プラットフォームサービスに関する研究会（第31回）

１　日時　令和３年12月23日（木）10時00分～12時00分

２　開催場所　総務省第１特別会議室（８階）

３　出席者

（１）　構成員

宍戸座長、新保座長代理、生貝構成員、木村構成員、大谷構成員、手塚構成員、松村構成員、宮内構成員、森構成員、山口構成員、山本構成員

（２）　オブザーバー・発表者

個人情報保護委員会事務局　参事官　赤阪　晋介

法務省人権擁護局　参事官　唐澤　英城

（一財）マルチメディア振興センター　担当部長　牧野　孝

（３）　総務省

竹内総務審議官、二宮総合通信基盤局長、北林電気通信事業部長、林総合通信基盤局総務課長、木村事業政策課長、小川消費者行政第二課長、丸山消費者行政第二課課長補佐、池田消費者行政第二課課長補佐

４　議事

（１）　今後の検討の進め方について（案）

（２）　有識者による発表（関西大学 水谷准教授、国立情報学研究所 越前教授、（株）Spectee（スペクティ））

【宍戸座長】　それでは、始めたいと思います。

皆様、お忙しい中お集まりをいただきまして、ありがとうございます。定刻でございますので、プラットフォームサービスに関する研究会第31回の会合を開催させていただきます。

本日の会議につきましては、新型コロナウイルス感染拡大防止のため、一部構成員及び傍聴はウェブ会議システムに実施させていただいております。進行しなければいけない私もオンラインで、会場にお集まりの構成員の皆様には御迷惑をおかけいたします。

それでは、事務局より、ウェブ会議による開催上の注意事項について御案内がございますので、よろしくお願いいたします。

【池田消費者行政第二課課長補佐】　宍戸先生、ありがとうございます。総務省の消費者行政第二課、池田でございます。

ウェブ開催に関する注意事項を幾つか御案内させていただきます。

まず、本日の会議の傍聴につきましては、ウェブ会議システムによる音声及び資料投影のみでの傍聴とさせていただいております。事務局において傍聴者は発言ができないよう設定させていただいておりますので、音声の設定を変更しないようにお願いいたします。

次に、構成員におかれましては、ハウリングや雑音混入防止のため、発言時以外はマイクをミュート、ビデオをオフにしていただくようにお願いいたします。御発言を希望される際には、事前にチャット欄に発言したい旨を書き込んでいただくようお願いいたします。それを見て、座長から発言者を指名いただく方式で進めさせていただきます。発言いただく際には、マイクをオンにして、映像もオンにして御発言ください。発言が終わりましたら、いずれもオフにお戻しください。接続に不具合がある場合は、速やかに再接続を試していただくようお願いいたします。その他、チャット機能で随時事務局や座長宛てに御連絡をいただければ対応させていただきます。

本日の資料ですけれども、資料、本体資料として資料１から４、また参考資料１を用意しております。

注意事項は以上になります。

それでは、これ以降の議事進行を宍戸座長にお願いしようと存じます。宍戸座長、お願いいたします。

【宍戸座長】　承知しました。本日は、まず、今後の検討の進め方について事務局より御説明をいただきたいと思います。

次に、有識者からの御発表として、関西大学、水谷准教授から、「オンライン言論ガバナンスの透明性と適正化に向けて」について、次に、国立情報学研究所、越前教授から、「顔を対象としたフェイクメディアの生成と検出について」、最後に、株式会社スペクティから、「SNSによるデマ情報拡散のメカニズム」について御発表をいただきます。それぞれ有識者の発表の後、質疑の時間を設けますので、構成員の皆様から御意見等をいただきたいと思います。

本日も盛りだくさんでございますので、円滑な議事進行に御協力を賜ればと思います。

それでは、早速、議事の１でございますが、今後の検討の進め方について（案）を事務局より御説明いただきたいと思います。それでは、お願いいたします。

【池田消費者行政第二課課長補佐】　宍戸座長、ありがとうございます。消費者行政第二課、池田でございます。

資料１につきまして、御説明をさせていただきます。

今後の検討スケジュールという資料で、ページ番号右上に１とあるページを御覧ください。

今後の検討スケジュール案をこちらでお示ししております。左下、12月23日、本日ですけれども、進め方というところでの案をお示しし、有識者の皆様からの御発表を伺おうと思っております。

今後のスケジュールでございますけれども、２月頃におきまして、誹謗中傷関係の事業者団体のヒアリングを行いたいと考えております。

また、３月頃においては、フェイクニュース、偽情報対策関係のヒアリングを行いたいと考えております。

これら、事業者団体、事業者からのヒアリング、モニタリングを踏まえまして、４月から５月頃、それらに対する御議論、また、海外制度の検討状況等についての御議論、論点整理等を行いたいと思っております。こちらにつきましては、モニタリングの結果等に応じて多少アジェンダの調整や時期の調整をさせていただきたいというふうに思っております。

５月中には、第二次取りまとめ骨子ということで、ヒアリングした内容、また、海外の制度等についての論点整理を行った結果の部分を骨子として取りまとめたいと思っておりまして、６月中には第二次取りまとめ（案）ということでレポートの形で取りまとめまして、パブリックコメントを行ってまいりたいというふうに考えておるところでございます。

今後、プラットフォームサービスにおける研究会につきましては、このようなスケジュールにおいて進めてまいりたいと事務局として考えております。

以上でございます。

【宍戸座長】　ありがとうございました。

それでは、次に、有識者から御発表をいただきたいと思います。そして、それに質疑をするということでございます。

それでは、資料の２につきまして、関西大学、水谷先生、どうぞよろしくお願いいたします。

【水谷氏】　ただいま御紹介にあずかりました関西大学社会学部准教授の水谷と申します。今日はよろしくお願いいたします。

それでは、本日は、「オンライン言論ガバナンスの透明性と適正化に向けて」というタイトルでお話をさせていただければと思います。

本日、大きく分けて４項目でお話をさせていただければと思います。

まずはじめに軽く自己紹介をさせていただきますと、私は、2019年の４月から関西大学の社会学部メディア専攻でメディア法と情報法の担当ということで移籍をさせていただき、現在に至ります。それまでは帝京大学の法学部におりました。慶應義塾大学の大学院で博士（法学）をとっており、専門は憲法やメディア法ということになります。もともとインターネットやAI技術の普及した世界で報道機関とかデジタル・プラットフォームの憲法上の機能をどう考えるかということを研究してまいりました。共著本として、『ＡＩと憲法』や『憲法学の現在地』といったものに書かせていただいております。

それでは、早速ですけれども、本日の発表の中身に入っていきたいと思います。

最初は、もうここにいらっしゃる先生方には釈迦に説法のようなお話になるかもしれませんけども、私の専門の憲法学の観点から現代の表現環境というのはどういうふうに位置づけられるのかというお話をさせていただければと思います。

私の見方ではメディア環境は３段階ぐらいあるかなという感じで考えておりまして、これまで、つまりデジタルメディアが普及する前の時代の表現環境ですが、この時代では、中間項としてのマスメディアが、特に新聞やテレビが中心ですけども、非常に大きな存在として表現環境の中に鎮座していました。これらの機関が、社会において情報発信のための媒体を基本的に独占してきた部分がありますので、多くの一般市民は、マスメディアにアクセスできない限りは発信の機会を持てないというような状態でした。受領も、基本的

にはどの新聞を読む方かとかどのテレビのチャンネルを見るかというようなレベルでしか選択権がないというような状況だったわけであります。

その後、デジタル革命というふうに言うと大げさかもしれませんけども、インターネットの普及によって環境は大きく変わりました。万人が公平に利用可能な情報発信、受信のインフラが誕生したわけです。これによって、マスメディアが社会的にも、あるいは法的にも、今まで占めていた地位が低下していくということが見受けられるのではないかと考えております。

さらに最近、ビッグデータあるいはAI技術の普及ということが言われ、それを前提にしてSociety5.0ということが言われておるわけですけども、この社会での表現環境というのは、個々のユーザーの選好や行動がデータによって予測をされるようになってきて、パーソナライゼーション、すなわちユーザー個々に最適な形で情報が届けられるようになると考えられます。

こうしたデジタルメディア革命以前と以後を比べてみたときに、一つポイントとなるのは、憲法学が、特に表現の自由の領域で非常に重視してきた思想の自由市場論という考え方です。これはアメリカ憲法判例の有名なフレーズで、基本的には言論環境というのは政府による介入・規制を入れずに、なるべく自由な状態にしておいて、そこで悪しき言論とか情報というのは言論同士で競争させればいいと。その結果、勝ち残ったものが真理である、あるいは、いずれ真理は勝ち残るはずだというような考え方で、基本的には表現空間を政府介入から自由な状態に置き、表現者に対する強力な権利を与えて、特にコンテンツ規制を可能な限り排除しようとした考え方であったわけです。

一見するとこういった考え方と、インターネット時代の万人が利用可能であるというようなインフラが普及したというのは非常に親和的に見えるわけです。アメリカのインターネットが普及し始めた草創期の有名な判決で、Reno判決というのがありますけども、この中にもインターネットのことを新しい思想の自由市場だというような言い方がされてきたわけであります。

一方で、私の中では、ちょっとこの考え方というのが現代においては様々な側面から楽観主義的なんじゃないかなと考えております。

大きく分けると３つぐらいあるかなと思うんですけども、１つはデジタル空間というのは非常にアルゴリズミックな性質を持つんだと。イーライ・パリサーがフィルターバブルと言ってからもうかなり時代がたちますが、憲法学者のキャス・サンスティーンもデイリ

ー・ミーを指摘してきました。デジタル空間の情報流通というのがますますアーキテクトによって設計された場の上で行われると。大体これが現代においてはデジタル・プラットフォームがそのアーキテクトになっているわけですけども、そうである以上、そこでのアーキテクチャの設計によって、ユーザーのオンライン体験が変わってくるということを前提にする必要があると思います。

もう１つは、これは私が論文で書いたのですけども、情報資本主義が到来してきて、思想の市場と経済市場の間の垣根がだんだんなくなってきているんじゃないかということです。今まで、アメリカではこの２つは一応理論上きれいに分けて考えて、思想の自由市場には非常に強力な自由を与えて、一方で、経済市場に関してはある程度の規制を認めるというような区別がされてきたわけですけども、ますます現代においては、より多くのデータを握って、そしてそれを的確に駆使できるものが表現活動のみならず経済活動をも有利に進めることができるようになるといういうことを考えると、この垣根があまり機能しなくなってきているのかなということが見受けられます。

さらに、この背景的要因として最近注目されているのがアテンション・エコノミーです。デジタル空間における市場競争での生き残りは粘着性が重要であるといういうのが、マシュー・ハインドマンが指摘をしていますけども、これに合わせて、山本龍彦先生も、アテンション・エコノミーが支配した表現空間を思想内容の競争から刺激の競争になっていると指摘しておられます。偽情報に引きつけて言えば、例えばより効率的に私たちの希少な注目を奪えるような、センセーショナル、またサブリミナルな形態でそういう情報が発信されるようになってくることを考えなければならないと思います。

こういった前提を考えると、今までのように思想の自由市場の機能を無邪気に捉えることはあまり得策ではないと思います。むしろ、思想の自由市場があまり機能しないことのほうが前提になるんじゃないかと考えるわけです。そうすると、何が困ったことが起きるかというと、憲法学的には、民主政を支えてきたのが、ある種理性的な市民間の熟議だということを考えると、そこへの世論形成における熟議みたいなものがうまくいかないと。例えばフィルターバブルも、この観点から見ると、自分と異質な他者と接触して理性的に意見を交わしていくんだというような機会が、やっぱりフィルターバブル的なもので失われがちになるというようなことを考えると、ますます民主政的なものを支えている熟議が、これから機能しにくくなるんじゃないかと指摘できると思います。

