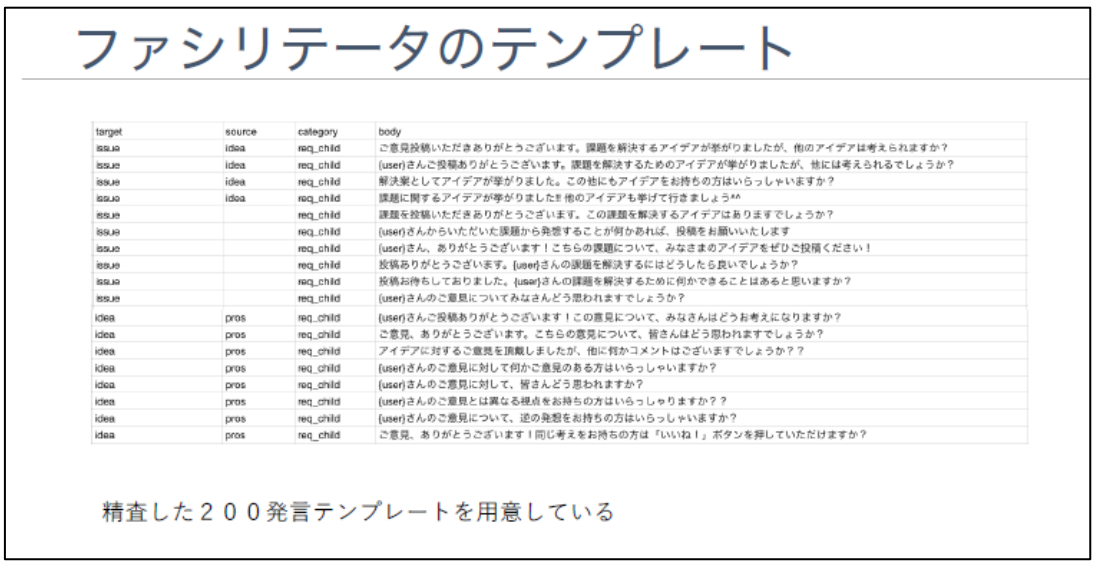


図 2-4-3　ファシリテーターのテンプレート

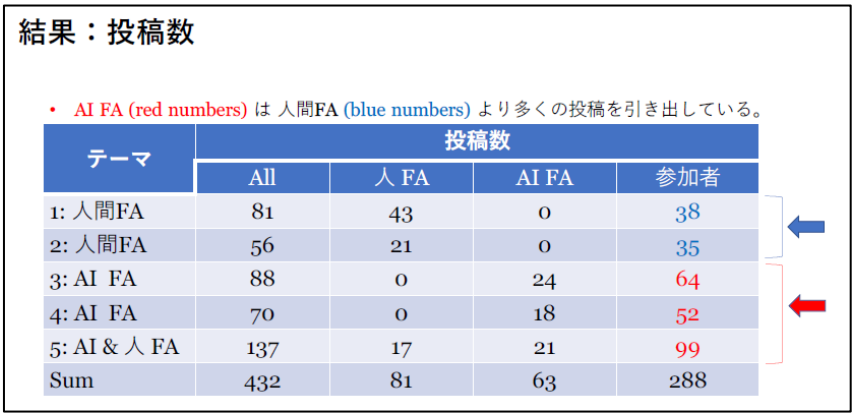
システムアーキテクチャー（実装）は、大学の中にあると社会実験がやりにくいので、AWS 上に構築した。深層学習は Python、ユーザーインターフェースは Java Script をベースに開発している。その他 AWS の監視の仕組みとかを使っている。

名古屋市と共催で、名古屋市次期総合計画の中間案に対する市民の意見集約を実施した（図 2-4-4）。名古屋市は、市民から意見を集約するために、12 区を一つ一つ回ってタウンミーティングを実施している。実際に、お金をかけて、場所もとって、人もかけているが、その中の一つをオンラインでもやりましょうということで我々のシステムを採用してもらった。

2018 年 11 月から約 1 カ月実施した。中間案について設けた 5 つテーマ中、2 つのテーマは人間が、別の 2 つは我々のプログラムが、最後の 1 つは我々のプログラムと人間が共同で、それぞれファシリテートした。

図 2-4-4　社会実験概要

いくつかの観点で結果を分析した。まず、投稿数から見た結果を図 2-4-5 に示す。人間がファシリテートした場合に比べて、AI を使った場合、投稿数を増やすことができたという結果が得られている。また、アンケートで、議論に対する満足度を聞くと、大体、どのテーマに関しても、AI であっても人間であってもそれほど変わらなかったという結果が得られた。



テーマ	投稿数			
	All	人 FA	AI FA	参加者
1: 人間FA	81	43	0	38
2: 人間FA	56	21	0	35
3: AI FA	88	0	24	64
4: AI FA	70	0	18	52
5: AI & 人 FA	137	17	21	99
Sum	432	81	63	288

図 2-4-5　社会実験結果：投稿数

成功事例がいくつかあるので一つ紹介する（図 2-4-6）。「名古屋の知名度をアップするには？」という課題に対して、「この課題を解決するためには何が必要ですか」と聞くと、参加者から「名古屋市はテレビで広告を出したらいいんじゃないか」という発言があり、我々のプログラムがこれをアイデアと認識して、「このアイデアの実現には何が必要ですか、深掘りしてください」と問いかけ、それに対して「名古屋市とテレビ局と協力を深めたらいいんじゃないですか」というような意見のやりとりが実際にできている。こういう成功例は幾つも出ており、それなりにファシリテートできることが分かってきた。

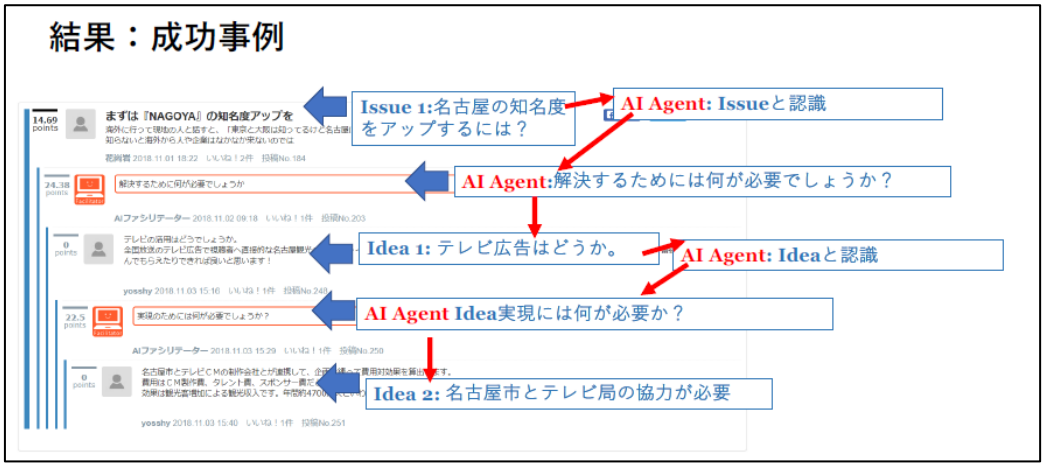


図 2-4-6　社会実験結果：成功事例

我々のプログラムはプログラムなので、参加者が投稿したらすぐ、それに対して反応ができる。そのため発言の頻度は非常に高い。0 時から 24 時までいつでも、参加者が投稿すると AI のファシリテーターも投稿できているということが結果に示されている。

別の事例では、ある方の発言でどきっとしたものがあった（図 2-4-7）。名古屋市に民主主義を取り戻したいという長めのご意見をいただいた。このときに人間のファシリテーターだと、どう答えたらいいかということで考えてしまうと思うが、我々のプログラムは、これはプログラムなので、これをちゃんと課題として処理する。課題を言っているんだが、ただ長いので、もう少し

まとめてください、現状は何が本当に課題なんですかと聞くと、発言者もきちんと短くまとめて回答する。その後、この課題に対して何かアイデアはありますかというようなファシリテートが段階的にできていた。

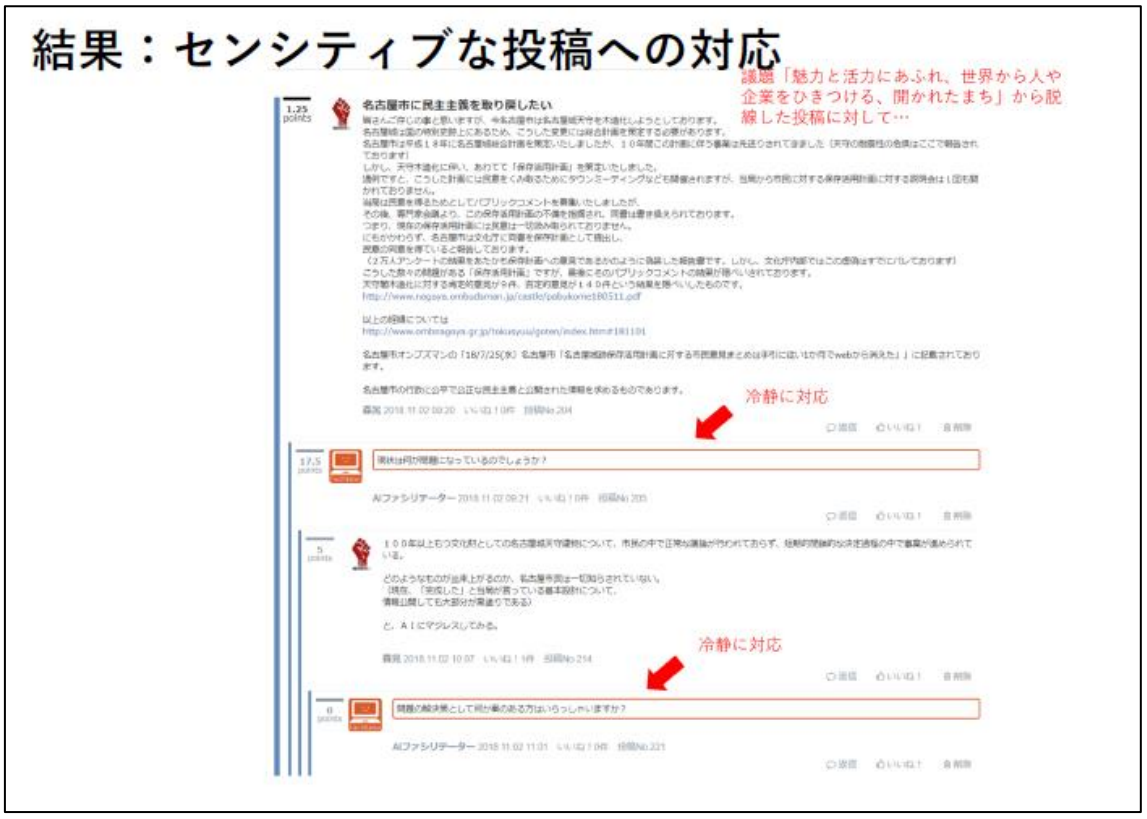


図 2-4-7　社会実験結果：センシティブな投稿への対応

構造化した結果は構造のビューアーとしてユーザーに示す。課題、課題を深掘りした課題、アイデア、アイデアに対する良いところと悪いところを表示する。間違うこともあるが、議論の構造がどういうふうになっているか見えるようになる。

暴力的な発言とか性的な発言をなるべく未然に防ぐためにフィルタリング機能をつけた。5 ちゃんねるとか、過去の我々の実験データの上でとってきて、その中でポジティブな発言とネガティブな発言､バイオレンスな発言というのを分けて学習させている｡Doc2Vec を使っただけだが、それなりの精度が出ている（図 2-4-8)。

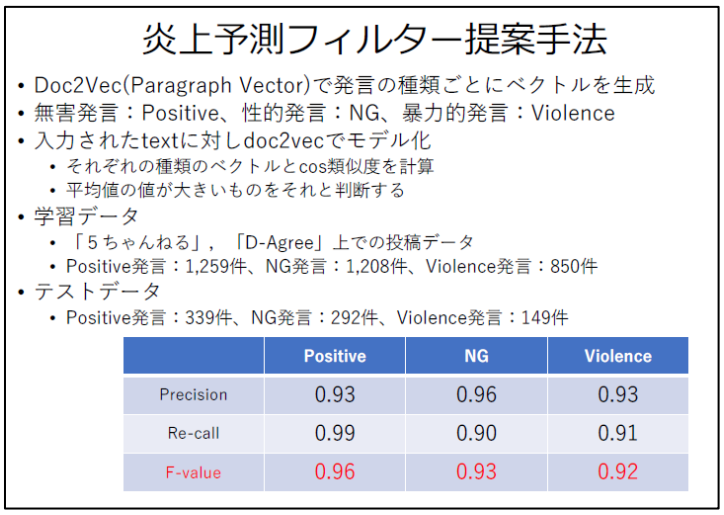


図 2-4-8　炎上予測フィルター提案手法

実際のシステムでは、このような発言があったときに、市民の皆様からいただいた発言をいきなり消してしまうのはよくないので、システムとしてはいったん非表示にして管理者が見る。管理者が見て、これは表示すべきだと思ったら、表示にするということをしている。今回の名古屋市との実験に関しては、そういうものがいくつかあって、いったんは非表示になっても、表示するというような手続をとっている。

まとめると、大規模に意見を集約するというプラットフォーム D-Agree を開発した。ファシリテーションを自動的に行うエージェントを作ってみた。議論のモデルとしては IBIS を使っている。これは一つの理想的なモデルであると考えている。名古屋市の次期総合計画に関する社会実験の結果を、部分的ではあるが紹介した。

今後は、ビデオにもあったように、発展途上国などに注力したいと思っている。インドネシアや、アフガニスタンのカブール市での展開は、国際化の事業として実施している。ほかにも、例えば医学会総会で実験をしたので、結果についてはまた違う機会にお話しできたらと思う。

今、研究の中で、私が一番注力しているのは、よいディスカッション、よい合意というのはどういうものなのかというような評価軸、評価手法の組み立てである。これについても次にもし機会があったら、これはまた、結果が出たところでお話ししたい。集合知の問題解決能力の測定法というのを、実は Thomas W.Malone 先生が 5 年前ぐらいにサイエンスに発表している。その方法に近しいやり方があるのではと考えている。

質問応答

Q： D-Agree を使ってみたい、どうすればいいか？

A： d-agree.com には商用バージョンがある。研究用バージョンは私にお問い合わせください。

Q： 課題から出発するが、議論の途中でそれは課題じゃなくて本当の課題はこっちじゃないのというようなことが出たら、どういうふうなカテゴライズ、枝分かれするのか？

A： 今の段階では個々の発言が課題かどうかを特定する。それに対して、上位のメタ課題があるのではという場合には、テンプレートでファシリテーターが処理するようにできる。

Q： ノード一つ一つのサイズに制限がないと議論が発散してしまわないか？

A： ノードのサイズの制限はない。
Q： たくさんのノードができたときにどうするか？
A： 現状の理想像として、IBIS は、課題があったときに、その見え方、観点をたくさん出してほしい、さらに各観点に対して、どういういいところ、悪いところがあるのかというのを全部出してほしいというのが基本的なアイデアなので、ノードが多ければ多いほどいいと考えている。
Q： 人間のファシリテーター、AI のファシリテーター、AI＋人間のファシリテーター、それぞれの特徴と短所は？
A： 人間のファシリテーションを見ていると、心を通わせるというか、ファシリテーションの言葉がすごく長い。参加者の発言の気持ちを汲み取った上で、こういうことかと言える。それに対して AI は、すぱっと言ってくれる。それはそれでいいときもあるし、それが誤解を生むときもあるかもしれない。
Q： ある意見を支持する人が多い、少ないは反映されるのか？
A： 現状は「いいねボタン」というのがあり、賛成が多いというのを取得することはできる。
Q： 議論が深まる、あるいは最終的に決定するということを AI はどう判断するのか。例えば発散していく場面と収束している場面のファシリテートはどうするのか？
A： 現状は、ある課題に対して広くカバーするという議論をしたいので、発散はするが、単に広がるのではなくて、IBIS の形に従って深めていく。直感的なまとめ、要約というところまではできていないが、IBIS の形に従って議論を全てカバーするように導くことはできている。
Q： 処理するのに得意なテーマ、苦手なテーマはあるか？AI のファシリテーターの話。
A： まさに、どういう議論だとうまくいくか、どういう議論だとうまくいかないかというのを今いろいろ試している。意見がすごく分かれてしまっている場合は難しい。この春に静岡で浜岡原発の立地に関する住民投票についてどう思うかという、そういうミニパブリックス型の社会実験というのをやった。そのときもエージェント自体はそれなりに動いていた。
Q： 合意しようと思っていない方々に対して、これを持ち込んでもうまくいかないのでは？
A： 難しいと思う、今のところは。
Q： ファシリテーターが AI と人間で、ユーザー側の対応に違いはあるのか？
A： 議論に対する満足度を調べた限りでは、そんなに変わらないという結果を得ている。
Q： AI が心理学をマスターすると、人間の意思決定や結論を誘導することも可能になるか？
A： 今でも、IBIS の理想的な議論の形で議論しなさいと、ある種誘導している。
Q： それが逆に言えば危険性もあるかもしれない。
A： モデルの使い方にある。
Q： 人間よりは中立なんじゃないかなと思ったがどうか？
A： 人間のファシリテーターにもストラテジックにやるような人がいる。例えば、ある会社で新規事業を考えるときのワークショップなどでは、その会社のための筋道を立ててファシリテートするというのがあるので、それよりは公平である。
Q：人間だと相手のキャラを読むことがあるが、AI だと中立的なので、むしろ、発言する人の本音が出るということはないのか？
A： それはそう思う。AI だとそういうことはできないので公平である。

2.5　自動意思決定・自動交渉：機械学習・最適化技術によるアプローチ

森永聡（ＮＥＣ）

この発表では、前半で自動意思決定システム、後半で自動意思決定システム間の挙動調整・自動交渉の話をする。既にいろいろな分野で商用化している自動意思決定システムの話は、いわば前振りで、その限界を突破するために取り組んでいる新しいテーマが、後半で紹介する自動意思決定システム間の挙動調整・自動交渉である。

2.5.1　自動意思決定システム

ここでいう自動意思決定システムは、伊藤先生の発表にあったような集団・コミュニティにおける共通認識を作るという話とは異なり、システムの利用者（我々から見たら顧客）に対して、この後どうするとよいのかを、システムが自動的に計算して教えてあげるというもの。そのような自動意思決定システムの基本構造は、およそ図 2-5-1 のようなものになる。

図 2-5-1　自動意思決定システムの基本構造

ここでは AI が手伝って 3 つのステップが実行される。まず「見える化」のステップ、情報を集めて、外界認識・情報構造化を行う。次に「予測」のステップ、構造化された情報をもとに将来予測や不明推測、この後に何が起きるのか、この後に何をするとどうなるか、といった予測を行う。最後に「最適化」のステップ、そのように少し先に何をしたらどうなるかが分かるのならば、利用者はどういう動作をするとよさそうか、AI が選んでくれて、実際にそれを実行する。

例えば電力調達運用の意思決定システム。これは実際に稼働している。そこでは、最初の「見える化」ステップでさまざまなセンサーやネット情報を集めて、それをもとに「予測」ステップでは、1 時間後にどの地区でどれぐらい電気が使われるか、2 時間後にはどの地区でどれぐらい使われるか、自分がどの発電機を回すと、それがどれぐらい賄えるかとかいった予測を機械学習を用いて作る。その結果、「最適化」ステップで、この後どのように運転すると顧客が決めた KPI の値がどうなるか、さらに、特別の事情で避けなければならない動作といった制約も考慮して、

その中で KPI の値を最大にできるような計画を選ぶ。そのようなさまざまな制約や KPI を考慮して計画を作ることは人間にとってとても大変なことなので、AI が選んでくれて、電力調達ミックスの計画まで作って実行してくれる。

もう一つの例は自動運転車。ここではまず「見える化」ステップで、さまざまなセンサーから、どこに車線があって、標識に何と書かれていて、交通信号機がどういう状態で、ほかの車や歩行者はどの辺にいるかを認識して、それに基づいて「予測」ステップでは、この後、この車はどちらの方向に動きそうで、この歩行者はどちらからどちらに出ていきそうかを予測したり、自分の車はどのように運転するとどうなるかというのも予測したりする。それを踏まえて、「最適化」ステップでは、絶対確保しなくてはならない安全性がどれくらい以上で、かつ、燃費が少ないとか、到達時間が短いとかの条件をミックスして最適な計画を作る。つまり、自分が何秒後にアクセルを踏んで、その後、どういうスケジュールでハンドルを切って、といった運転の動作を選んで、実際に実行する。

製造工程管理の例も同様。いま世の中はどうなっているか、工場はどう動いているか、情報を構造化しておいて、この後の将来の需要の予測はどうなるかとか、必要なリソースとか、時間とか、要因はどうなるかを予測しておき、それに基づいて、顧客が決めた KPI を大きくする生産計画を立案して実行する。

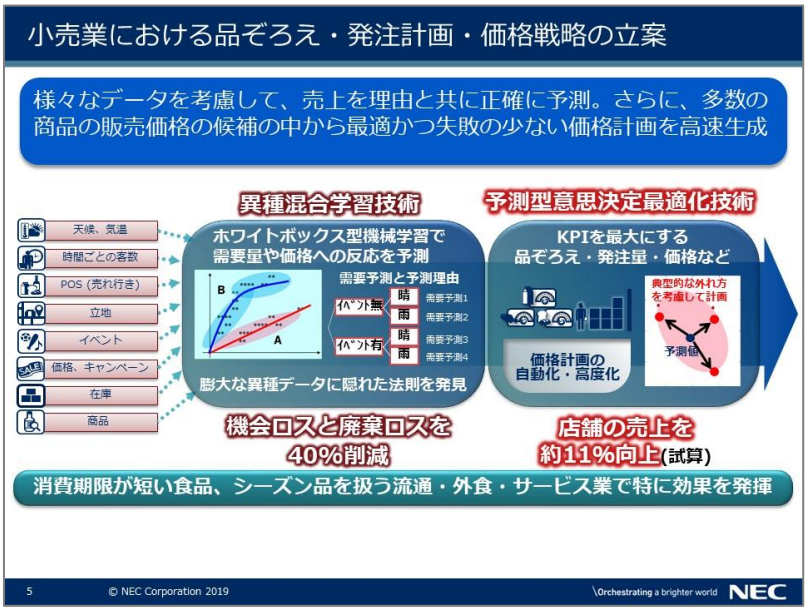


図 2-5-2　小売業における品ぞろえ・発注計画・価格戦略の立案

小売業に納めている自動意思決定システムでは、いろいろな情報を集めて、この後、何がどれぐらい売れそうかの需要予測をして、そのとき、どのような品ぞろえにするとどれぐらい売れるかとか、どのように各商品に値段をつけると、どれがどれぐらい売れるかとかまで予測する。そういうものを全部予測した上で、ならば小売業者にとっての KPI を最大にするには、いったいどういう品ぞろえにして、それぞれをどういう発注量にして、どういう価格づけをすればよいのか、という意思決定を全部 AI を使って最適化する（図 2-5-2）。

このとき、予測のところがブラックボックスだと、出てきた計画がどうしてそういう計画がよいのか、人間には全然分からない。この問題に対して、NEC には異種混合学習技術というホワイ