さてここまでお話したような表現環境を現代においてかなり多くの部分で支えているの

がデジタル・プラットフォームということになると思います。

このデジタル・プラットフォームの定義というのはなかなか難しくて、いろんなビジネスモデルが当然あるわけですけども、政府のほうで示された「プラットフォーマー型ビジネスの台頭に対応したルール整備の基本原則」をいつも僕は参考にしておりまして、「社会経済に不可欠な基盤を提供していること」、「多数の消費者（個人）や事業者が参加する場そのものを、設計し運営・ 管理する存在であること」、「そのような場は、本質的に操作性や技術的不透明性があること」という３つで大体特徴づけられていると考えています。今回のお話の中のデジタル・プラットフォームは、SNSとか動画共有サービスといった、直接的に表現活動の場として機能しているソーシャルメディアに着目をしたいというふうに思います。

そう考えますと、憲法学とか、あるいはメディア法とか領域でプラットフォームはどう位置づけられるんだろうかということがあると思うんですけども、これが今のところ日本の判例ではあまり定かではありません。アメリカにはパッキンガム判決という判決が出ておりますが、その中でソーシャルメディア、SNSは、「モダンパブリックスクエア」というような評価をされているわけですけども、日本で探してみると、最近のグーグル検索結果削除決定ですね、この中に、検索エンジンに関してですけれども、現代社会において「インターネット上の情報流通の基盤」として大きな役割を果たすのだというような指摘があります。一方でマスメディアの位置づけをした博多駅テレビフィルム提出命令事件というのがあるわけですけども、この決定の中では、報道は「国民の『知る権利』に奉仕する」ものであるという言われ方をするわけです。検索エンジンも、そのユーザーが知りたいものを提供するという意味でいうと、「知る権利」に奉仕しているというふうに言ってもいいはずなんですけども、前者には国民の「知る権利」という言葉は出てこないわけです。

この違いは何なのかという点なんですけども、やはりそれぞれが担っている機能が違うためじゃないかと私は見ています。プラットフォームは、グーグル検索もSNSもこの点では変わらないと思いますけども、資料にマーク・ザッカーバーグが語ったと言われている言葉を示しておりますが、ようするに、あなたが興味関心あるものをフィールド上に出しますよということだと思います。あなたが「知りたいと思っているもの」をいかに提示するかということが、SNSなどの機能を表しいると。

ですので、報道機関も情報の取捨選別は「編集」という形でしますけども、デジタル・プラッフォームによる選別は、今日のコンテンツ・モデレーションもそうですけども、基

本的にはユーザーの「知りたい」、もっと言うとユーザーが「心地よくなる」ものを目指すというのが前提であると思います。

他方でジャーナリズムを基幹としたアクターである報道機関は、資料に示したニューヨーク・タイムズのスローガンにも表れているように、こちらはどちらかというと、社会にとって必要なもの、権力批判ですとか、あるいは個人や社会の発展に資するというような報道価値といった観点に基づいて選別をしているわけです。ですから、どちらかというと「国民に知らせるべき」というか知っておいてもらわなきゃ困るみたいな情報を選別してきたわけです。このような違いが両者には見受けられるかなというふうに考えています。

さらに、デジタル・プラットフォームは、現代においては空間における情報流通の場の管理者としての側面を強く有しているというのが指摘できるかと思います。

それが資料の第二章になります。その意味でいうと、コンテンツ・モデレーションいう点が非常に重要になってくるかと思います。デジタル・プラットフォームのうち、特にSNSは、コンテンツの削除や、あるいはアカウントの凍結みたいなのを自社のポリシーに合わせて非常に手を入れているわけです。実はここが今回の発表の話にもつながってくるわけですけれども、そのプロセスが不透明であるというのが言われていまして、実際、DPFの中で何をどうやってモデレーションしているのか、どういう仕組みでやっているのかというのが外からあまりよく分からなかったんです。これを結構鋭く描き出したのが、アメリカのケイト・クロニックという研究者です。「the New Governors」という論文がハーバード・ローレビューに出ておりますけども、この論文を読むと、フェイスブック等がどうやってコンテンツ・モデレーションをしているのかというのが非常によく分かります。いろいろあるんですが、中央集権的な組織がやっぱりちゃんとフェイスブック等にはあって、その下、表に出しているコミュニティ・スタンダードよりもさらに細かい内部ルールが実はあると。しかも、人間のモデレーターがコンテンツ・モデレーションをやっているので、このモデレーターさんたちにちゃんとトレーニングをしているんだというんです。この訓練は法的な思考訓練するのと非常によく似ているみたいなんですけども、とにかくこのコンテンツが基準に合っているかどうかを細かくチェックしていくというようなトレーニングをしっかりやっていると。さらに、モデレーターも３層からなっているらしく、ティア１からティア３までいると。一番下のティア３の人たちが一番多いわけですけど、まず私たちがフェイスブック等でこのコンテンツは違反じゃないかといって違反通報すると、このモデレーターさんたちがそのコンテンツをチェックをすると。分からなかったら、

さらに上の段階にエスカレーションしていくんです。最終的には、本社のロイヤーがいるみたいな、そういう話でした。

さらに、こうした基準は結構外的な影響を受けて頻繁に変わっているという話なんです。外部からの影響を、民意と言っていいか分かりませんけども、フェイスブック等のSNSは実は外の意見を取り入れていると。ただ、その取り入れるポイントが結構万人から民意を吸い上げるというよりは、簡単に言ってしまえば、フェイスブック関係者と知り合いであるとか、あるいはマスメディアのように影響力があるところで指摘をできるとか、そういう回路が、結構外的影響として重視されてきたようです。

クロニックは、この点から、プラットフォーム・アクターは今までの表現の自由のアクターとはちょっと違って、やっぱり「新たな統治者」なんだという言い方をするわけです。そして彼らがコンテンツ・モデレーションを非常に大きなコストをかけてやっているわけですけども、それはなぜなのかというのは、やっぱりユーザーの期待に応えるためというのが大きいんだということです。これはなぜそうなるかというと、ユーザーの期待に応えると、さっきのアテンション・エコノミー的な部分も含めて、経済的な利益として自社に返ってくるわけです。彼らは経済的な論理が働いてこういうコンテンツ・モデレーションをやっている部分もあるんだという点は重要かなと思います。

さらにクロニックの論文ではあまり指摘をされてないんですが、ここがAIによってどんどん自動化されていく可能性が高いという点も指摘しておきたいと思います。

ですので、今の前提を見ると、プラットフォーム・アクターの情報流通の基盤というのは、ある種の多くのユーザーに心地よく場を利用してもらうために場をデザインして、そこからコンテンツポリシーも含めて管理運営をしていく機能だと言えると思います。

これが、ある日、ユーザーである私たちにとって牙をむく場合も当然あるわけで、結論的にはよかったと思いますけど、それで大騒ぎになったのがトランプ元大統領のSNSアカウントが凍結された問題なわけです。

こうした点から、私は、論文や新聞で、DPF事業者の「手のひらの上の自由」しか僕らユーザーにはないよねという話をしたことがあります。

このようなインターネット上のコンテンツのガバナンス、DPFが行っているようなガバナンスというのは、今まで法学者が前提にしてきたようなものと少し異なります。つまり我々には表現の自由の権利があって、場合によって名誉権を侵すような、権利侵害を侵すような情報を流すとそこで調整が入るというような、個々の「権利ベース」の議論から、

どちらかというと「パブリックヘルス」の時代に行っているんだというようなことが、最近ジョナサン・ジットレンによって指摘をされているわけです。このフレームワークは、どちらかというと個人の権利がどうかというよりも、そのプラットフォーム全体のリスクとかベネフィットを重視してバランスを取るというようなことが前提になっていると。

さらに、最近エブリン・ドゥエクという若手の非常に優秀な研究者がいるんですが、彼女も似たようなことを言っていて、コンテンツ・モデレーションの世界は、個人の権利ベースの世界というよりも、どちらかというと確率論、エラー率の世界だと。どれだけ違法なものとか有害なものを逃さずスクリーニングできるかというようなところで、これはエラーの割合が多かったらもうちょっとアルゴリズムを見直して調整しようとか、そういう観点で統治をしていくんだと。ですから、どちらかというと非常に「システミックな世界」なんだという指摘がされるわけです。

これは何でこういうふうになっていくかというと、２つのポイント、つまり規模の問題とツールの問題があると思います。

規模の問題では、ヤフーニュースのコメント投稿を例に取ると、ひと月で1,050万ある。そのうち、ひと月当たり35万件を削除しているという話なんです。これを権利ベースで１個１個見ていくということは当然不可能だろうという話になるわけです。

それから、スクリーニングしているツールが、さっきも申しましたように、AIを導入してきているので、そこでのコンテンツ・モデレーションというのは、基本的に個人主義ではなくてシステミックなものとして行わざるを得ないと。さらに、判断がそのプラットフォーム全体を考慮して行われるので、１個１個ケース・バイ・ケースで見ていくわけではないんだということになるわけです。そうすると、このプラットフォームの権利裁定は非常にアクチュアリアル（保険数理的）なものになるという指摘があるわけです。

ここでこういうシステミックなアプローチは、前提としてエラーが必然的に生じるので、このエラーを許容して、それをオンライン言論空間のガバナンス設計に組み込んでいくということが重要になってくるということになります。

こういった「ガバナンス」という視点で重要になってくるのが「プロセス」ということになります。オンライン言論ガバナンスの「プロセス」において、エラー率を低くするようなどういう取組をしているのかとか、そういったものを細かく透明性を高めていくということがガバナンスの点で重要になってきますし、プロセスの設計が誰によってどういう議論の下でつくられたのかというようなこともユーザーにとっては非常に重要になってく

るわけです。

すなわち、個々の決定には不満があっても、全体としてユーザーがこのプロセスには従えるなというようなある種の正統性ですね、これを持てるような形にガバナンスを高めていくというのが重要になってきます。

この点でちょっと御紹介したいのが、コンテンツ・モデレーションに対する国際的動向です。私は、最近オフィシャルな制度よりも、こうした市民団体とか学識者のグループが出している原則が面白いなと思ってちょっと注目をしているので、それを御紹介させていただければと思います。

ちょっと時間が少なくなってきましたので、マニラ原則の説明は飛ばして、次のサンタクララ原則の説明に参りたいと思います。

サンタクララ原則というのは、2018年にカリフォルニア州のサンタクララで行われた有識者のカンファレンスをきっかけに、その中のグループから提唱されたというふうに言われています。最初は資料にあがっている通り、３つだったんです。数値をいろいろ公開しろとか、ユーザーにちゃんと告知をしろとか、加えて興味深いのはアピールというが入っているんです。つまり、コンテンツ・モデレーションとかアカウント凍結に対して、まさにデュープロセスなんですけども、異議申立て機会をちゃんと与えるというようなことがこの原則では入っていたということです。

これが2021年現在にアップデートをしておりまして、今現在、サンタクララ原則2.0というのが出ているようです。

この2.0は、ちょっと項目が多いので全部御紹介し切れないんですけども、あと、私自身も研究中ではあるんですが、大きく分けて基本原則と運用原則という形からなっています。基本原則は、５原則ぐらい、企業に向けた理念と、ある種実践的なものを細かく書いているというような形でできているんですが、もう１つ、運用原則が、先ほど表した３つ、これをベースにさらに細かいところがいろいろ出てきているということで見ていきたいと思います。