トボックス型の予測器があるので、これを使うと、どうしてこういう予測になるか、こういう値づけをすると、こういうふうな価格反応曲線があるからだと説明ができる。例えば、ある価格戦略でビールを何銘柄か売るとき、定価で全部売ると何曜日にどれが何個売れるか予測できて、その結果、この週の総売り上げが予測できる。そこでさらに、価格を変更すると、どういう共食いが起きて、どれがどれぐらい売れるか、どれがどれぐらい売れなくなるかも予測できる。このように、何曜日にどの銘柄にいくらの値段をつけると、その週の店の総売り上げがいくらになるかを、AI を用いて予測できる。これを一々、人間が調べて考えるのはとても無理で、AI ならばこの最適化問題を解くことができる。

さらに、ドライバー管理のシステム。シンガポールの公共交通などで、ドライバーの事故リスク予測やトレーニング計画を行っている。つまり、ドライバーのトレーニング計画を立てるために、どのドライバーがどういう事故を起こしそうかを予測し、どういう教育を授けると、その事故を起こす確率がどれくらい下がるかについても予測する。その上で、限られた教育予算とドライバーのタイムテーブルを踏まえて、誰にいつ、どういう教育をするとよいか、AI が最適化する。

もう一つ、会員制ビジネスをやっている顧客向けのソリューション。どの会員が今後どれぐらいお金を払ってくれそうかを予測し、かつ、どの会員がやめそうかも予測し、どの会員にどういうサービスやおまけをつけるとやめないでいてくれるかということまで予測する。限られた予算のもとで、逃げてしまう利益を最小化するためには、いったい誰にどういうサービスやおまけをつけてあげればよいのか、という意思決定問題を AI を用いて最適化する。

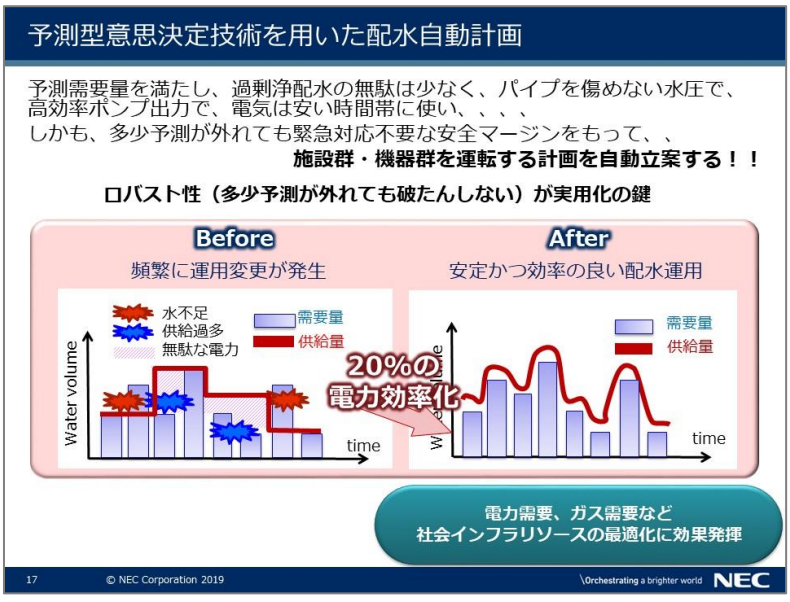


図 2-5-3　予測型意思決定最適化技術を用いた配水自動計画

水道施設の機器運転の計画自動立案もやっていて、水道事業者に使われ始めている（図 2-5-3）。今後どの地区でどれぐらいの水が使われそうかを予測し、かつ、水を浄水する施設をどういうふうに動かすとどれぐらいのコスト・時間・電気代がかかるか、各貯水池に水を送る際にはこれくらいの量だとどれくらいの電気代や時間がかかるかも予測する。さらに、パイプの劣化による水漏れは日本ではまだ少ないけれど海外ではしばしば起こることだが、どのパイプにどれぐらい圧をかけると、どれぐらい水が漏れるかとかいうことも予測する。そのようにして、予測された水

需要量を満たし、その一方で、水を作りすぎる無駄を少なくし、かつ、パイプを傷めない水圧で送るような運転計画を作る。その際、ポンプにかける電気代と、送れる水量の関係が実は線形ではないので、なるべく効率のよい条件でポンプを使った方がよいし、電気代は夜の方が安い。そういうさまざまなことを全部考えた上で運転計画を最適化する。

それから、このような自動意思決定システムにおいて、もう一つ重要なポイントとして、予測が的中する率は必ずしも 100%ではないので、予測が多少外れた場合でも緊急対応不要な安全マージンを持って、いろいろな施設群や機器群を運転する計画を自動立案するということがあげられる。ここまで述べてきたような基本構造に基づいて自動意思決定システムは動いている。必要な情報を集めて、この後どうなるか、何をどうしたらどうなるかを予測し、予測誤差も勘案して安全で有利な案を選択する。

2.5.2 自動意思決定システム間の挙動調整

自動意思決定システム自体は、ここまで述べてきたように、いろいろな応用に使われるようになってきたが、そういうものが普及した世の中を考えると、そういうもの同士が干渉し合うようになってしまう場面がきっと起きる。そこで、そのときに必要になる自動意思決定システム間の挙動調整という問題の研究開発に取り組んでいる（図 2-5-4）。

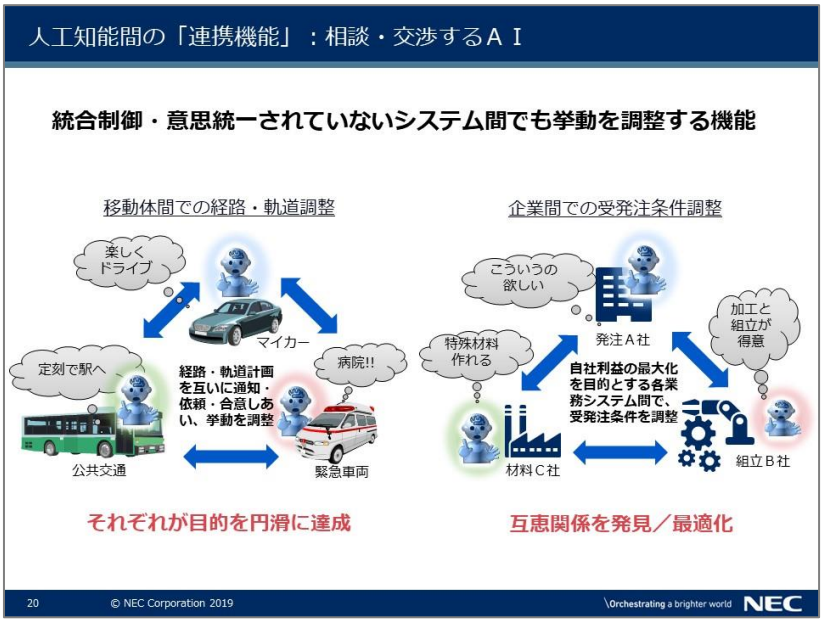


図 2-5-4 人工知能間の「連携機能」: 相談・交渉する AI

既に述べたように、自動運転車というのは自動意思決定システムの典型的なもので、周りの様子を見て、この後どうなるかという周りの変化を予測した上で、自分にとって有利なように次の動作を決める。しかし、こういうものが世の中に普及してくると、自動運転車同士が出合ってしまう場面は多々発生する。例えば、図 2-5-4 の左側の例のように、自動運転車 3 台が交差点に近づいてきたとき、自動運転車にとって、残りの 2 台がこの後どう動くかについて 100%の確信がない限り、少しスピードを落として様子を見るというのが正しい動作になる。しかし、それは他の車も同様で、相手がどう動くか分からないからスピードを落とし、この 3 台は近づけば近づく

ほど、ゆっくり様子見をし合って、お見合いすることになってしまう。これはとても無駄な状況で、より望ましい方法として考えられるのは、各自動運転車が自分はこの後どう動くか、どういう計画を持っているかを告白し合って、動作を決めることである。例えば、救急車からマイカーとバスに向かって「あなたの計画は困る、こちらは急病人が乗っているので、そこの交差点には3 秒間入らないでください」と計画の変更を依頼し、かつ、依頼された側は、それを受け入れられるのであれば「了解した、そうする」と約束してあげるというようなことである。このようなシステム間の動作調整ができると、互いに相手がどう動くか分からなくて立ちすくんでいるよりも、3 秒間待つだけでまずは救急車を行かせてしまい、その後、また 3 秒間、今度はマイカーが待ってバスを行かせてあげれば、一番遅くとも 6 秒したら行けることになり、それぞれの目的が円滑に達成できることになる。

もう一つ、図 2-5-4 の右側の例は企業のサプライチェーン。1 つの企業内のオペレーションは、先ほどの自動意思決定システムで最適化できる。しかし、企業と企業の間の商取引条件というものを自動的に交渉して調整できるようになると、互恵関係を発見して最適化するなど、1 企業内の最適化だけでは得られない、もっと高いレベルの KPI を各社が達成し得る。

先に述べた自動意思決定システムは、いわば、相談しない自動意思決定システムである。相談しない自動意思決定システムは、自分なりに必要な情報を集めて、周囲の動きを予測する。しかし、相手がどう出るかは想像に過ぎなくて、自分の予測に関する確信がないので疑心暗鬼になり、回避も過剰になりがちである。また、本当に KPI 上、得になるのであれば、相手がどう動くかは分からないけれども、やってしまえということにもなる。そのため、相談しない自動意思決定システムでは、もったいない状況や危ない状況が起きてしまう。これに対して、自動意思決定システム間、つまり、AI 間で挙動調整を行うようにすると、相手の挙動計画を知ることで、より精緻な予測分析ができたり、相手に挙動計画を変更してもらうことで、より安全な制御・誘導を可能にできたりする。

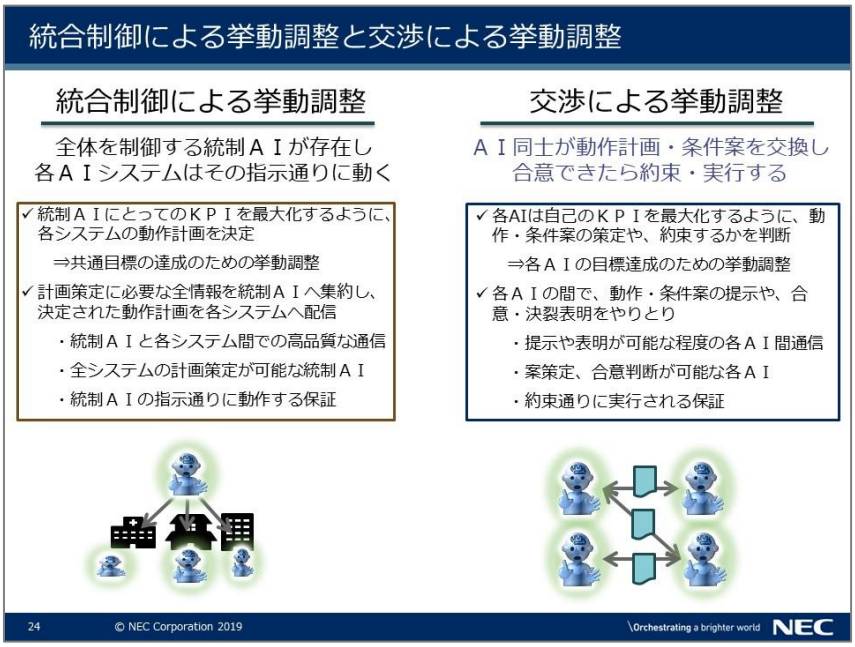


図 2-5-5　統合制御による挙動調整と交渉による挙動調整

複数システム間の挙動調整を実現するには、図 2-5-5 に示すように、大きく分けて 2 通りのアプローチがある。1 つめは、全体を制御する統制 AI が存在して、各 AI システムはその指示通りに動くというもので、もう 1 つは、分散した AI 同士が動作計画・条件案を交換して、合意できたら約束、実行するというもの。