まず、運用原則のナンバーと呼ばれるものですけども、ここではどういうものを情報公開、要するに透明性を高めればいいかということが示されています。情報の透明性を高めろといっても、何を公開すればいいのかというのがあまりよく分からないということで、一部だけ抜粋していますけども、例えば措置されたコンテンツとか停止されたアカウントの総数だとか、それに対して異議申立がどれだけあったかとか、異議申立が成功した、あ

るいは失敗した割合はどれぐらいだったかとか、そのうち自動検出がどれぐらいだったかとか、こういったものを公開せよということが言われています。自動化プロセスに関する情報も、いろいろこれから公開する必要があるよというのが運用原則で言われているわけです。

２番目のノーティスの部分は、これはユーザーに対してどういう情報を提供すればいいかという話です。これは当然措置を受けたユーザーに対して、コンテンツを識別するために十分な情報を与えたりとか、どのガイドラインに反したのかというのが分かるようにしたりとか、このコンテンツの削除が、要するに違法なものとして削除されたのか。そうではなくて、プラットフォーム上の有害ポリシーに違反したから削除されたのか、こういったものの区別をちゃんと分かるようにしろと。それから、国家の関与があったかどうかや、あるいは自動的な検出で消されたのか、他のユーザーからの通報を受けて人間がモデレーションして消されたのかというのとかをちゃんと説明しなさいということになっています。もう１つ重要だなと思ったのは、アカウント停止した場合でも、そういうものを見れるようにしなさいと。大体アカウントが止まっちゃうと、プラットフォーム上で何でアカウントが止まったのかというのを見るということが、なかなか難しくなる場合もあるので、そこをアクセス可能にしておきなさいみたいなことが出ております。

最後に、異議申立てという仕組みがあるわけです。この異議申立ては、コンテンツ・モデレーションが削除された後に、それに対して、これはポリシーに違反していないんじゃないか、おかしいんじゃないかというのをアピールするような仕組みがプラットフォームにはあったりするわけですけども、それに関して、今現在どうやって異議申立てを処理しているかというのを、タイムラインをちゃんと確認できるようにしなさいとか、どういう人がパネルとして関わっているか、特に最初のモデレーションに関わっていない人がちゃんと審査に加わっているかどうかといったところを公開せよと。これはユーザーに多分通知しなきゃいけないわけですけども、そういったものを見せなさいということになっているわけです。

興味深いのは、この後に御紹介しますけども、長期的には独立した審査プロセスが重要な構成要素になる可能性がありますみたいなことが、2.0で加わっていて、これは何なのかなというのが、実は次の部分と関係あるんじゃないかと思います。

それが、最近私も論文で書いてみたんですが、フェイスブックに「最高裁」ができたと。フェイスブック最高裁というような評し方がされるわけですけども、「監督委員会（オー

バーサイトボード)」というものがフェイスブックの中から出てきたわけです。この監督委員会のことを指して、さっき言った独立した審査プロセスというふうに言っているんじゃないのかなというふうに僕は見ているんですけども、まさにフェイスブックから、財政的にも、あと審査する有識者の人事的にも独立した監督委員会という別の外部審査機関をフェイスブックがつくって、そこに異議申立てとかを、場合によっては審査してもらうということを今つくっているわけです。

このフェイスブック監督委員会の発案者が、ハーバード・ロースクールのノア・フェルドマンという人だと言われていまして、彼がフェイスブックに最高裁が必要だとぱっと思いついて、シェリル・サンドバーグにメモを渡したところ、似たような考えが多分フェイスブックの中にもあったんでしょうけども、それらをきっかけに今のフェィスブック最高裁みたいなものが出来上がってきたということです。

もちろんこういったものがどこまで機能するかというのは、実は怪しいところもあって、まだ現在のところはさほど多く審査結果が出されてなくて、フェイスブックのコンテンツ・モデレーション全体からしたら全然数的にも不十分なんですけども、最近、先ほど挙げたエブリン・ドゥエクが、この仕組みはオンライン言論ガバナンスをするうえで重要な意味があるんじゃないかというのを２つ挙げているわけです。ひとつは立法プロセス、つまりポリシーを形成するときに、いろいろ慌ててつくるから抜け穴があったりするので、それをこういう外部機関が除去してくれたりといった機能がある。あるいはもう一つ、さっきの正統性とも関わってきますけども、ユーザーが受入れやすいような公共的な理由づけのプロセスを議論する場所になりえるというような指摘がされているところです。

最後ですが、現代においてこういうコンテンツ・モデレーションのプラットフォームが積極的にやっているという前提を踏まえた上で、偽情報や、あるいは誹謗中傷に関しても、コンテンツ規制をどれだけやっていくかということがあるわけですけども、基本的に政府による発信者のコンテンツ規制は注意しなきゃいけないと。かなり慎重に慎重を重ねるべきだろうというのが私の見解です。

それは、やっぱりコンテンツ・モデレーションの世界は、さっき言ったようにシステミックなものですので、政府が発信者に対する法規制をやっても、あまりメリットとリスクが釣り合わないのではないかなというようなところがあると思います。

同時に、DPFにコンテンツの削除義務を課すというのも、このコンテンツ・モデレーションを政府が強化してしまうということになるので、DPFの検閲代理人化やオーバーブロ

ッキングの懸念が出てくるので、やってはいけないわけではないと思いますけど、かなり慎重に考えなきゃいけないと。特にこの場合は、さっきのサンタクララ原則のところで実はちょっと飛ばしてしまいましたけども、政府自身もこういうコンテンツ・モデレーションにどれだけ関与しているかを透明性高めろというようなことの趣旨が書いてありますので、そういったことが当然、こういうことをやるんだったら必要になってくると考えます。

また、コンテンツ・モデレーションのやり方は削除以外にもいろいろあるわけです。警告表示とか、誹謗中傷対策のReThinkなんかもまさにその典型ですけども、そういった点から考えると、DPFの機能的な自律性を確保しながら、さっき言ったようなサンタクララ原則などを参考にして、透明性とかデュープロセスの実現を促進していくというような、その面で政府はルール形成に関与していくということが１つあるんじゃないかなと考えております。

ちょっと時間オーバーしてしまいました。大変申し訳ありませんでした。

私からの発表は以上になります。

【宍戸座長】　水谷先生、最新の研究成果を御披露いただきありがとうございました。

それでは、ただいまの御説明について、構成員の皆様方から御質問、御意見をいただければと思います。例によりまして、御質問、御意見のある方はチャット欄で私にお知らせをいただければと思いますが、いかがでございましょうか。

宮内先生、お願いいたします。

【宮内構成員】　宮内でございます。大変興味深い内容をありがとうございました。

１点ちょっと質問させていただきたいんですけども、このプラットフォーマーが、仮に、ある種の思想的な言論を封鎖するような、極端なことを言えば、共産主義のやつを全部消すとか、あるいは逆に国粋主義の部分だけ消すとか、そういうようなことをやろうとしたときに、それが公然とであれ、暗黙のうちであれ、そういうのは、言ってみれば私企業の自由というふうに見るべきなのか、これも言論の自由に反するから、それはある程度抑えられなきゃいけないのか、この辺については、どういうふうにお考えでしょうか。

【水谷氏】　御質問ありがとうございます。大変難しい御質問をいただいたというふうに思います。

まさにおっしゃるとおりで、コンテンツ・モデレーションを、現状はプラットフォーマーは自分たちの場をユーザーに多く安心して利用してもらうためにやっていますので、ルール上、誹謗中傷対策とか、あるいは最近だとCOVID-19の偽情報、こういったものの対応

をするという形でやっているので、そこの点は今のところは懸念になってないわけですけども、確かにおっしゃるとおりで、そういう、ある特定の思想とかを排除するというような形というのは当然あり得るわけですよね。共産主義もそうかもしれませんし、あるいはナチズムのようなものに対しては、ヘイトスピーチにどう対応するかという観点からも当然あり得ると思います。

難しいのは、まさに監督委員会もそういうところで壁にぶつかっているわけですけども、どういう言論を違法とするか、アウトとするかというのが、やっぱり世界各国で違うわけですよね。当然、ヨーロッパとアメリカでヘイトスピーチに対する許容というのは全然違ってくるわけですし、国によっては特定の思想が駄目だということで、まさに検閲代理人化して、政府からの圧力でその特定の思想を消すみたいなことがあるとも言われていますので、そういった面でも、プラットフォーム自体がグローバルな側面を持ちますので、その国の表現の自由に合わせて、この範囲しか消しちゃ駄目というような規制をかけるというよりは、どちらかというと何をどれぐらい、どういう割合で消しているのか、消しているものがどういう内容で消されているのかという透明性を１社１社高めていくと。それによって、まさにプラットフォーム間の多様性を確保するというのも重要になってくると思うんです。

つまり、あるプラットフォームでは、ある特定のコンテンツは厳しく規制されるけど、別のSNSではその部分は緩いというような、プラットフォームごとの多様性みたいなものを確保すれば、ユーザーとしては、どこのプラットフォームを利用するかみたいな選択が可能になるので、そういった点でちょっと確保していくべきかなというふうには思っております。

【宮内構成員】　大変よく分かりました。どうもありがとうございました。

【宍戸座長】　ありがとうございます。時間の関係もありますので、この後、新保先生、手塚先生、森先生から御質問希望をいただいていますので、それぞれ御質問、御意見いただいて、それでまとめて水谷先生にお答えいただくことにしようかと思います。

それでは、新保座長代理、お願いいたします。

【新保座長代理】　新保です。水谷先生、ありがとうございます。対面で御礼を申し上げることができてうれしいです。

私は、質問ではなくて、意見として、今日の御発表大変参考になりましたので、それについての意見を述べさせていただきたいと思います。御紹介いただきました原則のように、

原則策定の取組とともに、いわゆる原則ブームといいましょうか、原則という名のブレーンストーミングが非常に多くなってきている状況かと思いますけれども、それに伴って、実効性ある規律としての機能を重視しない検討が増えているように感じております。

一方で、表現の自由の保障について、そのガバナンスの透明化と適正化は極めて慎重な検討が求められる問題であると言えると思います。

この問題について考えるに当たって、例えば従来プライバシーの権利をめぐる問題については、いわゆるビッグブラザーの懸念というものが示されてきたわけであります。ところが、いつの間にかこのビッグブラザーの懸念以上のことをDPFが実現しているようにさえ思えるわけですけれども、つまり、今後、表現の自由の保障における問題、この問題、プライバシーの権利をめぐる問題の従来の議論を踏まえて考えてみると、ビッグブラザー規制であれば、憲法上、より明確に個人の権利を保障するための取組ができたのではないかと思います。

ところが、政府によるビッグブラザーではなく、民間企業による同様の取組については、政府が介入するには様々な課題があるというところで、なかなか難しい問題が生じているということかと思います。そうすると、これは以前、AIに関する原則を活用するに当たって、この検討会における例えばファクトチェックの問題は千載一遇のチャンスであると申し上げたわけでありますけれども、今後AIを用いた、こういった自動検知機能とかファクトチェックの活用、これをどのように機能させるかということについて、第三者による検証の仕組み、これを政府が行うのか、それとも民間が、いわゆる第三者認証的に行うのかということを検討するということが重要になっているかと思いますけれども、そうすると、今日のまさに表現の自由の保障との関係における問題も、原則を機能させる仕組みを今後検討するに当たって、原則を複合的にどのような視点から見ていくのか、それをどのように規律を生かしていくのかということを考える上で大変勉強になりました。