前者の統合制御による挙動調整をもう少し詳しく説明すると、ある種マザーコンピューター的な統制 AI が存在し、各システムはその統制 AI に自分の持っている情報とか、やりたいことを上げて、統制 AI が全体計画を計算・決定し、各システムにそれぞれ何をどうすればいいかを命令するという構成になる。実は、これは普通の自動意思決定システムのアクチュエータが分散しているのとほぼ同じである。これはこれでよいのだが、この形が適さない場面がある。例えば、前に述べたような公道上での自動運転車が出合った場面では、このようなアプローチは難しそう。また、企業間での商取引条件の調整という場面でも、マザーコンピューターが「おまえの会社はあの会社に何をいくらでいつ売れ」のような命令をすることになるので、資本主義経済においては社会制度的に受け入れられない。

一方、後者の各システムが主体性を持って、自分と相手との間で、合意条件の調整を交渉によって行うという方法が実現できれば、公道上での自動運転者間の調整場面でも何とかなるし、資本主義経済でも会社間の商取引条件の調整ができるようになる。

このような 2 通りのアプローチはそれぞれ向いているユースケースが異なる。統合制御が向くユースケースは、社内オペレーションやグループ企業内オペレーションの最適化、リソース融通など、全てを 1 つの管理配下に置けるようなケース。社会主義・共産主義経済のもとでの会社間調整はこちらのパターンになり得る。一方、交渉向きのユースケースは、資本主義社会における会社間受注調整やリソース融通、あるいは、マルチエリア／モーダルでの交通・人流制御など。

私は現在、後者のアプローチ、つまり自動交渉 AI の研究開発に取り組んでいる。これは、双方が合意できる Win-Win 条件を相談・交渉する AI エージェントを開発するという取り組みである。相手側に提示すべき合意条件案を自動生成する AI、それから、提示された合意条件案に対して受け入れるか、拒否するかを自動判断する AI を作ることになる。この目標に関して、毎年、IJCAI という国際会議の中で、自動交渉 AI の交渉のたくみさを競う競技会 ANAC（Automated Negotiation Agent Competition）が開催されている。今年も 8 月半ばにマカオで IJCAI2019 があり、自動交渉 AI の競技会も開催される。

それから、Facebook もこのようなエージェント間の交渉によって合意点を見つける技術を開発している。自然言語による交渉ゲームで、Facebook が作ったディープラーニングを用いたシステムが人間に勝利し、話題になった。

このような自動交渉 AI の技術が実用化できたら、自動運転車間での挙動調整によって、合流やすれ違いはスムーズになり（図 2-5-6）、会社間の商取引の調整も、物を買いたい会社の交渉エージェントと、物を売りたい会社の交渉エージェントの間で、物の売買条件を自動的に調整してくれて（図 2-5-7）、その後、決まった売買条件に沿って、製造計画を最適化し、実行することも可能になる。また、いろいろな有識者にヒアリングしたら、日本では物の売買ではなく、まず物流手配にこの仕組みを適用して効率化するのがよいという意見が多く、物流手配にも取り組んでいる。その際、自動で決めてしまうモードだけでなく、最後の判断は人間がするというモードも持つようにしたい。そして、なぜこの相手とこの条件で合意したかを人間が知りたかったならば、

きちんとその理由を説明できるようなエージェントを作りたいと考えている。

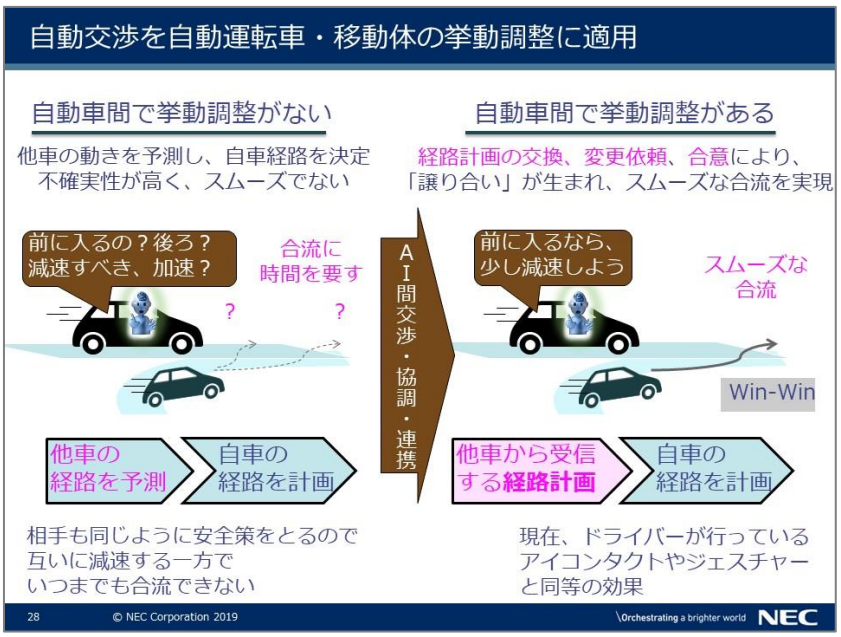


図 2-5-6　自動交渉を自動運転車・移動体の挙動調整に適用

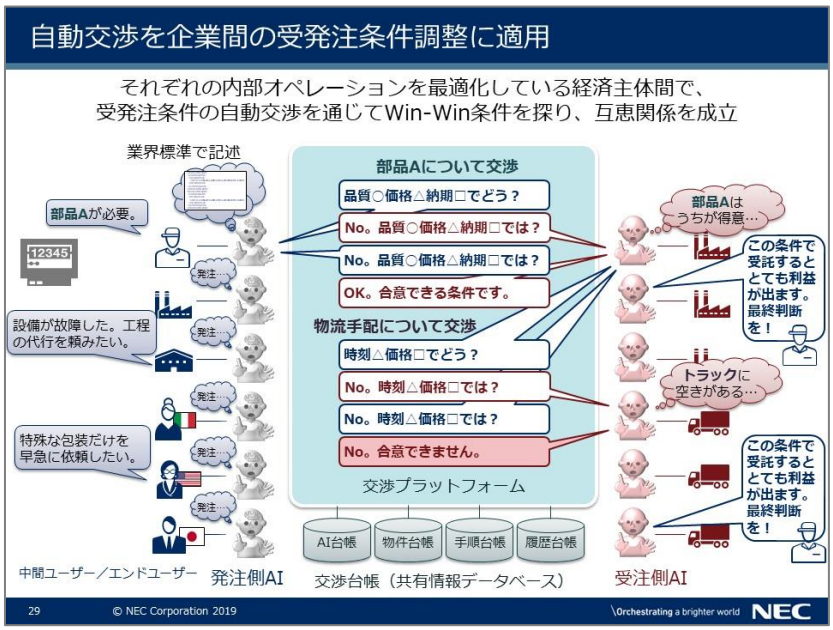


図 2-5-7　自動交渉を企業間の受発注条件調整に適用

企業間で Win-Win になる条件が見つかるので、1 企業内だけで最適化するのを超えた可能性が生まれる。1 企業内の最適化だと、余計な設備は稼働させないように運転計画を工夫して最小コストで生産実行という方針になるのが、企業間で自動交渉を行うようになると、自社の利益を最大化するように条件交渉をして、もっとたくさん設備を稼働させてビジネスを拡大するような意思決定が可能になる。それを広げていければ、従来人間による商談で構築されてきた企業間連携を超えた効率と幅で、もっと広いエコシステム圏を作れるはず。Society 5.0 でマスカスタマイゼ

ーションの話が出てくるが、そこでは、大量生産社会ではなく、さまざまな人たちがさまざまなものを欲しがっているとき、必要な量だけ必要な人に必要なときにすぐ届けるという経済システムが目指されている。それを実現しようと思ったら、大量生産社会に比べて、交渉・調整の頻度や回数が 1 千倍とか 1 万倍とかに増加するわけで、AI による自動交渉が不可欠になる。

図 2-5-8　自動交渉のユースケースと Society 5.0

政府が Society 5.0 として目指すべき将来社会像を示しているが、それは独立の意思決定システムがばらばらに最適化していては実現できないようなものが多い（図 2-5-8）。Society 5.0 は AI に限らず、目指すべき社会像が示されたものだが、人工知能技術戦略会議がまとめた「人工知能の研究開発目標と産業化のロードマップ」に示された将来像についても、個別のシステムにおける意思決定だけでは実現できないようなサービスが多数あげられている。つまり、このような将来社会像を実現するためには自動交渉 AI 技術が不可欠であり、この研究開発を強力に推進していきたいと考えている。

質問応答

Q： 挙動調整に関して未解決の課題は何か？

A： 多数ある。交渉によって定めなくてはいけない論点が 1 個だけ、例えば価格だけというケースについては、セカンドプライスオークションなど、非常に効率的な方法が知られているが、論点が複数になり、合意できる多次元空間内の 1 点を見つけないといけないケースになると、自分の全情報をさらけ出さずに、互いに合意できる点を見つける効率よいアルゴリズムは知られていないため大変。

Q： 局所最適化になり、必ずしも全体最適化にならないのではないか？

A： そこが結構問題で、全体最適化にはならないことが分かっている。また、合意点が必ずしもパレート最適にいかない。全体最適は、個々のシステムの KPI を犠牲にしてでも、何かの値を大きくするところで合意することなので、何らかの強制力がないとほぼ原理的にできない。なので、さらにそれを保証するようなもう一個の自由度があって、胴元がいれば何とかなるかもしれないという研究もされている。

Q： システム間の挙動調整の際に、全体としては発散し、想定外のことが起きたりしないか？

A： AI によって売買するアルゴリズミックトレードで、金融マーケットが不安定になってしまったという話がよく知られている。似たようなことは起こりえるかもしれないので、AI と金融マーケット市場に詳しい東大の和泉先生と一緒に、こういうものが導入されたときのマーケットの安定性に関する解析や悪意の及ぼす影響などの研究も進めている。

Q： 例えば、勝手に AI にやらせておいて、独禁法に違反するようなことが起きないか？また、Win-Win といっても利益の配分だけ考えておけばよいのか？もう少し外側の枠組みもいろいろ考える必要があるのではないか？

A： ご指摘の通りである。実際、セルフプレーで交渉アルゴリズムを学ばせると、別に人間が教えたわけではないのに、談合とかを学んでしまう。この問題に対してどうすればよいか、実は法学部の先生にも一緒に入っていただいて研究している。また、独禁法や談合に加えて、下請法とかも問題になってくる。くるくると条件を変えるというのは優越的立場の濫用で、これを AI が学ぶと、めちゃくちゃやり始めるかもしれない。

Q： 救急車が来たら一般車が待つという例があったが、それは何かルールを書くことになるか？

A： ルールを書いて全部できればよいが、相手が自分と同じルールを持っていることの保証や、このルールをいま適用してよいという条件を相手も同意しているかの確認が必要になる。それなら、むしろ自分はどうするかを伝えてしまった方が話が早いかもしれない。もちろん、ルールベースで、かつ、お互いの認識が一致していることが確認できるならば、ルールベースでできる範囲はたくさんあると思う。

3．総合討論

3.1　意思決定のための情報科学は必要か

小林正啓（花水木法律事務所）

法律家という立場から、フェイクニュースや政治的な意思決定に関する議論の背後にある「思想の自由市場」という法律的な概念について説明する。

まず、わが国でフェイクニュースは違法なのだろうか？例えば、虚偽の風説を流布して、他人の信用を毀損した場合には、信用毀損罪が適用されうる。偽計業務妨害罪、名誉毀損罪、外患誘致罪や、内乱罪の摘要も考えられる。公職選挙法においては、候補者の経歴等に関して虚偽の事実を述べると虚偽事実の公表罪になる。民事訴訟法上はフェイクニュースによって名誉を毀損されたとか、信用を害された場合には損害賠償請求ができる（図 3-1-1）。

ただし、フェイクニュースそのものを処罰する法律はない。

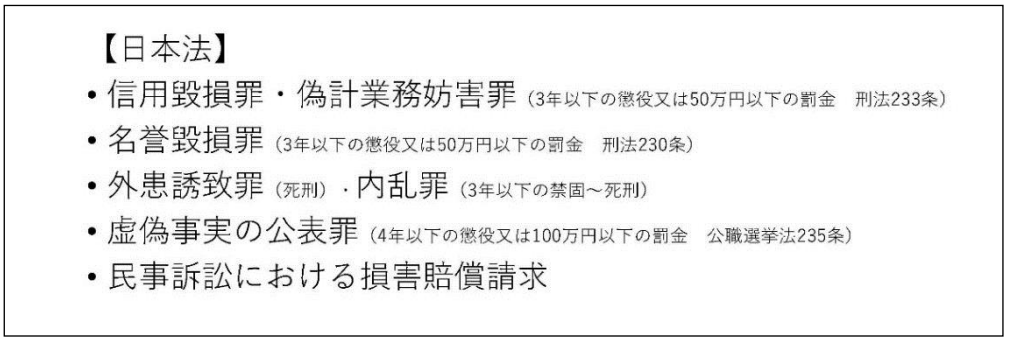


図 3-1-1　フェイクニュースは違法か？（日本）

ドイツでは、2017 年にネットワーク執行法が制定された（図 3-1-2）。200 万人以上の利用者を要する SNS 運営事業者に対して、特定の犯罪を含むコンテンツに関する苦情処理などの義務を負わせるものである。ただしこれも、苦情処理の義務を怠ったとか、苦情処理をするための制度を作らなかったとか、うまくそれを運営しなかったという場合に、裁判所の判断を経て罰金を課すものであり、フェイクニュースそのものを処罰するものではない。ヨーロッパでもドイツだけがこのような規定を設けており、他国にはない。

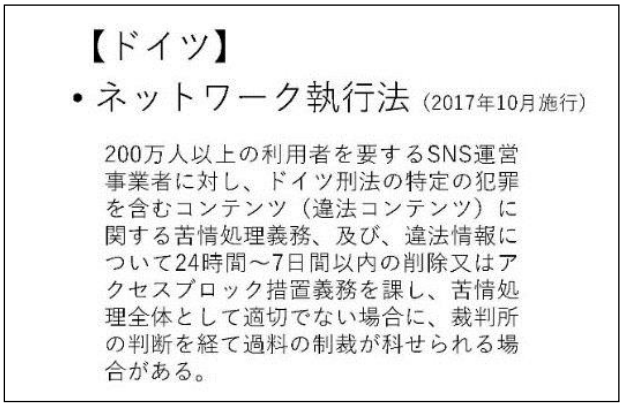


図 3-1-2　フェイクニュースは違法か？（ドイツ）

わが国にはフェイクニュースそのものを処罰したり、それを違法としたりする法律は存在しない。虚偽のニュースが何らかの犯罪に当たるという場合がもちろんあるから、そういう場合には罰するけれども、フェイクニュースを使って、あるいはそれで世の中を混乱させたというだけで人を処罰したり、損害賠償を払わせたりするという法律は存在しない。ドイツを含めた世界的な状況としても、そのような法律は存在しない。いわゆる自由主義国家には、そのような法律は存在しない（図 3-1-3）。

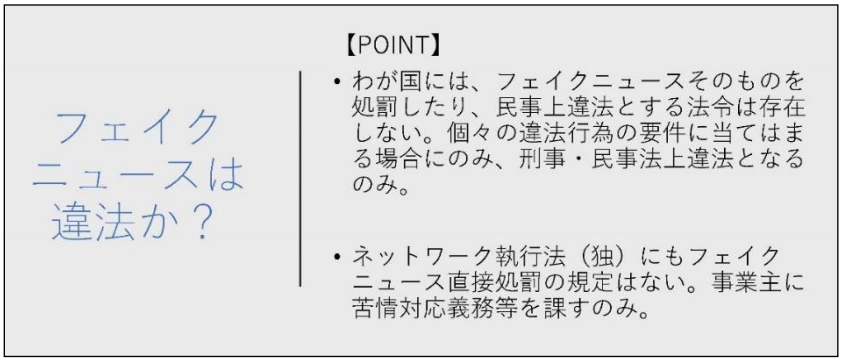


図 3-1-3　フェイクニュースは違法か？

昔はフェイクニュースというのはなかったのかというと、そんなことはない（図 3-1-4）。関東大震災における流言飛語や、太平洋戦争末期の大本営発表、ラジオドラマのせいで火星人が襲来したと思い込んだ人たちがパニックを起こしたという事件などがある。例を挙げれば切りがないほど、世の中にはフェイクニュースが存在した。なぜこれを処罰する法律がないのか？

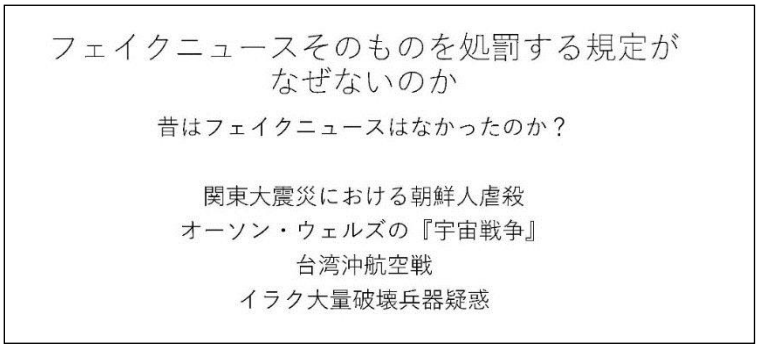


図 3-1-4　昔はフェイクニュースはなかったのか？

これはなぜかというと、運用が大変難しいからである（図 3-1-5）。フェイクニュースで処罰するといったところで、まず、その事実と評価をどう分類するかという問題があり、虚偽事実であるということの証明の困難さ、虚偽事実であるという証明が正統性をもつことの難しさもある。さらに公平な運用ができるのかという問題があり、悪用の危険性もある。フェイクニュースそのものを法律で規制するというのは極めて危険であるということである。

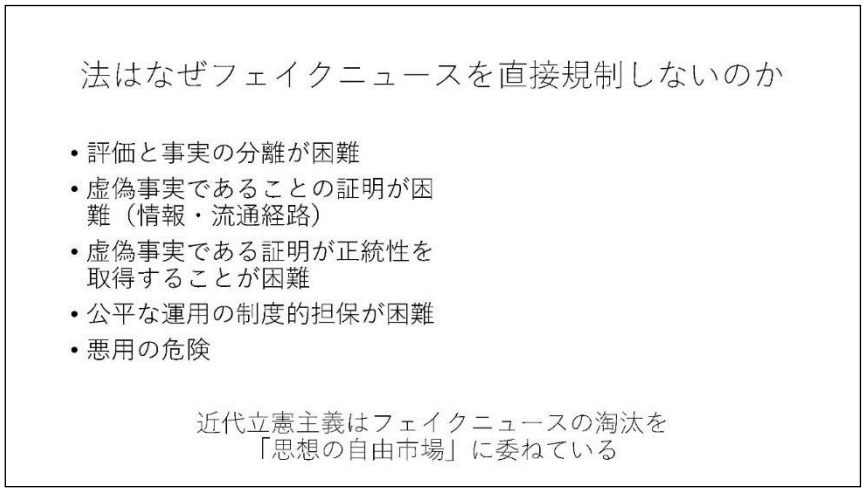


図 3-1-5　法はなぜフェイクニュースを直接規制しないのか？

これを法律家は、「近代立憲主義はフェイクニュースの淘汰を思想の自由市場に委ねている」と説明する。思想の自由市場というのは、もともとは John Stuart Mill が言い始めたが、有名なのは Oliver Holmes というアメリカの最高裁判所の判事を務めた元弁護士の言葉である（図 3-1-6）。自由市場を勝ち抜けば、その思想は正しいし、負けてしまえば、それは正しくないのだという考え方が一番いいのだという主張であり、これが立憲主義あるいは民主主義の基本をなしている考え方だとされている。

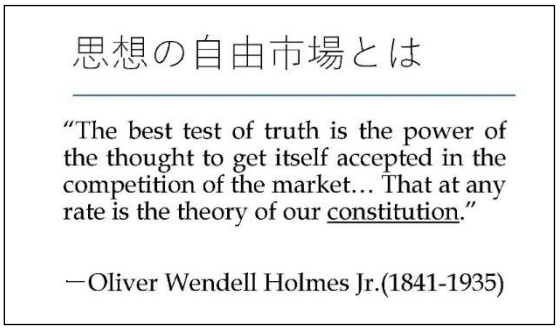


図 3-1-6　思想の自由市場とは？

しかしながら、実際にはそう簡単ではないということが問題の根本にある。フィルターバブルであるとか、エコーチェンバーであるとか、フェイクニュースだと言って言論を弾圧する人間の登場などがあり、思想の自由市場というのは本当にあるのかという疑問が起こってきている（図 3-1-7）。これを技術なり、科学の力で補完した方がいいのではないのかという議論が出てくるのはある意味、当然である。

フェイクニュースをめぐる情報技術に関して、我々法律家の立場からすると、思想の自由市場を回復するために必要なことであると言える。

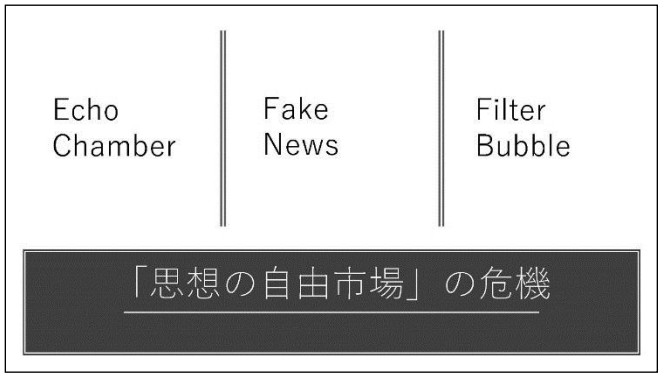


図 3-1-7　思想の自由市場とは？

つまり、思想の自由市場が機能不全を起こしているときに、このニュースはフェイクである、このニュースは正しくない、このニュースはどうも言語処理技術から見て怪しいニュースだというような情報を投入していくことは、思想の自由市場の復活・回復には意味がある、非常に意義のある研究である（図 3-1-8）。

図 3-1-8　フェイクニュースを研究する実益は？

次に、集団的意思決定を支える情報技術と法律、そして思想の自由市場との関係についても説明する。情報技術は集団的意思決定に貢献できるかということである。立憲主義国家は「思想の自由市場」を根本原理としている。思想の自由市場において、さまざまな議論を経て作られてきた意見が正しいという考え方で政治体制が作られている。

一方、AI がアドバイスするということは、「思想の自由市場」という考え方とは対立すると思われる。しかし、せっかく AI という技術があるのに、一切使わないで思想の自由市場に委ねていくのがいいのかというと、そうではないということも認めざるを得ない。思想の自由市場という重大な原理と AI をうまく折り合いをつけていかなければいけないということになる（図 3-1-9）。

現在、思想の自由市場というものが危機に瀕しているのだが、それが SNS によって、あるいはインターネットによって始まったかというと、決してそうではない。

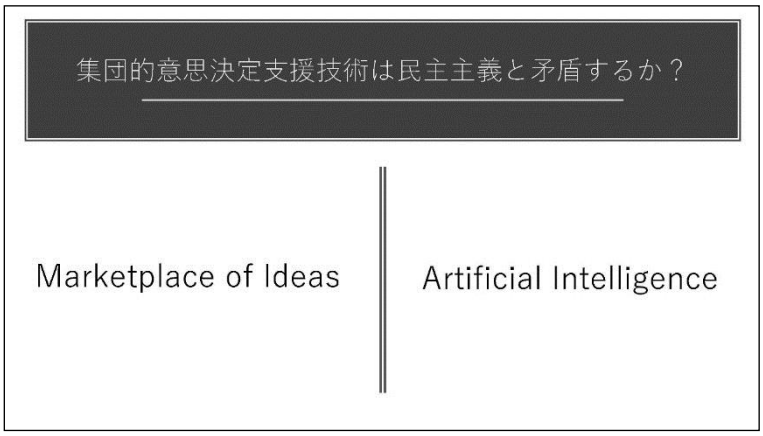


図 3-1-9　集団的意思決定支援技術は民主主義と矛盾するか？

これは Winston Churchill の言葉である（図 3-1-10）。最初の方は、民主主義に対する希望を述べた言葉である。一方、後ろの方は、民主主義に対する絶望を述べた言葉である。衆愚政治とか、大衆民主主義とかいうものはこの時代からあることを示している。