以上、私からの意見であります。

【宍戸座長】　ありがとうございました。

それでは次に、手塚構成員、お願いいたします。

【手塚構成員】　手塚でございます。非常に重要な内容のところをおまとめいただきまして、ありがとうございます。

非常に国と、こういうDPFとの関係性、これが明確に分かってきたかなという気もしています。その中で、今ちょうどこのページです。ここすごく、最後のところ重要だなと思

っていまして、政府自身も透明性を確保しなければならないという、こういう表記があるわけですけど、この透明性というのは、先ほどからの説明ではDPF側の透明性というお話でやっていましたけども、ここの政府側というのは、先生のお考えではどういうことを透明性ということで考えていくべきなのかというところもお教えいただければと思います。

以上です。

【宍戸座長】　ありがとうございます。

次に、森構成員、お願いいたします。

【森構成員】　ありがとうございます。大変勉強になりました。私は個人的に非常に関心を持っていた分野ですし、お教えいただいたことが大変勉強になりまして、ありがとうございました。

そしてまた、最後のページでお示しいただいた政府がプラットフォームによるモデレーションの基準自体について介入するというのではなくて透明性について介入する、それがプロセスの正当性を確保するものであるということは、恐らくこのプラットフォームの研究会の、これまで公表してきた考え方にも非常にマッチしているものであって間違ってないというふうに言っていただいたのではないかというふうに思いまして、心強く思いました。

私がお尋ねしたいのは、そのプラットフォームのコンテンツ・モデレーションに関するアプローチです。そこでシステミックなアプローチということと権利ベースのアプローチですね、そういうことを13ページで御説明いただいたと思いますけれども、確かにシステミックなアプローチとしてプロセスの正当性を保障するということが求められていると思いますけれども、もう権利ベースのアプローチというのはできなくなったんだということでいいのかということをちょっと教えていただきたいと思いまして、例えば本件ではまずモデレーションが行われて、それに対して消された人がアピールするという考え方かと思いまして、このプラットフォーム研究会もそこを重視しているわけなんですけども、表現の自由との関係でそこを重視しているわけなんですけども、権利侵害を受けた人、あるいは受けたと思っている人のほうから消してくれという手続というものができないものかと思っておりまして、そういった部分についてしっかり消してほしい、被害者側の権利に配慮したプロセス、しっかり消すべきものは消すということが必要になる必要性があるのではないかと思いまして、そういったことについても教えていただければと思います。すいません、申し訳ありませんでした。

【宍戸座長】　ありがとうございます。それでは、山本先生、お願いします。

【山本構成員】　ありがとうございます。山本です。

私のほうから２点、１つが質問で１つはコメントのような形になるんですけれども、先般フェイスブック文書なるものがアメリカにおいてかなり報道されているという状況で、その中で言われていることは、フェイスブックが10代の女性のメンタルヘルスを害するような、そういうコンテンツを送っていたとか、憎悪とか対立を増幅するようなコンテンツを促進するアルゴリズムを使っていたといったような報道がなされているところかと思います。

そういう状況を踏まえたときに、先ほどおっしゃっていただいた最高裁、フェイスブックのいわゆる括弧つきの「最高裁」が果たすその役割や権限というものがどこまでなのか。つまり、コンテンツ・モデレーションのような、削除する、しないの問題だけを審理していくのか、それとも邪悪なアルゴリズムというんですか、そういうアルゴリズムの精神的影響ですとか、そういったようなことも、最高裁なるものがチェックしていくということになるのか。つまり、その最高裁の監督する領域とか範囲というのがどこまでのものとして想定されているのか、あるいは水谷先生がどこまでのものを監督すべきだというふうに思われているのかということを伺えればと思います。

２点目は、先ほどのシステミックなというところとも関連するんですけれども、なかなかあるコンテンツが直接誹謗中傷にあたり誰かの権利を侵害しているとか、１つのコンテンツが表現の自由として問題だとか、そういう１つのコンテンツに着目をして、それが白なのか黒なのかを判断していくのは難しい。その意味で、全体のシステミックな影響というものにより着目していくというのは、確かにあり得るだろうと思いました。

その点、最近私自身、インフォメーションヘルス、情報の健康ということを最近、東大の鳥海先生たちと一緒に言っているんですけれども、ある種、言論空間における情報的な健康、要するに情報の偏食ではなく多様な情報をバランスよく摂取して、偽情報などへの免疫を獲得させていくということも重要なのかなという気がしてきております。それは権利侵害とか、具体的な何か損害、侵害の救済ということではきっとないないだろうというふうに思うんです。

政府も、例えば21条の観点だけではなくて、25条の「健康で文化的な最低限度の生活」というものを守っていくということになるのではないかと。言論空間の「衛生」は、まさに健康で文化的なというところにも関わってくる問題なのかなというふうに考え始めまし

た。

そういう意味では、政府が情報の健康に直接関わるということになると、これは全体主義的にもなるということで抑制的であるべきだと思いますけれども、やはりプラットフォームがユーザーの情報的健康について配慮していく、政府もそういった取組を奨励していくというような２段構えの構造というのはあり得るのかなというふうに、システミックな観点と関連して思っていると。この点、もし御意見があれば伺いたいなというふうに思いました。

すいません、以上です。

【宍戸座長】　ありがとうございます。

時間も限られておりますので、可能な範囲で、水谷先生のほうから御質問への回答、あるいはコメントがあればいただければと思います。よろしくお願いします。

【水谷氏】　たくさんのコメントとご質問、本当にありがとうございます。

時間も限られておりますので、お応えできる範囲でお話させていただければと思います。まず最初に、政府側の透明性というのはどういう部分を高めるのかというお話が出ましたので、これについてお答えをさせていただければと思います。

サンタクララ原則の基本原則と運用原則の後に、政府、あるいは国家アクターのための原則というのが新しくついていまして、２つぐらいあるんです。１つは、これはどちらかというと政府自身の透明性を高めるというよりは、政府が、企業が透明性を高めようとしていることを妨害してはならないとか、妨害するような障害があればそれを取り除いたほうがいいというような原則なんです。恐らく２点目のほうが政府の透明性の観点に関わってくると思うんですが、要するにここで言う政府の透明性というのは、コンテンツの修正なんかをプラットフォーム側に、直接的に法律に基づいたり、あるいは間接的に日本とかのような民主主義国家じゃない国の場合は、そういう圧力的なものはありますから、そういった形で修正を求めていくというようなことをやっているのであれば、どういう関わり方をしているか、つまり、プラットフォーム側のコンテンツ処置とかアカウント停止に対して、国家側がそれをどれぐらい要求しているかみたいなところです。あるいは、要請に法的根拠があるのかどうかとか、こういったことをオープンにしていきなさいというような点があると考えられます。

僕もサンタクララ原則を見て、なるほどと、この部分も必要だなというふうに思った次第でして、まだこの部分は、そんなにたくさん原則にも書かれているわけではないので、

ほかの原則なんかを見て、こういう部分について、もうちょっと私自身の研究も深められればなというふうには思っております。

それから、権利ベースとシステミックなアプローチの違いの話が出て、私の説明の仕方がよろしくなかったかなというふうにもちょっと思ったんですけども、別に権利ベースの仕組みがこれから無くなっていくわけでは恐らくないと思います。現状、誹謗中傷対策も踏まえて、プロ責法の発信者情報開示の部分が大きく改正をされたわけですけども、権利侵害を受けた被害者が、それに対して発信者情報開示をして、そこで得た情報を基に、従来どおり司法の場において救済を受けるというような仕組みというのは、当然これは残ると思うし、残さなければいけないと。なぜならば、そこで我々の表現の自由とほかの権利、名誉権とか、そういったもののバランス調整の法理形成を判例によってこれからもしていかなきゃいけないので、その部分は確保しなきゃいけないと思います。一方で、恐らくそれで救済できるものというのが、実は全体から見てやっぱり一部だろうということで、こういうシステミックなアプローチみたいなことが出てきていると思います。プラットフォーム上で毎日、非常にたくさんの数が、誹謗中傷その他で消されたりしているので、プラットフォームのコントロールを受ける側にとっても、どちらのアプローチが実はプラスなのかと。発信者情報開示の仕組みがある種簡易化したことはありますけども、それでもやっぱり司法スキームで解決していくというのはなかなかハードルが高いわけです。その一方で、それをプラットフォームが効率的に消していけるのであれば、それを高めていくというのも一つのあり得る方向性なのかなとおもいます。

そういう意味で、プラットフォーム側には違反通報みたいな仕組みはちゃんとあるわけですから、これをもうちょっと活性化するというか、うまく使えるような仕組みというのを考えていく必要があるかなと考えております。繰り返しますが、決して権利ベースの仕組みが今後もいらなくなるというわけではないと考えております。

それから、フェイスブック文章とオーバーサイトボードの役割の話なんですが、オーバーサイトボードが現在の仕組み上、主に担っているのは、明らかにコンテンツ・モデレーションで消された投稿についてです。これを復活させるかどうかという点に主眼を置いた、実はあまり広くない部分を役割として担っているということです。

ただ山本先生がおっしゃられたとおり、フェイスブック文章の中でそういうアルゴリズムも含めて権利侵害というより、どちらかというと我々の社会や人間にとっての有害さを持つようなものというのが問題視されるということがあると思います。そういうある種の

広い意味でのユーザー体験、このアルゴリズムによって形成されるユーザー体験を、これからオーバーサイトボードが担っていくべきかどうかはちょっとまだ分かりませんけども、しかし同様の外部審査機関みたいなところがチェックしていくという仕組みは、これは当然あってもいいんじゃないかなというふうに考えているところでございます。

ちょっと時間もありませんので、全てにお答えできなかったんですけども、私からはここまでとさせていただきたいと思います。ありがとうございました。

【宍戸座長】　水谷先生、的確に分かりやすくお答えいただきありがとうございました。

それでは、申し訳ありませんが、時間の都合上先に進ませていただきます。

資料３につきまして、国立情報学研究所、越前教授から御発表いただきたいと思います。よろしくお願いいたします。

【越前氏】　越前でございます。

それでは、始めさせていただきます。

国立情報学研究所の越前でございます。本日は、顔を対象としたフェイクメディアの生成と検出ということで、私からはフェイクメディアの生成と検出に関する技術的な内容について御説明させていただきます。どうぞよろしくお願いいたします。

時間もありませんので、アウトラインを、ざっと御説明いたしますと、今日は４点御説明させていただきます。生成と検出と、最終的なインフォデミック等の克服に向けてのプロジェクトについても御説明いたします。

まずは、イントロダクションということで、実際のフェイクメディアの例を御紹介いたします。

簡単なクイズ的なものですが、２つの写真をまず示しております。実はそのどちらか一方はAIが成した架空の顔画像になります。どちらがフェイクか分かるでしょうか。これは、左のほうがフェイクで、これは敵対的生成ネットワーク（GAN）と呼ばれる機械学習モデルを使って生成した架空の物の顔だということです。このように、本物と見紛うような顔画像をAIが生成することが可能になっております。

こちらはいかがでしょうか。こちらは、右がフェイクになります。

これは、一昨年グーグルが公開したリアルとフェイクの映像のペアになります。実は右側が、違う人の顔に置き換えるFace swapと呼ばれる機械学習モデルによるものです。Deepfakeという名称でも知られております。近年は、静止画だけではなく映像でも、AIが自然な形でフェイクメディアを生成することが可能になってきております。

それでは、全体の背景でございますが、御存じのとおり、AIの技術進化と計算機資源の充実によりまして、顔、音声、身体、自然言語などの人間由来の情報をAIが学習し、本物と見紛うフェイクメディアの生成が可能になりつつあります。ここに挙げていますように、フェイク顔画像・映像であれば2018年頃からDeepfakeという名の下に知られるようになり、フェイクニュースの生成だとGROVERというモデルが一部紙面等をにぎわしたかと思います。

実際の事例も起きておりまして、例えば、このウォールストリート・ジャーナルの記事にありますように、フェイク音声で実際に企業の幹部になり済まして現金を搾取した事例や、こちらは、先ほどのGANと呼ばれる機械学習モデルを使って生成した架空の人物の顔なのですが、この顔画像を使って偽のSNSのプロファイルを作成して、企業の株価操作を目論んだ事例などがあります。こちらは、イーロン・マスクになりすましたフェイク顔映像でZoomに参加して、他の参加者をびっくりさせている映像ですが、こういったことが技術的にできるというデモでございます。

御存じのように、国内では2020年、Deepfakeによるアダルトビデオの公開で逮捕されたという事例がございます。

このように、フェイクメディアを用いることで、幾つかの深刻な脅威が起きているということでございます。

では、どうやって顔を対象としたフェイクメディアというのはつくられているのか、技術的な内容を簡単に説明したいと思います。

大きく分けて５つのタイプがございます。これを順次、最新の技術動向に基づいて御説明させていただきます。

まず、１番目の顔の全体の合成でございます。これは、ある潜在変数というノイズから、実世界に存在しない顔画像を機械モデル、学習モデルで生成するものでございます。最初にお見せした２枚の静止画像は、まさにこの顔全体の合成というものでございまして、代表的な手法でよく知られているのは、この左にお示ししているStyleGANと呼ばれるものです。これは、高品質な顔映像の生成に特化した敵対的生成ネットワークというモデルでございます。

２番目は、顔の属性操作と呼ばれるものでございまして、これは、入力は顔画像なのですけれども、その顔画像の髪の色だとか肌の色とか表情などを違和感なく修正した顔画像を生成するものでございます。右の画像にありますように、左の列にインプットと書いてありますが、これが入力となる顔画像でございまして、これをブロンドヘアに変えるとか

ジェンダーを変えるとかというのを、違和感なく修正することができてしまうというものでございます。

それで、次は３番目でございます。顔映像の表情操作というものでございます。これは、攻撃者の表情に合わせて、ターゲットの顔映像を生成する手法でございます。代表的な手法は、2016年にコンピュータービジョンのトップ会議で発表されたFace2Faceと呼ばれるものであります。これは左側に示してございますが、左上が攻撃者の表情の映像になります。左下のターゲットの顔映像に対して、攻撃者の表情を転写して、攻撃者の表情に合わせたターゲットの顔映像を生成することが可能です。Face2Faceでは、攻撃者側とターゲット側の双方で映像が必要なのですが、右側に示しているNeural Talking Head Modelというのは、ここでいうとモナリザの１枚の顔画像から攻撃者の表情を転写したモナリザの映像を合成可能です。これは、先ほどお見せしたイーロン・マスクの合成と同じような手法になります。

続きまして、４番目ですけども、顔映像の話し方操作というものでございます。これは、音声やテキストを合成して、その音声やテキストを話すターゲットの顔映像を生成する方法になります。有名なものとしてSynthesizing Obamaという手法がございまして、これはオバマさんの任意の合成音声を使って、その音声をあたかも話しているようなオバマさんの顔映像を生成しようというものでございます。要するに、著名人に好きなことをしゃべらせることを音声を入力としてできてしまうということでございます。

右のモデルは、話者の音声、しゃべっている映像からテキストを抜き出して、そのテキストの一部を改変すると、それに合わせて話者の映像の口元が自然な形で修正され、話者が話している内容の一部を違和感なく改ざんした顔映像ができてしまうというモデルでございます。

最後の５番目、これはよく知られたDeepfakeに関わる手法でございますが、顔の入替え、これは特にFace swapと呼ばれております。これは、映像の顔部分をほか人物の顔に置き換えるという手法でございます。左が初期のFace swapのモデルでございまして、これは、AさんとBさん用の２つのオートエンコーダーと呼ばれる機械学習モデルを用います。オートエンコーダーというのは、簡単に御説明しますと、このAさんの顔画像をエンコーダーによってぎゅっと圧縮しまして、表情などの特徴を一旦抽出した後、その特徴からAさんの顔画像を、今度はAさん用のデコーダーを使って再構成する仕組みです。Face swapでは同じようにBさん用のエンコーダーとデコーダーも用意します。顔の交換には、このAさん

の特徴を、Aさん用のデコーダーではなくBさん用のデコーダーを使って再構成することで、Aさんの表情を維持しながらBさんの顔に置き換えるということになります。これがAさんからBさんへの顔の入替え、Face swapになります。この顔交換方法は、ポルノビデオ等の生成に利用されたことで、社会的に非常に問題になりました。その後、メディアで、Deepfakeという言葉でこの手法が説明されています。右は、改良版のFaceswapーGANというものでございますが、これは先ほどオートエンコーダーを２つ使っておりますが、２つから１つに減らすことで学習時のコストを抑えております。

次に、顔を対象としたフェイクメディアの検出（真贋判定）手法について、最新の動向も含めて御説明いたします。

実は、このDeepfake等のフェイクメディアの検知ですが、私たちのグループが世界で初めて、2018年にFace swap、いわゆる顔を入替えしたフェイク顔映像と、Face2Face、顔の表情操作をしたフェイク顔映像を検出する手法を開発しました。私らの手法は、AIを用いて、Face swapやFace2Faceによって生成されたフェイク顔映像を自動的に検出することができるということです。採用したのは、３万個のパラメーターを持つ非常にシンプルな４層のCNN（畳み込みニューラルネットワーク）です。この左にお示ししているものでございますが、４層の非常に単純なネットワークでございます。このモデルの中間層を見てみますと、真贋判定にどのような特徴、実は幾何学手法モデルはブラックボックスなので何をやっているかは分からないのですが、中間層等を見てみると、目の辺りとか口元の特徴を見て真贋を判定しているように見えます。検出率は、当時、Face2FaceとFace swap、それほど多くの生成手法がなかったというのもありますが、２つの手法で生成されたフェイク顔映像で90％以上の検知率という高い値を示しております。

さらに私たちは、より高い精度でフェイク顔映像の検出を行う手法というのを提案いたしました。この手法では、CNNではなく、カプセルネットワークという機械学習モデルを採用しています。カプセルネットワークというのは、簡単に申し上げますと、顔のパーツ間の幾何学的な関係性を保持しまして、僅かなゆがみを検出することできるため、フェイク顔映像の検出をさらに高精度に実現することができるということです。

少し模式図を使って御説明しますと、ここに２枚の顔面を表した模式図があります。左の顔は通常の顔ですが、右の顔は目、鼻、口といった顔のパーツが顔面上にランダムに配置されております。実はCNNでは、この２枚の画像の違いを見分けるのは難しいのですが、カプセルネットワークでは、この違いを容易に見分けることができます。技術的な説明に

なりますが、カプセルネットワークは、カプセルという名が示しますように、複数のカプセルから構成されており、各カプセルでは、入力画像を様々な観点から分析します。右図に示したものは３つのカプセルにおける中間層の結果なのですが、入力した顔映像が同じでも、カプセル毎に異なる顔領域がアクティブになっていることが分かります。それぞれのカプセルが真贋判定を行い、上位のカプセルで総合的にリアルかフェイクか判断するというモデルになります。そうすることで、フェイク顔映像に存在する微小な人工ノイズや位置関係のゆがみ等を検出することが可能になりまして、高精度な真贋判定が可能になります。

これは、Face swapの検出結果を示しております。私の顔が他の人に入れ替わった動画が上でございまして、下の動画がその逆になっております。どちらの動画も映像も高い精度で検出できていることが分かります。

これは、Face2Faceの検出結果を示しております。顔映像の表情操作になりますが、一定の強度で映像圧縮がかかっても80%以上の精度で検出できていることが分かります。

さらに、私たちはこれも世界で初めてなのですが、フェイク顔映像の真贋判定と同時に、顔の改ざん領域を推定する手法を提案いたしました。下の図は、様々なフェイク顔映像の生成手法によって生成された顔になりますが、手法によって改ざん領域が異なっていることが分かります。私たちの手法によって、真贋判定と同時に改ざん領域を推定することで、どの手法によって改ざんされたのだろうかということをある程度知ることが可能になります。

これが実際のデモ映像になります。真贋判定と同時に改ざん領域を推定しておりまして、この場合は、Face2Faceのsmooth maskという手法で生成されたのではないかということが推測できるわけです。

それでは、私たちが現在進めております、JST CRESTのフェイクメディアのプロジェクト及び国立情報学研究所に新たに設置したシンセティックメディア国際研究センターの取組について少し御紹介をさせていただきます。

まず、JSTのCREST FakeMediaと呼ばれるプロジェクトでございますが、昨年12月からJST（国立研究開発法人　科学技術振興機構）の戦略事業であるCRESTの下で進めております。CREST FakeMediaですが、フェイクメディアとインフォデミックの関係が研究プロジェクトのモチベーションとなっております。先ほど御説明いたしましたが、AIの技術進化と計算資源の充実によりまして、顔、音声、身体、自然言語などの人間由来の情報をAIが学習し、

本物と見紛うフェイクメディア、以降はFMとも呼びますが、これらの生成が脅威になりつつあります。特に、新型コロナウイルスの感染拡大でフェイクメディアによるインフォデミックが世界的に問題になっているという状況でございます。

インフォデミックというのは、SNSなどを通じて社会に恐怖や混乱を引き起こす、不確かな情報の氾濫という意味でございまして、今回のコロナ禍では、御存じのように、科学的根拠のない予防法に関わるフェイクニュースや、右の写真に示しておりますように、ソーシャルディスタンスを保っているのですが、望遠レンズを用いて特定の方向から撮影することで、意図的に密集状態を演出したと思われる写真がSNS上で拡散され、社会に混乱が生じているところでございます。

今後、攻撃者が、AIにより多様なFM、フェイクメディアというのを生成しまして、インフォデミックを意図的に発生させ、不当な利益を得る可能性があると私たちは考えております。

多用なFMですが、具体的には３つのタイプを想定しております。１つ目は、先ほどから御紹介しています、本物に限りなく近いが本物ではない、フェイクメディア、これをメディアクローン型FMと定義しております。２つ目は、世論の操作のためにメディアを意図的に加工する。これは、加工前のメディアがリアルかフェイクかは問わないのですが、このようなものをプロパガンダ型FMと定義しております。３つ目が、人間をだますのではなく、AIに対して誤動作・誤判定をさせる、敵対的サンプル型FM。これらのFMを使ってフェイク映像、フェイク音声、フェイク文書等を作成し、サイバー社会に大混乱を与える可能性があると、私たちは考えております。

そこで、私たちは、人間中心の健全なサイバー社会を実現するためには、このような脅威に適切に対処することで情報の信頼性を向上させ、同時に、多様なコミュニケーションや意思決定支援が不可欠であると考えております。

検討している課題は、以下の３つとなります。

まず１つ目が、高度なFM検出技術。Real／Fakeだけではなく、FMの種別など、説明可能な形式で情報提供する技術の確立を目指します。また、このFM検出の前提となる多様なFM生成技術についても検討いたします。

２つ目が、FMの無毒化技術、これは、思考誘導や誤動作・誤判定が生じないように、フェイクメディアを無毒化するという全く新しい考え方でございます。この無毒化によってSNS上で通常のメディアとしての視聴に加えて、AIによる学習が可能になります。学習デ

ータに攻撃者によって人工的なノイズが付加されている可能性がありますので、学習前にできるだけ品質を維持しながら、そういったノイズを除いてあげるというのが無毒化のもう１つの目的です。

３つ目が、意思決定支援技術になります。これは、今述べたFM生成検出無毒化技術を最大限に活用して、情報の信頼性を高める、社会システムの原理と技術を確立するものでございます。右の映像に示しておりますが、主たる共同研究者の東工大の笹原和俊 准教授が、SNSを模した意見形成モデルによるエコーチェンバーの発生といったことに関して高い専門性を持っておられまして、こういった計算社会科学の知見も融合しながら、このプロジェクトを進めております。