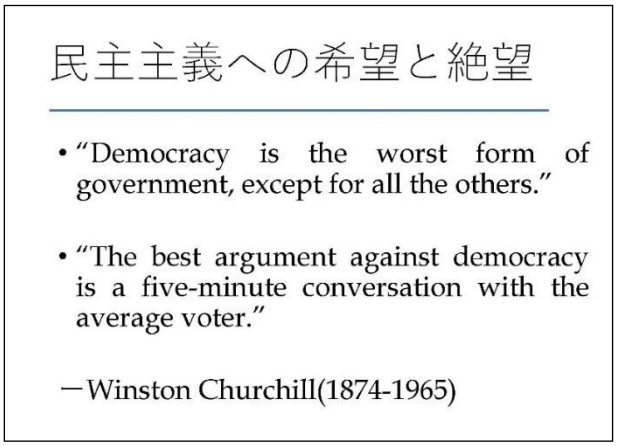


図 3-1-10　民主主義への希望と絶望

大衆がある種の情報に惑わされて集団として間違った決定をするということ自体は、少なくともここ 100 年ぐらいは普通に起きていることで、何も SNS が普及したからそうなったというわけではない。それにもかかわらず、この 100 年間、人類が「思想の自由市場」という政治原理は捨てずにやってきたということも、また事実である。

したがって、集団的意思決定を支援する技術は、「思想の自由市場」の機能不全のメカニズムを解明するということに意義がある。そして、さらに積極的には「思想の自由市場」を回復するという意味がある。ただ、思想の自由市場論を前提にする限りは、最終的な決定権は人間側にあるということを制度的に担保する仕組みが、どうしても必要である（図 3-1-11）。

したがって、集団的意思決定を支援する技術は、「思想の自由市場」の機能不全のメカニズムを解明するということに意義がある。そして、さらに積極的には「思想の自由市場」を回復するという意味がある。ただ、思想の自由市場論を前提にする限りは、最終的な決定権は人間側にあるということを制度的に担保する仕組みが、どうしても必要である（図 3-1-11）。

例えばこの秋に消費税を 10%に上げるかどうかという一つの政治的論点がある。そのときの日本についていろいろな予想がある。しかし、これはあくまで予想であって、正確に予言できる人はいない。しかし、AI が 6 割の確率で予想するようになったらどうなるのか？そのとき、我々は AI の判断を無視して 4 割側の決定を本当にできるのか？

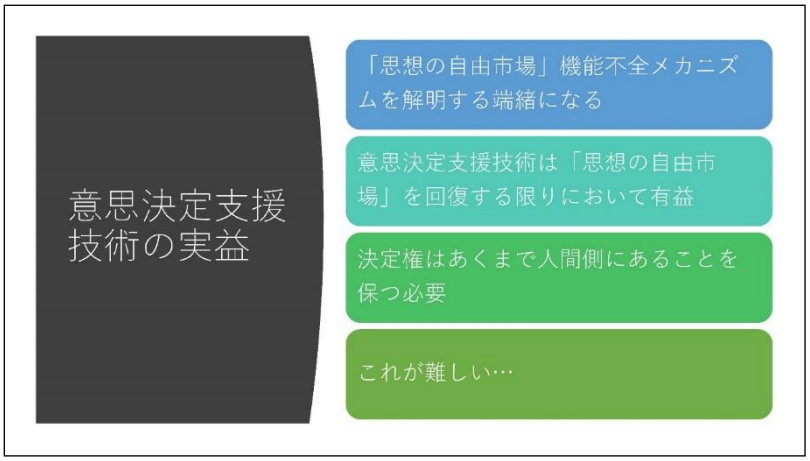


図 3-1-11　意思決定支援技術の実益

3.2　総合討論

福島　総合討論に入る。今回のワークショップ、あるいはこの総合討論で一番議論したいポイントとしては、意思決定が難しくなっている状況に対して、技術開発や研究開発をどういう方向性・やり方で進めていくべきかということである。

小林先生のお話では、技術だけで何とかなるということではないが、一方で技術に対する期待もあると。そこで、この総合討論では、まず、それぞれの発表内容について、技術だけで何とかなるような領域ではないが、こういう領域を攻めていくときの狙いどころなり、考え方などをサマライズしていただきたい。

笹原　発表した内容を簡単にまとめると、私がやっている研究は、基本的には計算社会科学という方法で、まずはフェイクニュースの実態を科学的に知るということである。特にフェイクニュースの拡散現象に着目し、認知バイアスなどの認知的な側面とプラットフォームの相互作用などについて研究をしている。

特に注目している現象としては、エコーチェンバーという現象である。放っておくと人間は同質性が高まり、集合知が生まれてこなくなる。そこに技術的に介入し、人間のもともと持っている性質を促進するような形で同質性が高まる方向を抑制し、多様性が自然に高まり集合知が生まれるようにしようとしている。

山岸　機械学習によって生成された音声、動画もしくはオンラインショッピングサイトのレビューのような偽情報を検知してフェイクを見破る技術をいくつか紹介した。音声は精度が高く、映像の精度も上がってきている。テキストはまだまだ問題がある。

この研究と並行して人間の教育を行うべきである。フェイク動画と一緒に、これが偽の動画だという情報を一緒に出すと、人間をある程度教育できる。

最終的には、人間には分からないが、機械には識別できるという状況になるかもしれない。そんな状況になったとしたら、セキュリティ側が勝つかもしれないし、そうではないかもしれない。そのような状況を見越して研究を進めていく。

今はフェイク動画の話であるが、これをフェイクニュースと融合する必要もある。例えばフェイクニュースがその動画に基づいているのであれば、ニュースが間違っている。しかし、そんな簡単なことでは全ての問題は解決できないので、フェイクニュースとの融合についても今後、考えていかなければいけない。

乾　私はファクトチェックに自然言語処理技術がどう関われるかという例を紹介した。基本的にファクトチェック自身は警察でもないし、裁判官でもない。何かを判断するのではなく、思想の自由市場の一つの情報源として機能すると考えている。たくさんある情報源の中の一つであり、間違っているかもしれない、ファクトチェックのさらにファクトチェックということもあり得ると考えている。

そうした中で、技術がどういうふうに関われるかというと、AI が得意なところと、まだまだほとんどできていないところがある。AI でできることというのは限られていて、人間に任せた方がいい、今の技術では全然、AI にはまだやらせられないことがたくさん残っている。新しいシステムを作る場合、人間に何をやらせて、機械に何をやらせるかというデザインが重要だと感じている。

伊藤　マルチエージェントシステムは社会を対象とした研究であり、私のやりたいことは、今にはない未来の社会システムを提案したいと思っている。D-Agree は民主主義の次の形、スマホとかインターネットというような技術がある上で、どんな民主主義を作り出せるのかというトライアルでもある。

これを実装し、実際に人に使ってもらって、社会で本当に使えるところはどこかというのを研究として進めたところ、ファシリテートするというのもすごく難しそうに思えるが、意見集約という中で AI と言われるプログラムがある程度は貢献できるということを示せた。

私としては、決定権が AI に移ってもいいと思っている。

森永　ビジネス的な観点から自動意思決定の話、自動意思決定システム間の挙動調整の話、特にセントラルな制御ではなく、交渉ベースの話をした。最終的に人間が関与できないと、ビジネス上も売れないので、自動意思決定にしても自動交渉にしても、どうしてこういう決定になったのか、どうしてこういう合意をしたのかというのを必要なときに人に説明できる、ホワイトボックス型になるようにアプローチしている。

この技術はファクスとか電話と同じで、世界で一人だけ持っていても、その人にとって何の得にもならないもので、できるだけたくさんの人が持っていて初めて意味がある。当社においても秘密研究ではなく、現時点からほかの会社とも一緒に、相互接続性とか相互運用性を持たせるようにしている。

機械学習で自動交渉をうまくできるようにするとすぐ談合を覚えたり、最初のうちはわざとばかな振りをして自分のストラテジーを相手に読ませないエージェントが世界チャンピオンになったりするので、本当の世の中でひどいことをやられてしまう前に、できるだけたくさんの悪いことやひどいことを安全な砂場の中でやらせてみることも必要だと思う。

福島　発表していただいた皆さんに、いろいろ振り返っていただいたけれども、それを聞いて小林先生からも何か一言お願いする。

小林　AI に合理的に動くようにプログラムをすると悪いことを平気で始めるという森永先生の発表は、大変示唆的なことだ。「思想の自由市場」論も、資本主義も、人間は合理的な存在であるから、自由な環境でそれぞれが合理性を追求すれば、世の中は全体としてよくなるのだという思想に基づいて作られた政治経済システムである。有史以来 16 世紀に人類がたどり着いた考え方が、AI によって打ち崩されつつあるということはひしひしと感じている。

いまのところ、フェイクニュースに関する法律家の態度は、比較的冷淡である。その理由は、一つは法律家が信じてやまない思想の自由市場に対する挑戦を含んでいるということであり、もう一つの理由は、結局のところ、SNS が出てきてからそういうことが話題になっているけれども、それほどたいしたことではないのではないのかという感覚があるように思える。

この感覚を少し説明すると、笹原先生の説明でうわさは早く広まるというご発表があった。だが、我々は例えば「悪事千里を走る」とか、「人のうわさに戸は立てられない」というように経験的に知っている。エコーチェンバーについても、「井の中のカワズ」ということわざがあり、ある意味、経験的に知っていることを今になって科学の方々がやり始めているということで、おそらくそんなところからあまりたいしたことではないという感覚があるように思われる。

しかし、私自身の考えは違う。フェイクニュースは SNS より前にあったことは事実だが、SNS の登場・普及によって、フェイクニュースの科学的な研究が可能になった。今まではうわさや経験、ことわざでしかなかったもの、つまり、後から検証できなかったものが、SNS の登場によって科学的な検証に耐える情報として登場した。SNS によって、どういうふうにフェイクニュースが発生して、どういうふうに広まるのかということが、経験則的にではなくて、初めて科学的に解明できるようになったことが、フェイクニュースの情報的な研究の意義であろうと考えている。

笹原　デジタルトレース、つまり、オンライン上にそういう痕跡として行動が残るようになって、初めて定量的に研究ができるようになり、計算社会科学が始まった。SNS 以前のうわさと現在のそれが圧倒的に違うのはスケールである。うわさの伝搬の分布やネットワーク構造について、ことわざは教えてくれない。有効な対策を打つためには、数理的に解明しないといけない。計算社会科学は単に定量的にやっているだけではないというのが私の考えである。

福島　リテラシーに関する質問がオンライン[1]に来ている。リテラシーの向上や教育などについて、アイデアとかコメントがあったらお願いする。

山岸　難しい質問である。リテラシーを上げるとなると、何でこれは偽の情報なのかということを説明して、見ている人が納得し、覚えていただかなければいけない。現在はそれはできていないし、実は人間の方も能力を欠いている。まず、機械の方は全部ディープラーニングを使っているので、基本、ブラックボックスになって説明可能ではない。したがって、説明可能な人工知能技術が必要になる。それがリテラシーを上げる上で役に立つ。

難しいのは人間側である。実は人間の感覚はあまり当てにならない。聴覚も視覚もチューリングテストをパスできるところまできているので、こっちが偽物、こっちが本物だと見せても人間には分からない。そういった状態で本当に見ている人が納得してくれるかどうかが