実際に、我々の取り組んできたものをデモ映像にて御説明いたします。

これは、先ほど御説明したリアルとフェイクの真贋判定と、真贋判定と同時に改ざん領域を推定する技術でございます。このプロジェクトでは、これら技術をさらに進めまして、音声なども含め、どの手法で改ざんされたのかというのを説明可能な形式で情報提供したいと考えております。また、今後のメディア流通やデータ学習の際に必要となる無毒化、この映像は究極ですけども、フェイクからリアルを復元するというようなものについても検討していきたいと考えております。

さらに、２０２１年３月にCREST FakeMediaのウェブサイトを公開し、プレプリント、プログラム、データセットなどの研究成果の公開を積極的に行っております。

それで、Deepfake等のメディアクローン（MC）型のFMの脅威というのが、さきほども少し御紹介しましたが、２０２１年に入ってからも、国内を含め様々な事案が生じております。特にこのMC型のフェイクメディアは、スマホアプリなどを使って誰もが生成できるようになってきているということです。学習済みの機械学習モデルが公開されておりますので、それらを使って、スマホ等のアプリというのはそれほど難しくなく開発できてしまうということで、フェイクメディアを生成する敷居が非常に低くなってきていることを私たちは問題視しております。

それによって、今までは専門的知識を持った組織や人物による詐欺、詐称、思考誘導が一般的だったのですが、一般の個人による思考誘導や小さなコミュニティー内での名誉毀損やいじめ等に用いられるようになっているということを懸念しております。例えばこの左下にあるように、娘のライバルを蹴落とすために母親がライバルのDeepfakeを生成したとか、親の目が行き届かないコミュニティーでのいじめの発生とか、そういったことも大

きな問題ではないかということを、私たちは認識しております。

そこで、このMC型のFM検出の早期の社会実装を検討することにいたしました。私たちは、このMC型FMの中で、フェイク顔映像を対象として、多くの企業が簡単に利活用することが可能なフェイク顔映像検出プログラムを開発いたしました。私たちのこれまでの企業関係者とのやり取りから、Deepfakeの自動検出に非常に興味を持っている企業が多いのですが、自前で検出モデルを開発するためには高度な深層学習技術等が必要で、また相応規模のGPUサーバー等のリソース投資が必須となるので、参入障壁が高いという問題がございました。

そこで、この学習済みモデルと前処理・後処理を包含することで、すぐにでも利用可能なフェイク顔映像検出のサーバープログラムとAPIを開発いたしました。企業は、サーバーを用意することで、このプログラムを用いて、単独でAI as a service (AIaaS)として利用可能ですし、既存のサービスと組み合わせることも可能です。

この開発したプログラムをSYNTHETIQ：Synthetic video detectorと名付けまして、これは判定対象の映像のクライアントからサーバーへのアップロード、サーバーによる真贋判定、そこから判定結果を示した映像をサーバーからクライアントへダウンロードするまでの全てのプロセスをWeb APIとして使用可能ということです。さらに、これはWeb APIですので、AIを利用したウェブサービスを容易に実現することができます。

さらに、２０２１年７月に発足しました国立情報学研究所（NII)、シンセティックメディア国際研究センターについて少し御説明いたします。

先ほどご説明したCREST FakeMediaプロジェクトの重要性が、私の所属機関であるNIIに認められまして、NII内に研究センター設置を認めていただきました。このセンターのミッションでございますが、ここにありますように、人間中心のAI社会を実現するために、多様なメディアの生成、メディアの信頼性確保、意思決定支援のための研究開発を、実世界の課題を取り上げながら、国際的な拠点として推進することになっております。本プロジェクトであるCREST FakeMediaに加えまして、音声メディアを主な対象としてその利活用と保護を実現する基盤技術の確立を目指しましたNIIの山岸先生によるCRESTが2018年に発足しております。これら２つのCRESTが相補的に連携することで、多様なメディアの連携対象としたシンセティックメディア生成、フェイクメディアの検知、メディアの信頼性確保のための研究開発を、国内だけではなくて海外の研究機関とも密に連携することで、多用なメディアの利活用と信頼性確保を、国際的体制で追求していきたいと考えております。

これがセンターのウェブサイトでございます。

最後に、今まで御説明してきましたフェイクメディア検出の課題について述べたいと思います。

最初に画像や映像などのフェイクメディアというのは、SNS共有や再配信のタイミングでリサイズや再圧縮などが施されますので、補間や符号化によるノイズによってフェイクメディアの生成時と比べて品質が劣化しているということです。これによって、真贋判定の精度が低下するという問題がございます。

２点目が、実はもっと本質的な問題なのですが、現在、非常に多くのフェイクメディアの生成手法が出現しているということです。御存じのように、機械学習モデルというのは、学習モデルという名があるように、未知の手法で生成されたフェイクメディア、すなわちモデルが学習していないフェイクメディアに対しては、高い精度による検出が苦手なのです。そうしますと、定期的な学習データのアップデートやそのモデル学習が必要になってきます。

さらに問題になるのは、今まで覚えていた手法で生成されたフェイクメディアを高精度で真贋判定しながら、新たな手法で生成されたフェイクメディアに対しても、高精度の真贋判定を実現しようとすると、これはかなり難しくて、それ相当の知見がいりますし、モデル学習のコストも非常に膨大になってきます。

そこで、今、私たちが検討しているのが、画像を対象とした自動ファクトチェックです。自然言語分野ですと、自動ファクトチェックというのは既に検討や実装が行われているのですが、画像を対象としたらどのように実現できるのかということを今考えております。これはどういうことかというと、クエリーの画像から類似のオリジナル画像を検索し、オリジナル画像とクエリー画像の比較からオリジナル画像の真贋判定、改ざん箇所を推定するというものでございます。今後は、フェイクメディアの生成手法は多岐にわたるということでございますので、将来的には機械学習モデルによる真贋判定とファクトチェックを相補的に活用することが重要ではないかと考えております。ただ、このファクトチェックも、ファクトとなる画像と映像のデータセットをどうやって維持するかというのが課題になってきますので、それらの課題についても今後取り組んでいきまして、これらの相補的な活用ということも含めて検討していきたいと思います。

発表は以上でございます。御清聴ありがとうございました。

【宍戸座長】　越前先生、貴重な御発表ありがとうございました。

それでは、ただいまの御発表につきまして、構成員の皆様方から御質問、御意見があれば承りたいと思います。また、チャットで私にお知らせいただきたいと思いますが、いかがでございましょうか。

それでは、手塚先生、お願いいたします。

【手塚構成員】　越前先生、大変貴重な内容を発表いただきまして、ありがとうございます。

テクノロジーサイドの話と、これはもう１つ、ユーザー、利用する方たちの経験値というかUI／UX的な、それらの接点のところの大きな問題だと思いますが、あともう１つ、データといいますか、そういう視点から見ると、データの動画とかそういう画像もそうですし、静止画の画像もそうですし、みんなそういうコンテンツのデータまたインフォメーションのデータという視点でこういうものをどういうふうに捉えるかといったところで、一般には、こうやって画像になると、人間のほうは視認性でそれでチェックするわけですけども、システムの仕掛けとして、こういうところのデータに対するバリデーテェッドなデータ、または検証ができるようなデータ構造にこういう世界をしていかないと、なかなかフェイクかどうかというところの、非常に深遠なテクノロジーだけでカバーしていくというのには限界が出てくるんじゃないのかなと。もうほとんど同じにできてしまえば、もう全く見分けつかなくなっちゃうわけですよね。そうすると、そのデータに対しても見分けをつけるようにするとなると、究極はやっぱり、単なるローデータではなくて、そこにバリデーテェッドな情報を付加するという、そういうデータという形で流通させていく世界を考えるというのは、もう１つ一方で重要かなとも思っていて、そういうところはどうなんでしょう。先生からの視点ではいかがなんでしょうか、そういうところのものの見方。

【越前氏】　御指摘の点、大変ありがとうございます。その点は、あとどの程度そういったデータ生成の際にコストをかけるかということが問題になるのではないかと思います。最近はNFTとかございますが、再配信の際の符号化ノイズの重畳などを考慮すると、電子透かしによって権利情報等をメディアに不可分に重畳するとか、そういったことでメディアの真正性を検証していこうという流れは、御存じのように過去からもあるのですが、今を含めてこれからは、画像や映像だけを見ても、どれがリアルかフェイクかも分からない状態になってくるということがございます。

ですので、そういったプラットフォームというのは構築する必要性は高まってきているのだと思っているのですが、一方でプラットフォームを運用する企業とプラットフォーム

を利用するユーザーから見てのコスト感というのは多分非常に大きくになってきますし、それらを踏まえまして、プラットフォーマーもユーザーも納得がいくコスト感のところに落ち着くんじゃないかと思っております。

ただ、フェイクメディアの脅威というのは、まだまだこれから多く出てくると思いますので、そういった事情も踏まえながら、どの技術について注力していくのかということは、技術サイドとしては常にウオッチしながら検討していきたいと思っております。

ありがとうございます。

【手塚構成員】　それで、そういう点から、あと１点、先ほどアライアンスの話があって、検出率の話があったんですが、生態認証系だと検出率何％とかそういうパーセンテージで結構議論があるんですが、これについては、どのくらいの検出を想定して、世の中的にもどのくらいであればリーズナブルなのかとか、その辺は今、先生の御見識ではどんな感じでしょうか。

【越前氏】　これは私たちが開発したフェイク顔映像検出プログラムに興味を持たれた複数のユーザー企業からも問合せいただいているのですが、まさにユーザー企業さんと、どの程度であれば実サービスとして運用できるのかということを詰めている状況でございます。

ただ、これは先ほど御説明しましたように、フェイクメディアの生成手法は本当に多様にありまして、現状、機械学習モデルで多様なデータセットを学習しなければいけないのですが、いわゆる研究者が用いることができるフェイクメディア検出用のデータセットというのは、やはり偏りがあります。今現在、多様なフェイクメディアの生成手法が出ている中で、それらをデータセットとして逐次収集して学習するというのはかなりコストがかかりまして、これから実サービスとして運用可能になるためには、やはり企業と協力してデータセットを集め、アップデートしていく、それによって適宜モデル学習をすることが重要と考えます。データセットにおける多様なフェイクメディアのブレンド率なども常に考える必要があり、例えば特定の用途に特化したアプリケーションに対して精度を高めたいのであれば、それに特化したデータセットを構築して学習し直すというようなことが必要になってきます。精度に対しては、具体的には、ちょっといろいろあるのですが、企業によって違うので、まだ検討中ということでございます。

【手塚構成員】　どうもありがとうございます。大変貴重な御発表をいただきまして、ありがとうございました。

【越前氏】　ありがとうございます。

【宍戸座長】　越前先生、ありがとうございました。時間の関係上、越前先生との質疑はここまでとさせていただき、後ほど構成員が先生の御発表を咀嚼して、メール等で事務局通じて御質問させていただくこともあろうかと思いますが、どうぞその節はよろしくお願いいたします。

【越前氏】　ありがとうございます。

【宍戸座長】　それでは、次に、資料４につきまして、株式会社スペクティ、村上様のほうから御発表をお願いいたします。

【村上氏】　よろしくお願いいたします。株式会社スペクティの村上です。今、資料を共有します。

よろしくお願いいたします。スペクティの村上です。前の２つの発表を見ていて、水谷先生も越前先生も非常に興味深いなと思いまして、我々、SNSの情報を日々解析しながら、基本的に防災に関わる情報をいかに自治体であったりとか防災関係の方々に届けていくかというようなことを行っています。