[1] 今回の公開ワークショップでは、パソコンやスマホから質問を入力できる Web 上のサービスを用いて、会場の参加者からの質問を受け付けられるようにした。

問題である。いくら説明をしても見た目が同じだから、そんなのはうそだと思われてしまうかもしれない。そういう人間の感覚と機械の感覚の違いをどう埋めていくのかというところも、今後、大事なところである。

乾　教育は非常に重要な問題である。常に違う見方があるということを肌で感じて、自分の中に定着させていくということが必要である。しかし、それを全て人手でやろうとすると難しく、機械がそれを助けて、まさにこういうエビデンスがあり、こういう説明があるから別の見方も成り立つということを説明付きで人間に提供する、そういう環境をうまく作れれば、教育という意味でも重要なことになる。

AI というのがブラックボックスであってはいけないと多くの研究者が考え始めている。どうやって説明つきの技術にしていくかということに関して、今、さまざまな取り組みが世界中でなされている。技術の説明性とか、解釈性とか、そうしたことと意思決定なり、情報の真贋みたいなものの判断というものは、非常に密接につながっている。

森永　リテラシーや教育の意味で、自動交渉エージェントに人間として勝ってみなさいと言う人たちもいる。将棋や囲碁が人間に勝てるようになった後に、それに勝てるように頑張って人間の人たちが研究して、新しい世界が開けたかのような話を聞いているので、そういう観点では、リテラシー教育という意味で、AI に勝てというのは一理あるのかもしれない。

乾　機械が得意なことと、人間が得意なことがある。ある種の数量的な最適化、あるいは予測は機械で随分できるようになってきたし、データさえとにかくたくさんあれば、人間よりはるかに合理的に最適解を求めることができる。この点においては、人間は追いつけない。

一方で、例えば物語を理解して、なぜ太郎はこのときにこういう行動をしたのかとか、そういう素朴な予測や最適化の形に落ちないような問題が世の中にはたくさんある。人間の能力ではなかなか知覚できないようなフェイク画像まで作れるようになっている一方、人間は機械が作った物語に不自然さを感じてしまう。そこには、機械が得意なところと人間が得意なところとの間には大きな溝がある。したがって、AI が全てにおいて人間の能力を超えているわけではないということを常に考えておくべきである。

福島　今回の 5 人の先生方にそれぞれのアプローチをお話ししていただいた。会場の方で、これ以外にこういうアプローチで取り組んでいるとか、自分はやっていないがこんな観点で研究を進めたらいいとか、また別の切り口から考えられるようなことがもしあったら、マイクのところで話していただきたい。

会場 1　自動交渉において合理性を追求させると悪者になるという話があった。それは倫理観を持てないところに問題があるのではないか。AI あるいはロボットに倫理観を持たせられるかどうかというのが議論なっているが、どう考えられるか。

森永　倫理観は難しいと思う。禁止というものを制約条件として外から与える、あるいは明示的ではないにしても、テストをするというのはあり得るかもしれない。しかし、談合の例もいわゆる人間社会における談合、つまり事前に会って何かを相談するというのとは違い、公開されている掲示板に人間には意味不明なビット列を投げて、それをほかのエージェントが読んだらオファーする金額を上げた方がいい、というようなことを何回もプレーするうちに偶然見つけるので、何を禁止するかは厳しいものがある。倫理観を与えて、それに基づいて自分で判断するというのは相当難しいと思う。

会場 2　最近、企業において利益追求のために検査をごまかすという事件があった。それは合理

性追求の余りに一番大事なことを忘れ、そこで倫理観を失うということがあったようである。そういうことを防ぐ方法はあるのか。

小林　自動交渉エージェントの技術は進んでいると思っているが、弁護士の仕事が奪われるとは考えていない。例えば離婚事件一つをとってみても、お金のやりとりだけであれば弁護士が出てくるまでもなく、裁判になる前にほぼ決着がつく。なぜ弁護士が裁判までやって交渉をするのかというと、人間には大事なことがいくつもあり、その中の優先順位がついたり、つかなかったり、本人の中で揺れていたり、本人が合理的だと思う優先順位が実は端から見ると違ったりということがあり、それを第三者がうまく調整しなければいけないからである。例えば親権はお母さんがもらう、お金はこのぐらいだけれども、マンションは出ていって、というような形で弁護士は交渉する。すごく生々しい話であるが、それがまさしく人間にとってのいろいろな価値観の問題である。

商行為に関する自動交渉エージェントはある意味、すごく進んでいるように見える。なぜかというと、商行為の場合、経済的合理性を追求しろという価値観が1個だけだからである。だから、AIとしては非常に仕事がしやすい。ところが、商行為ではない政治的意思決定については、例えばごみステーション一つにしたところで近くにごみ捨て場があれば、歩く距離は短いけれども、くさいとか、いろいろな価値観があり、さまざまな議論を呼ぶ。そうなると、AIには処理できなくなる。そういうことは結局のところ、人間しかできない。だから、AIが得意なところはAIに任せつつ、最後の調整は人間がやるというのがおそらく今のところの折り合いの点ではないかと思われる。したがって、倫理とか政治、正義とか環境などの最終決定権人間の方にホールドしておかないと、まだ当面は無理だと思う。

山岸　私は機械学習の方からコメントする。マルチタスクラーニングなど、一つの指標ではなく、複数の観点から学習させようという動きが最近出てきている。枠組みは既にある。できないのは価値観をどうやって定義するか、私たちの価値観をどうやって具体的な形にするか、明文化するかというところがであり、もしくはそれらをうまく統合できないので、モデルが学習できないという状況である。

伊藤　倫理というのを考えるときに有名な実験で、MITがやったモラルマシンという実験がある。各地域、例えばアフリカとか日本とかヨーロッパとか、世界全体で倫理に関する問題をいくつか参加者に解いてもらって、各地域でどんな違いがあるかというのを見たが、相当上位のモラルでないとみんな世界中で共有できない。例えばおなかにあかちゃんのいる女性を殺してはいけないというような、本当に本質的にだめなものでないと、倫理観は共有できないようである。文化とか地域によって全然違う。自分の研究に当てはめてみると、議論や意見の集約に関しても、ファシリテーションの仕方がどこまでそれぞれの社会に受け入れられるのかが重要になる。倫理観、価値観をAIにインプリメントできたとしても、それが本当に社会に対してどれぐらい受け入れられるのかというのが大事になる。

会場3　開発というのは合理性を追求していく行為である。しかし、実際にものを作って社会で試すと、ネガティブリストがたくさん出て来る。こういうふうに経験の上でのネガティブリストがモラルであり、我々社会が学んできたことであると思う。

森永　実際に、世に放って初めて分かるネガティブリストと、研究室の中でやって見つかるものとの差が大きくて、世の中で試していいのかという問題もある。

会場4　技術が進んで、AIを100%信じてAIを批判しなくなる、そして、AIに依存してしまう

人間が非常に多くなる。するとさまざまな問題が起きる可能性が出てくる。そうすると、AIの要はリテラシーになるように思える。また、対象に対して何の知識も持ち合わせていないAIが単に統計解析的に予測したものと、そのドメインに関わる非常に深い知識や、モデルを持っている学者が予測したものとを比べると、まだまだ、知識を持った人間が予測した方が正しいと思われる。その区別はつけられるのか？

福島　1点目の方は、リテラシーを上げるためにいろいろ考えなければいけないという話と関連がある。

伊藤　両方の答えになるかもしれないが、プログラミングをやると、大体、AIでできることとできないことが分かると思う。AIも基本的にはプログラムなので、できる範囲というのは分かると思う。

森永　データから帰納的に何か言うのと、物理法則とか、知られていることから演繹的に何か言うのというものを上手に組み合わせて、補い合うようなことをやるというのが結構ホットなエリアである。対象相手それぞれによってその比率とかは全然違うので、個別になる。見分ける方法は、デザインした人に聞かないと分からない。

笹原　ビッグデータで何ができるかはビッグデータを触ってみることが重要である。ビッグデータには質感というのがある。それは触った人にしか分からない。

福島　普通ならば最後に一言まとめていただくところだが、今回は一番最初に全体的なまとめ的なことを話していただいたので、これで終わりにしたいと思う。

4．まとめ

公開ワークショップ「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～」を開催した。

冒頭、CRDS から開催趣旨として、情報の氾濫、社会のボーダーレス化、フェイク情報の流通等によって、人間の主体的意思決定が難しくなり、社会の分断・混乱さえ引き起こされつつある等の問題意識を共有した。

続く 5 件の発表では、5 つの技術的アプローチが紹介された。まず笹原和俊氏（名古屋大学）から計算社会科学によるアプローチとして、フェイクニュース問題を中心に、人間の認知特性や SNS（Social Networking Service）における拡散傾向の分析や、それを踏まえた対策の可能性が示された。次に山岸順一氏（国立情報学研究所）からメディア解析技術によるアプローチとして、フェイク動画やフェイク音声の検出技術の現状が示された。乾健太郎氏（東北大学）からは自然言語処理技術によるアプローチとして、肯定・否定の両面でネット上の意見を整理して見せる言論マップや、人間が行うファクトチェックを支援するシステムの事例等が紹介された。伊藤孝行氏（名古屋工業大学）からは集合知・エージェント技術によるアプローチとして、エージェントによる自動ファシリテーションを行う大規模意見集約システム D-Agree の技術と適用事例が紹介された。最後に森永聡氏（NEC）から機械学習・最適化技術によるアプローチとして、自動意思決定システム、および、そのようなシステム間の自動交渉技術への取り組みが示された。

総合討論では、まずコメンテーターの立場で小林正啓氏（花水木法律事務所）から「法はなぜフェイクニュースを直接規制しないのか？」「フェイクニュースを研究する実益は？」「集団的意思決定支援技術の実益は？」等について、法律家の視点からの見解や問題提起が示された。これを踏まえた総合討論では、課題の重要性や技術的対策の必要性が確認されるとともに、人間・社会の側の変化や受け止め方も含めて、多面的な取り組み・議論を進めていくことの必要性が再認識された。総合討論で出された意見・考え方の中からいくつか代表的なものを、以下に示す。

- 16 世紀以降の「皆が合理的に動けば社会が良くなる」という前提で作られ到達した社会が、AI を含む情報技術の急速な発展で打ち崩されつつある。
- SNS 以前からフェイクニュースはあったが、SNS の登場で科学的な検証につながった。それが計算社会科学だが、単に定量化だけでなくスケールも違う。
- AI が得意なのは経済合理性等、単一の価値観で処理できる場面。そうではない場面は多いし、人間は合理性だけで動いているわけではない。人間と AI のそれぞれに何をさせるか考えることが大事。現状できることにムラがある。
- 機械学習では談合等の悪いことも覚えてしまう。サンドボックス上で AI に悪さをさせてみることで対策を考えるという手もあるかもしれない。
- 対策技術に取り組むとともに、一般の人間のリテラシー向上が大事。そのためさまざまな見方を知ることや、説明することが効果的。
- 一方で、フェイク音声・動画は、人間が見破ることは難しく、機械ならば見破れるような状況になりつつある。その状況で人間は納得するか、次にどうすべきか。
- MIT（マサチューセッツ工科大学）のモラルマシン実験で示されたように、文化・地域によってモラルや価値観が異なる。AI に倫理観やルールを実装したとき、それらが社会にど

れだけ受け入れられるかを考えることが重要。

フェイクとファクトを見分けることが人間に難しくなるということは、犯罪捜査・裁判等における証拠能力や、さまざまな取引や人間関係における信用・信頼が揺らぎ、また、社会的・政治的な主張や科学技術的な研究成果における捏造のリスクも高まることになる。それが悪意・扇動・干渉意図をもってソーシャルメディアで拡散されるならば、選挙を含めた社会の方向性の選択や国の安全保障にも影響を与えかねない。

このような問題は技術だけで解決できるものではないが、今回の公開ワークショップでは情報科学技術によるアプローチが状況改善に貢献し得る可能性が示された。人間は認知限界・認知バイアスを持っており、それらを乗り越えるために、情報科学技術をうまく活用していくことが役立つはずである。

本ワークショップには 150 名の参加者があり、開催後のアンケートではこのような取り組みの意義を支持する回答・意見が多数寄せられた（付録 2・付録 3）。引き続き、この問題に対する幅広い視点からの議論と、「意思決定のための情報科学」の研究開発の活性化を進めていきたい。

付録

付録１　開催概要

ワークショップ名称：

JST CRDS 公開ワークショップ
「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～」

開催趣旨：

わが国が描く社会像 Society 5.0 は「人間中心の社会」であり、社会に広がる人工知能（AI）技術に関して「人間中心の AI 社会原則」を掲げている。AI 技術が進化し、さまざまなタスクの自動化が可能になりつつあるが、自動化するタスクの定義や結果の最終判断等、上位の意思決定に責任を持つのは人間である。しかし、そのように人間の主体的意思決定が求められる一方で、情報の氾濫、社会のボーダーレス化、フェイク情報の流通等によって、人間の主体的意思決定が難しくなり、悪意・扇動意図を持った情報操作によって社会の方向性が左右される危険性さえ生じる。本ワークショップでは、このような問題・懸念は必ずしも技術だけで解決できるものではないことを踏まえつつも、人間の主体的意思決定を助けるために情報科学はどのような取り組みが可能かを論じる。5 つの技術的アプローチを示した上で、推進すべき研究開発の方向性を考える。

開催日時・場所：

2019 年 7 月 25 日（木）13:00～17:50
JST 東京本部別館（東京都千代田区五番町 7　K's 五番町）1F ホール

プログラム：

13:00～13:10　開催挨拶
　　木村 康則（JST 研究開発戦略センター）
13:10～13:30　趣旨説明
　　福島 俊一（JST 研究開発戦略センター）
13:30～14:00　発表①「フェイクニュース問題：計算社会科学によるアプローチ」
　　笹原 和俊（名古屋大学）
14:00～14:30　発表②「フェイク動画問題：メディア解析技術によるアプローチ」
　　山岸 順一（国立情報学研究所）
14:30～14:45　休憩
14:45～15:15　発表③「言論マップ、ファクトチェック：自然言語処理技術によるアプローチ」
　　乾 健太郎（東北大学）
15:15～15:45　発表④「大規模意見集約：集合知・エージェント技術によるアプローチ」
　　伊藤 孝行（名古屋工業大学）
15:45～16:15　発表⑤「自動意思決定・自動交渉：機械学習・最適化技術によるアプローチ」
　　森永 聡（NEC）
16:15～16:30　休憩

16:30～17:45　総合討論
モデレーター：福島 俊一（JST 研究開発戦略センター）
①～⑤の発表者、および、コメンテーター（花水木法律事務所 小林正啓）
17:45～17:50　閉会挨拶
木村 康則（JST 研究開発戦略センター）

付録２　参加者情報

参加者総数：150 名
内訳　6 名：招聘者（上記プログラムの発表者①～⑤、および、コメンテーター）
7 名：JST 研究開発戦略センター システム・情報科学技術ユニットメンバー
137 名：一般参加者（参加募集に申し込み、当日の参加が確認された方々）

一般参加者の専門分野[2]
26%：AI（機械学習、自然言語処理、ロボット、画像処理、データ科学等）
25%：ICT（AI を除く計算機科学、セキュリティ、通信、電子工学等）
10%：社会科学（ソーシャルメディア論、経済学、社会心理学、認知科学等）
10%：政策・行政（省庁関係者・JST 職員等で政策・行政関連に特に関わる者）
7%：業種応用（ヘルスケア、防災、生産システム、システムエンジニア等）
5%：メディア・広報
5%：教育・コミュニケーション
4%：ライフ・環境・エネルギー・材料
4%：知財・法律
2%：理学（物理・化学等）
2%：その他

付録３　参加者アンケート結果

公開ワークショップ実施の 1 週間後までの間に、Web フォームにて、取り組みの意義に関する 5 段階評価と自由記述による感想・意見を集めた。53 名から回答があり、その概要を以下に示す。

取り組みの意義についての 5 段階評価[3]
72%：大変意義がある　5 段階評価で「5」
26%：意義がある　5 段階評価で「4」
2%：どちらでもない　5 段階評価で「3」
0%：あまり意義がない　5 段階評価で「2」
0%：意義がない　5 段階評価で「1」

[2] 参加登録時に参加者本人に自由記述で記述してもらった専門分野の名称と本人の現所属の情報をもとに、人手でおおまかにクラスタリングして分類ラベルを付与したもの。厳密性には欠けるが、AI・ICT が多いものの、全体としては多様な分野から関心を持たれたという傾向が見て取れる。
[3] アンケート（Web フォーム）での質問文は「本ワークショップで取り上げた「主体的意思決定の困難化に対する情報科学技術の研究開発」に取り組む意義について、どのように感じられたでしょうか（5 段階評価）」

また、自由記述欄[4]についても、多数の感想や今後の取り組みに対する助言・示唆をいただいた。熱心・丁寧に書かれた回答が多く、今回のワークショップで取り上げたテーマが、参加者の関心にミートし、思考や議論の契機につながったように思えた。多くの示唆が得られ、今回いただいたフィードバックを今後の検討に活かしていきたい。

参考：アンケート自由記述欄の回答文字数の分布

記載なし	：19 名
1～100 字	： 8 名
101～200 字	： 9 名
201～300 字	：11 名
301～400 字	： 2 名
401～1000 字	： 2 名
1001 字以上	： 2 名

複数人からあげられた感想・意見としては、以下のようなものがあった[5]。

- 人間の意思決定やフェイク問題に対して、計算社会科学、画像・動画解析、自然言語処理、意見集約・自動交渉エージェント等の技術、および、法律家の観点からも取り組みを知ることができ、大変勉強になった。異なる分野の専門家の間の議論の機会はとても意義がある。
- 時機を得た重要な研究テーマである。日本が先導していける分野という可能性も感じた。
- 情報科学に加えて、認知科学、倫理学、神経経済学、人文社会科学等、より人間に目を向けたアプローチも重要。ゲーム理論やメカニズムデザイン等、社会デザインに関わる視点からも取り組まれている。
- 技術開発と並行して、人々のリテラシーを高めることが必要。技術による対策が考えられても、人々がその限界を理解しつつ、それを適切に使いこなせるスキルが必要になる。
- 質疑や総合討論など、議論にもっと時間を取ってほしかった。第 2 弾もやってほしい。

付録４　本テーマに関するこれまでの報告資料等

[1]　「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」、JST CRDS 戦略プロポーザル：CRDS-FY2017-SP-03（2018 年 3 月）
http://www.jst.go.jp/crds/report/report01/CRDS-FY2017-SP-03.html

[2]　「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」、JST CRDS ワークショップ報告書：CRDS-FY2017-WR-05（2017 年 10 月）

[4] 自由記述の質問文として「研究開発の方向性や重要な研究課題など、取り組みの推進に向けたご意見・コメントがありましたら、ご記入お願いいたします（自由記述）」「公開ワークショップに参加されて、ご感想がありましたら、ご記入お願いいたします（自由記述）」という 2 問を設けた。

[5] ここでは、自由記述の 2 つの設問の両方を通して、多かった感想・意見をピックアップした。ここに掲載した感想・意見は、アンケートに記入された生の表現ではなく、共通するものをある程度まとめた表現にしている。なお、多く寄せられた感想・意見だけでなく、いただいた全ての意見はワークショップ後の継続検討の中で大いに参考にさせていただいている。

http://www.jst.go.jp/crds/report/report05/CRDS-FY2017-WR-05.html

[3] 「研究開発の俯瞰報告書：システム・情報科学技術分野（2019 年）」、JST CRDS 研究開発の俯瞰報告書：CRDS-FY2018-FR-02（2019 年 3 月）
http://www.jst.go.jp/crds/report/report02/CRDS-FY2018-FR-02.html
⇒2.1.5 節「意思決定・合意形成支援」

[4] 「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」、福島俊一、人工知能学会誌 34 巻 2 号（2019 年 3 月）、pp.131～138

[5] 「OS-05: 複雑化社会における意思決定・合意形成のための AI 技術」、福田直樹・福島俊一・伊藤孝行・谷口忠大・横尾真、人工知能学会誌 34 巻 6 号（2019 年 11 月）、pp.863～869

付録５　意思決定・合意形成支援の技術俯瞰図

関連技術を広く捉えるための参考として、趣旨説明の発表の付録に掲載していた技術俯瞰図を示す。

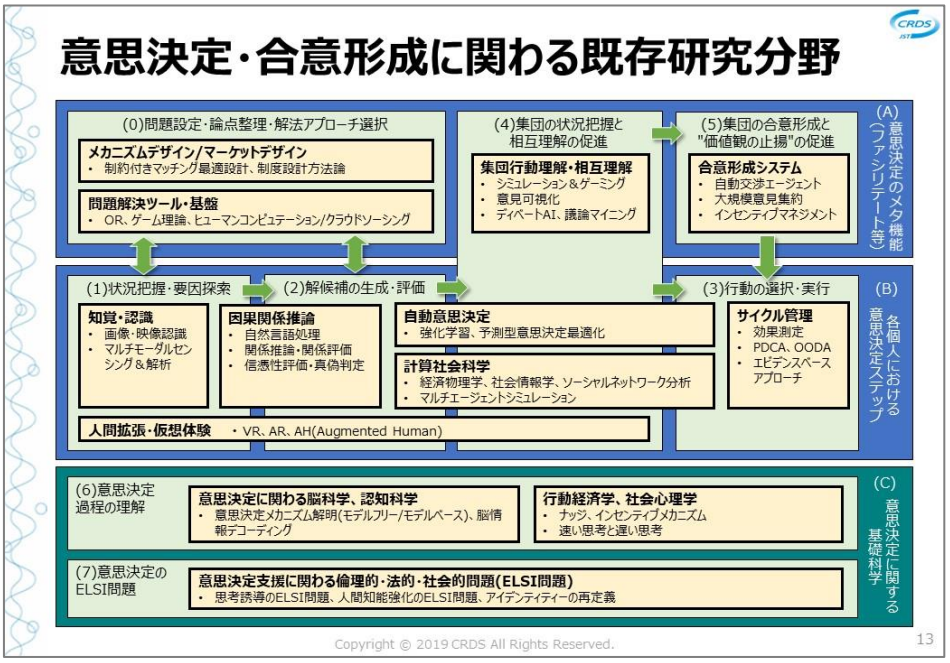


図 5-1　意思決定・合意形成に関わる既存研究分野の俯瞰図

■ワークショップ企画・報告書編纂メンバー■

木村　康則	上席フェロー	（システム・情報科学技術ユニット）
青木　孝	フェロー	（システム・情報科学技術ユニット）
井上　眞梨	フェロー	（システム・情報科学技術ユニット）
嶋田　義皓	フェロー	（システム・情報科学技術ユニット）
高島　洋典	フェロー	（システム・情報科学技術ユニット）
福島　俊一	フェロー	（システム・情報科学技術ユニット）
的場　正憲	フェロー	（システム・情報科学技術ユニット）
茂木　強	フェロー	（システム・情報科学技術ユニット）

※お問い合せ等は下記ユニットまでお願いします。

CRDS-FY2019-WR-02

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～

令和 2 年 2 月　February 2020
ISBN 978-4-88890-658-6

国立研究開発法人科学技術振興機構　研究開発戦略センター
システム・情報科学技術ユニット
Systems and Information Science and Technology Unit
Center for Research and Development Strategy, Japan Science and Technology Agency

〒102-0076　東京都千代田区五番町 7　K's 五番町
電　話　03-5214-7481
E-mail　crds@jst.go.jp
https://www.jst.go.jp/crds/

ISBN 978-4-88890-658-6