その中で、やはりデマ情報とかフェイクとか、そういったものがいっぱい出てきますので、そういったものを扱っているというような、そういった会社になります。

ちょっと自己紹介になりますが、スペクティという会社は、2011年、ちょうど東日本大震災が発生したときにできた会社になります。今は、AIと防災を組み合わせるというところで、いろんなソリューションを展開しているというような会社になりまして、ちょうど創業したときというのが本当にサラリーマンをやっていて、そのときに災害が発生して、ボランティア行きながら、いろいろ現地の情報と実際にメディアなんかで伝わってくる情報の差というのは非常に大きいなというのを感じて、その差をどうやってなくせるかというところからサービス開発をしたというようなものになっています。

今、その関連もあって、ファクトチェックイニシアティブという団体の理事をやっていたり、あとAI防災協議会の理事をやっていたりします。

今お話ししたとおりで、我々の会社、本当に防災と危機管理、ここの分野に特化したスタートアップでして、AIと防災という組合せたソリューションを展開しています。

機器を可視化するというミッションを持って、様々な機器を見える化して、さらに予測をしていくというようなことを行っています。

その中で、SNSを解析をして、また災害が発生したときにその情報をいかに早く伝える

か、現場でどんなことが起きているかというのを伝えていくかというようなことを行っていまして、そのサービスをいろんなところに展開をしているというようなものになります。

我々、SNSの解析だけではなくて、気象データとか道路カメラ、河川カメラとか、あとは人工衛星のデータであったり、あと自動車のプローブデータと呼ばれるもの、走行データですね、そういったもの、あとは人流のデータだとか、そういったものから解析を行った上で、リスクの可視化と予測を行っているというようなものになります。

今日は本題に入る前に、ちょっと簡単にサービスの紹介だけさせていただきます。

こちら、今見えているのがスペクティのトップ画面になります。IDとパスワードがあると、ログインするとこんな感じで見れて、いろんなところから上がってくるSNSの情報というのを見れるというようなものになっています。どこで何が起きているかというのが瞬時に分かるような形で、地図と連動しながら情報を見るような感じになっています。これは今岩手県で雪が降っていますよというようなものです。見たい情報をいろいろ絞り込みをして、エリアを選んでいただいたりとか、あとはカテゴリーを選んでいただく。どんなことが発生しているか、気象災害であったりとか火災であったりとか、そういったようなものを選んで、そこで情報収集をするというようなものになっています。全体はこんな感じで、地図と連動して見せるような、そんな仕組みになっているものです。

ちょっと時間もないので飛ばしちゃいます。

裏側のAIの技術というのをどうなっているかといいますと、SNSに上がってきた映像、画像がついていた場合は、その中に何が映っているのかというのを自動的に判別をしていくというようなことを行っています。今、これ、煙が映っていますと、炎が映っていると。そうすると、これは火事の映像だなというのをAIが判断して、これは火事ですよというカテゴリーに振り分けるというような仕組みになっています。映像、画像だけではなくて、書かれたテキストの情報も、いろんなキーワードであったり全体の文脈であったりとか、そういったようなところから、これは何の情報ですというのをどんどんどん振り分けていくというようなことを行っています。

それで、発生から本当に瞬時に、現場の映像、画像、またはつぶやかれているテキストの情報と、地図と連動しながらどこで何があったかというのを見える化していくというようなものになっています。

国内だけではなくて海外の情報なんかも扱っていまして、海外で発生する災害であったり、海外だとテロとかいろんなことが起きますので、そういったようなものも、それを見

える化していくというような、そんなサービスになっています。

まとめますと、SNSの情報、SNSも、扱っているのはツイッターだけでなくてフェイスブックだとかインスタグラムとか、そういったような情報と、あと、SNSだけで情報判断しているわけではなくて、気象データであったりとか、あとは交通物流のデータであったりとか、そういったようなものを組合せて、それをアラートで配信をしたり地図に連動したりとか、あとは分析ができたりとか、そういったようなことができるような、そんな仕組みになっています。

今日特に話題になっているファクトチェックのところなんですけど、我々どんな対応しているかといいますと、SNSに上がってきた情報を、先ほど、この前の話で越前先生のやつもありましたが、基本的にAI側のほうで、いろんな学習を行った、過去にあったいろんなデマ情報を基に学習を行ったAIを基に、自然言語解析だったり画像解析を行っています。物すごく、先ほどあったDeepfakeのような形でかなり高度なものをチェックしているというよりは、よくあるSNSのデマ、特に災害なんか発生したときにあるデマはこういう画像が使われやすいとか、こういう文章を書かれやすいとか、そういったものをかなり分類化して、物すごくスピードが大事ですので、瞬時に判定をしていくというようなことを行っています。

例えば画像であれば、よくあるケースというのが画像の使い回しなんです。過去にあった、どこかで発生した災害の映像、画像を、それを貼りつけてSNSに上げて、あたかも今それが起きているような形で上げてくるというような、そんなことがよくあるんですが、そういったものを、同じ映像、画像というのがSNSの過去のデータの中にあるかないかとか、また、インターネット上にそういうものがないかとか、そういったものを判定して、どうも類似なものがあるよというものがあれば、それを判断をしていくと。ただ、全てAIで100％チェックができているわけではなくて、必ずAIで解析した後に、人の専門チームがついていまして、そこで分析を行った上で、お客様のほうに届くときは、必ず我々のほうでファクトチェックをしたものを配信するというようなことを行っています。

なので、AIと人でダブルチェックをしているというようなもので、ヤフーのコメントのところの分析なんかもそうだと思うんですけど、AIで解析したやつと、あと人のチェックも入れているというのと非常に似たような形ですが、我々もAIで解析した後に、やはり人の確認を行った上で、どうしても対価をいただいているサービスですので、我々のほうで正確性を担保してあげてお客さんに届けるというようなことを行っています。

今、実際にいろんな自治体さん、御利用いただいていまして、都道府県庁の防災ですと大体７割以上のところが使っていただいていたりとか、あと市区町村とか、あと官公庁さんも結構使っていただいているような、そんなサービスになっています。あとは、警察、消防なんかも非常に多くなっているような、そんなサービスになっています。あと民間の企業さんなんかも、インフラ系の会社さんであったりとか、あとは建設会社さんとか、あとは小売りチェーンさんみたいな、店舗をたくさん持っているお客さんなんかは、店舗が被災したりとか、そういったようなことがありますので、そういった災害対応で使っていただいているというようなものになります。

おかげさまでいろんな評価をいただいているというところで、ここからは今日の本題になりますが、SNSにおけるデマ情報、特に災害が発生したときというのは、災害が発生したからデマ情報が多いというわけでは実際にはなくて、どちらかというと、災害時とかそういった緊急時というのはデマが拡散しやすいというふうに考えたほうがいいかなと思うんです。デマだとかフェイクというのは、一定量必ず平時も常に上がっていたりはするんですけど、災害だとかコロナだとか、そういった緊急性が高いものがあると、やっぱりそういった人間の心理状態もあると思いますが、拡散しやすいというようなことです。これは我々がアンケートをとった結果なんですけど、SNSのいわゆるデマ情報によって困った経験があるかというと、大体半分ぐらいの人はそういう経験をしているというようなことを言われています。

特に、災害だとかそういったときには多いんですけど、いろんなタイプのデマがあって、それぞれいろんな形があるというところで、結構デマパターン分けみたいなことを我々やっていて、ちょっと名前も付けたりとかしているんですけど、ここで言うとオオカミ少年タイプとかヘイトとか勘違いとか伝言ゲームとか、いろいろあるんですが、ちょっと簡単に説明をすると、オオカミ少年タイプは本当に世の中を騒がせてやろうというある意図を持って、こんなことが起きたというので注目を浴びたいとか、そういった形でうそを並べていくと。有名なところでいうと、熊本の地震のときのライオンが逃げたというようなやつです。そういった投稿をすると、やっぱり注目を浴びますので、SNSとしては拡散しやすいと思うんです。

そのほか、ヘイトというのはある特定のグループだとか、そういったところにかなり否定的な思いを持っているような方々です。そのような形で、例えば外国人の方とか特定の団体に対して、ちょっとふだんからあんまりよく思ってない方なんかは、こういう災害み

たいなものに乗じてそういうデマを流したりとか、あとはもう単純な勘違いみたいなものがあったりします。

これは大阪北部地震のときですけど、大阪ドーム、京セラドームですね、京セラドームにひびが入ったというのがバーッと出てきて、ネット上にすごく拡散したんです。確かにぱっと見ひびっぽく見えるんですけど、これはよく見ると足場が組んであって、ひびでも何でもないです。ふだんからも大体こうなっているんですけど、ふだんはあんまり気にしてないんですけど、災害があったというときに、みんなぱっと見たときに、これはひびじゃないかというふうな、そんな感じで、ひびがあったみたいなもので拡散しちゃったりとか。あとは、コロナなんかでよくありましたけど、科学的にちょっと怪しいようなものなんかも結構拡散しやすかったりすると。

そのほか、伝言ゲーム的に広がっていくというのは結構あって、これは宮古島でコロナが発生したという2020年３月ぐらい、このときはまだ宮古島でコロナというのは全然出てなかったんです。宮古島は病院がすごく少なくて、コロナの病床というのはほとんどないような状態なので、もし広がったら大変だというようなことを言われていました。それが、３月の終わりぐらいに、急に宮古島でコロナが出たというううわさが広がったんです。その中で、どうも関西から来た人がコロナを持ち込んだとか、あとは大学生の旅行でとか、そういったものがどんどん変わってきて、最初に出てきた情報からどんどん情報が書き換わっていって、いろんな情報が拡散してったというようなことがあって、もともとは、実はコロナが出たのかなみたいなものが、一言つぶやかれていたんですけど、それが日がたつにつれて、もう宮古島でコロナが出たということが事実のような形で広がっていったというような感じだったんです。そういった伝言ゲームのような広がり方、こちらがありましたね。もともとそんなのがあったよみたいな話から、関西から来た人が持ち込んだとか大学生の旅行でとかいう話になって、そのほかまた大きな話になって、宮古島コロナ確定みたいなことが出てきたりとか、そういうような話のどんどん書き換わって広がっていくというような、伝言ゲームみたいな形というのは、よくあるタイプです。

あと、犯人探しみたいなものでも結構デマが流出しやすかったりすると。何か注目を浴びるような事件があったりすると、その犯人を特定しようというのがネットの中というのは結構起きやすいんです。これは有名なガラケー女の事件という、常磐道であおり運転した人というのを、その投稿を基に、その人が誰なのかというのみんなで犯人探しをして、全く違う人を犯人とつるし上げて、ネット上に拡散していったというようなものです。裁

判になって女性の方が勝ったんですけど、最終的にそういうような形になったというような、そういう犯人探しみたいなものも結構それによっていろんなデマが広がりやすかったりすると。

災害時によく起きるものでいうと、タイムリミットがあるとかいうようなものとか、関係者が言ったというようなものとかです。これは北海道胆振東部の地震ですけど、そのときは、６時間後に断水するというの自衛隊が言っていましたとか、そんなデマが広がっていくんです。そうすると、何時間後に何かが起きるみたいな情報というのは、ユーザーの心理からすると、これは早く誰かに伝えなきゃいけないというのをすごくせかされている情報共有になりますので、より拡散力が強くて、どんどん広がりやすかったりすると。あとは、何か関係者が言ったというような、ちょっとお墨つきを与えるような、自衛隊が言っていましたとか、そういうようなものがやっぱり広がりやすかったりすると。より信憑性が上がってくるというところで、そういったデマなんかも広がりやすいような環境にあるかなと。

これ、時限的なものというのは、人の心理に、行動に大きく影響するので、何か起きると、何時間後に水がなくなりますとか言われると、みんなそこで水を買いに走るとか、そういったようなことになりますので、特に災害時は結構そういうものが多く出てきやすいというようなところで、我々のほうは結構そういうのをチェックしているというような感じです。

続きまして、拡散のメカニズムというところです。先ほどいろんなタイプがあるというところ、特に災害時に起こりやすいタイプみたいなものに注目をして、先ほどあった越前先生のようなDeepfakeをどう見抜くかというような、そんな高度なところでは実はなくて、もっともっとSNSの情報は結構単純な形で情報が広がったりとか、単純なものがいろんなデマとして上がってくるので、そういったものに対して、こういう傾向があるものは大体デマの可能性が高いというのをある程度踏んでやっていくというのは、我々のやり方です。

そういった意味で、タイプだったり、拡散するのはどういうときだとか、そういうのを見ていたりします。その中で、拡散のメカニズムと書いていますが、大体５つのところが急激に拡散が広がるというようなものになったりします。インパクト投稿というのがありますけど、いわゆるインパクトがある画像、映像、そういうものというのは非常に広がりやすいですね。ライオンが逃げたみたいなもの。こちらは2019年の台風19号、関東を通過していったやつですけど、それが来たときに、小笠原諸島でこんなになっているというよ

うな写真がついて投稿されていたんです。確かに、この画像を見ると非常にやばいみたいな感じになりますし、こういう画像を作ると必ず広がってくるんです。実際にはアメリカの画像みたいなんですけど、そういったものとか、何かその情報自体インパクトがすごく高いような、そういうような感じになってくると。

なので、こういった何か強いインパクトがあるもの、インパクトがある情報なんかが出ると、すぐに拡散してくるというところで、そういったインパクトの高いものに関してはデマの可能性というのを見ているというようなことをしています。

あとは、広がる方法としては、このインフルエンサー型の拡散というのも結構あって、コロナ禍であったんですけど、ちょうど2020年の４月ぐらい、ゴールデンウイークでサザエさん一家が旅行に出かけるという内容があって、それが不謹慎だというのがSNSですごく広かったんです。ただ、実際にサザエさんが放送してからずっと、そんなに実は広がってなかったんです。あるところから急激に拡散していったというようなものなんです。それをたどってみると、実は有名人の方が、この内容は不謹慎なんじゃないかなというのをつぶやいたんです。それをきっかけに一気に広がっていったんです。こういったインフルエンサー型の拡散というのは、まさにSNSの特徴みたいなところだったりしますし、デマが広がっていくときとかも、やっぱり有名人の方とかSNS上フォロワーが非常に多い方が発信をすると、それが一気に拡散しやすくなるというようなものだったりします。有名人が発信すると、より本当なのかなと信じやすくなるんですけど、そういったものはちょっと疑ってみるとか、そういうのも大事なのかなと思ったりします。

あとは、メディアによって広がるケースなんかもあったりするので、これはデマとかとはまたちょっと違うんですけど、４月に入って「#東京脱出」という投稿が出始めた。これもコロナ禍です。2020年のちょうど４月です。緊急事態宣言が出る直前ぐらいとかですけど、「#東京脱出」というような投稿が出ていたと。ハッシュタグがついて「#東京脱出」というのがいろいろ出てきたと。それが、ある新聞社が「#東京脱出」というのを拡散しているよという記事を出したんです。それをよく見てみると、「#東京脱出」というのを、我々ツイッターのデータ検索をすると、４月１日から４月７日までの間は28件ぐらいしかなかったんです。４月７日の午前７時に、ある新聞社が「#東京脱出」という記事を出したんです。そうしたら一気に広がっていったというような感じで、この「#東京脱出」は本当は実は広がっていなかったのに、あるメディアが拡散しているというような記事を書いたから拡散したというような形で、メディアが記事を書くとそれが広がるというのはす

ごくあるので、やっぱりSNSの拡散みたいなところでいうと、メディア自体が拡散させちゃっているというのも結構あるのかなと。

あと、そのほかに意見が対立するようなつぶやきが出てくると、それも結構SNSの拡散につながるというようなものだったりします。何か対立するような意見、Aが正しい、Bが正しいみたいなものだと、そういったものというのは炎上しやすいので、双方の意見がいろいろぶつかって、どんどん広がっていくというようなものがあったりします。

トイレットペーパー騒動、皆さん覚えているとは思うんですが、コロナでトイレットペーパーがなくなるというのがあったんです。トイレットペーパーが不足していますよというツイートを見ていると、トイレットペーパーが不足する、しているというツイートが上がってきたのと、実際にトイレットペーパーがどれだけ本当に不足していたかというのは、最初の頃はそんなに相関してなかったんです。これはPOSデータから取っているトイレットペーパーの売上げ数ですけど、SNSのほうは、先にトイレペーパーがどうもなくなってきているというのがつぶやかれていると。実はこのトイレットペーパーが不足するという投稿自体は、実際には不足しているのではなくて、どうも不足するというううわさが出たんです。その背景には、トイレットペーパーの紙の原料は中国からやってきている。実際は中国から来てなかったんですけど、というデマが流れていて、それで、コロナによって輸入が止まって入ってこなくなると。だから、トイレットペーパーが不足するというううわさが流れたんです。そのときは全然トイレットペーパー自体は減っていなかったと。ただ、これに対して、そんなことないよというような否定する投稿が実は始まり始めたのと、トイレットペーパーは国産製なのでそんなことはあり得ないよということが出てきたのと合わせて、トイレットペーパーが実際に一気に需要が上がってなくなっていったというようなことが起きたと。

これがどこまで相関しているか分からないんですけど、やっぱりデマをきっかけにいろいろトイレットペーパーはやばいんじゃないかなというようなうわさが流れている中で、さらにそこを否定するような話が出てくると、それがまた一気にSNS上で情報が拡散して、みんなが知るところになると、やっぱり今のうちにトイレペーパー買っておいたほうがいいみたいな、そういう流れになっていくんじゃないかというようなことです。なので、意見が対立するというのは非常に拡散しやすいです。

というようなことが起きるというのがSNSの特徴かなと思うんです。人間は不安な状況とかそういったものというのは非常に行動に変わってくるので、実際に本当にやっぱりな

くなっちゃったというようなことが起きたと。

あと、ちょっと時間ないんでここぐらいにしますが、最近よく広がってきているのが、ラインのようなクローズドなSNSというのがだんだん広がってきたと。ツイッターのようなオープンな場所ではなくてクローズドな場所で拡散するケースというのが結構多いと。災害時は、意外とここもかなり危険であったりすると。何か発生したときに、ラインの友達同士の間でその情報が広がっていくと。うわさがどんどん広がっていって、気がついたらかなりの人に広がっているというような、そんな状況になっているということです。クローズドの拡散は結構危険でして、ツイッターみたいなオープンな場所だと誰かがこれはうそだというのに結構気づくんです。それに対して、これはうそだよというのを言いやすいということです。あと、元の情報というのが探すと見つかるので、情報の書き換わりというのはあんまり起きづらいんです。

なので、何かうそ、デマみたいなものが出てきたときも、これはちょっとおかしいんじゃないという人は必ず出てくるし、それの元ソースはこれだよとか、そういったものもやっぱり見つけやすかったりすると。

ただ、クローズドな場所というのは、誰からも見れないような状態で、知り合いだけの間で広がっていくので、なかなか見つけるのが難しいと。気がついたら物すごい数に広がっていって、それが、時間たつ上にSNSのほうに広がって、オープンな場所のほうに広がっていくというような、そんなことが起きるかなと思います。

ということで、最後ちょっとスペクティの取組だけ御紹介して。先ほどやっているというところで、ちょっとテレビで取り上げられたものですけど、映像、音声が出ないですか。ちょっと飛ばしちゃいます。

では、ちょっと時間だということなので、こちらで終わりにしたいと思いますが、我々スペクティは、こういったいろんなパターンのSNSの情報をより正確な情報を基に、災害情報というところで自治体に届けていったり企業さんに届けたりというようなことをやっているというような、そんな会社になります。

すいません、じゃ、時間ということで終わりにしたいと思います。ありがとうございます。

【宍戸座長】　村上様、ありがとうございました。

それでは、時間が限られておりますが、ただいまの御説明に御質問や御意見があれば若干承りたいと思いますが、いかがでしょうか。ある方はチャットでお知らせいただきたい

と思いますが。

生貝先生、お願いします。

【生貝構成員】　貴重な御説明、ありがとうございました。

１個だけ御質問なんですけれども、まさにこういった非常に膨大なSNSのデータを活用して分析をされている中で、例えばSNSのデータにアクセスするに当たって、何か困難や制約があったりといったような課題はあるのかということと、あるいはSNSが持っているデータ、あるいはそれ以外でも、もっとこういうデータにアクセスできればより充実した分析やサービスが提供できるのだがというふうにいったようなところがあれば教えていただきたいです。よろしくお願いします。

【村上氏】　ありがとうございます。SNSの情報、例えばツイッターの情報であると、基本的にそのツイッター社から買うというような形になるんです。なので、費用としてはかなり高くなってくると。我々は災害というところに絞っているので、ある程度絞り込みをした上で情報を取得するというようなことを行っているんですけど、その場合のこの取得の仕方というのが、基本的にキーワードをベースに取得するというものしかなかったりするんです。先ほどDeepfakeの話もありましたけど、結構映像、画像なんかで加工されていたりとかするケースがあって、必ずしもテキストに書かれているキーワードだけでの情報抽出だけではやっぱり難しいんです。

ただ、SNSのプラットフォームのほとんどがAPI公開しているんですけど、情報を絞り込もうとする場合は、大体キーワードでの絞り込みになってくるというような形なので、何かもうちょっと別な情報の取得の仕方をプラットフォーム側もできるといいなというのは、我々のほうは思っていたりします。

あと、ほかの情報との組合せというのは結構いろんな情報との組合せを行っていて、やっぱりオープンデータが大分整備されてきているんですけど、実際のデータだとかそういった情報なんかは、より活用できるような状態であるとすごくやりやすいですし、特に災害という意味でいうと、そのことが起きているかどうかだったり、災害が起きやすい地域かどうかだったりとか、あと過去のデータだったりとか、そういったものと組み合わせることで、より正さを判断できるとかいうのがあるので、そういったデータがよりオープンになるといいかなとは思います。

【生貝構成員】　どうもありがとうございました。

【宍戸座長】　ありがとうございました。まだまだお伺いしたいことはいろいろあろう

かと思いますが、予定の時間でございますので、申し訳ありませんが、御質問等ございましたら事務局のほうに構成員の皆様、お伝えをいただき、村上さんには、お手数ですが、可能な範囲で御対応いただければと思いますので、よろしくお願いいたします。

本日は貴重なインプットをありがとうございました。

本来であれば、ここで自由討議の時間も設けたかったところでございますが、予定の時間でございますので、特に何かここで御発言ということはありますでしょうか。よろしいでしょうか。

よろしければ、時間の関係もございますので、本日の会合はひとまずここまでとさせていただきたいと思います。

事務局から連絡事項のほうをお願いをいたします。

【池田消費者行政第二課課長補佐】　事務局でございます。次回会合につきましては、別途事務局から御案内をいたします。

以上です。

【宍戸座長】　ありがとうございました。

それでは、本日の議事はここまでといたします。以上で、プラットフォームサービスに関する研究会第31回会合を終了とさせていただきます。年内はこの会合が最後だということだろうと思いますので、皆様、どうぞよいお年をお迎えください。御出席いただきありがとうございました。これにて散会といたします。

以上