## 【様式２】　研究開発プロジェクトの要旨

| | |
|---|---|
| プロジェクト名 | 陰謀論への流入経路の特定と人々の傾倒を未然に防ぐフレームワークの開発 |
| 研究代表者 | 村山　太一（大阪大学産業科学研究所　産業科学 AI センター特任助教） |

【プロジェクトの要旨】

１．プロジェクトの概要

陰謀論は、健全な社会的意思決定や信頼関係醸成のためのトラスト社会の構築を阻害している。本プロジェクトでは、陰謀論に対し事前的な介入プレバンキング手法を確立することで、日本における陰謀論の蔓延を防止しトラスト社会を形成することを目標とする。具体的には、ウェブ上の行動データを活用し、陰謀論信者の行動パターンや陰謀論を信じるに至る経路を明らかにすることで、人々が陰謀論を信じることを未然に防ぐ技術を開発する。信じた陰謀論を否定することは困難だが、本技術は人々が誤った信念を持つ前に介入することで、陰謀論の被害の発生を未然に防ぐことが期待される。これにより、政治や社会の混乱を未然に防ぐだけでなく、陰謀論の被害者の発生も防止できることから、社会厚生に与える公共貢献は大きい。**本プロジェクトは陰謀論の根本的な解決に取り組むもので、情報発信やコミュニケーションを安心して行える社会を目指す画期的な取り組み**である。

２．研究開発要素①②③の概要

**（１）研究開発要素①「トラスト形成のメカニズム理解、阻害要因の分析」**

ウェブ上の行動データを対象とし、陰謀論信者の行動パターンや人々が陰謀論に陥ってしまうメカニズムを理解するための研究開発を行う。まず、研究リソースの構築のため、日本における複数のソーシャルメディアプラットフォームから陰謀論に関連するデータを収集し、陰謀論に関する投稿を自動的に検出するアルゴリズムを開発する【研究項目 1】。次に、構築した研究リソースを活用することで、人々がどのような経路で陰謀論に陥るかを視覚的に理解できる「陰謀論経路マップ」を作成する【研究項目 2】。陰謀論を信じる人々の契機を網羅的に把握する陰謀論経路マップの作成は、陰謀論の蔓延を防ぐ上で非常に重要な情報なリソースとなり政策決定などの助けになることが期待される。

**（２）研究開発要素②「分析結果を踏まえた対策の開発」**

前述の分析結果に基づき、行動データから陰謀論に陥る可能性のあるユーザを早期に発見する前兆行動発見アルゴリズムを開発する【研究項目 3】。このアルゴリズムにより、陰謀論との接触を未然に防ぐべき対象を特定することが可能となる。

**（３）研究開発要素③「社会実装手法と効果測定法の提案」**

これまでに構築した経路マップや前兆行動発見アルゴリズムに基づき、人々が陰謀論に陥る可能性のある経路への進行を未然に阻止し、人々が陰謀論へと傾倒を防止するフレームワークを提案する【研究項目 4】。未然に陰謀論の抑制を行うためにユーザの行動に介入する提案フレームワークは、公開または社会実装することで、社会での陰謀論対策に活用されることが期待される。

# 【様式３】　研究開発プロジェクトの構想

## 1．プロジェクトの目標

日本における陰謀論の蔓延を防止しトラスト社会を形成することを目標とする。このために、人々が陰謀論に陥る経路を明らかにし、経路を封じるための根本的な対策を実現するフレームワークを開発する。具体的には、ウェブ上の行動データを活用し、陰謀論信者の行動パターン分析し、陰謀論に傾く要因と前兆行動を明らかにすることで、人々が陰謀論を信じることを未然に防ぐ技術を開発する。一旦信じた陰謀論を否定することは困難だが、誤った信念を持つ前であれば陰謀論信者になることを防ぐのは比較的容易である[1]。そのため、本研究開発プロジェクトで高い効率と効果をもったフレームワークの開発が期待される。

## ２．研究開発プロジェクトで対象とする具体的な問題とその背景

ソーシャルメディアの普及により、誰でも情報の発信と受信ができるようになった。人々のコミュニケーションは密になったが、フェイクニュースの拡散といった新たな社会問題も生み出している[2]。フェイクニュースへの対策は世界的に多くの研究によって分析と開発が行われており、日本においてもファクトチェックの充実[3]や政府による啓蒙活動[4]など、実践的な活動が多く行われてきている。フェイクニュースは誤った情報に基づいて構成されるので、ファクトの指摘や公的機関による情報発信で抑制ができる。しかし、類似しているにも関わらず近年その有効な対応策の開発が問題になっているのが、**陰謀論**である。陰謀論はその定義上、不十分な証拠や思い込みによって構成されるため、事実の指摘以上の対策が必要となる。しかも、陰謀論は実際に現実に深刻な被害をもたらしている。たとえば、2021 年アメリカ合衆国議会議事堂襲撃事件は陰謀論の拡散が契機で起こったことが知られている。陰謀論は健全な社会的意思決定や、信頼関係醸成のためのトラスト社会の構築を阻害している。このような背景があり、陰謀論の研究と対策が必要にもかかわらず、陰謀論はフェイクニュースと比べ実情が明らかになっておらず、 Semantic Scholar に登録されている陰謀論に関する論文数もフェイクニュースのものと比べ 10 分の 1 以下など知見や明確な対策方法が充実していない。

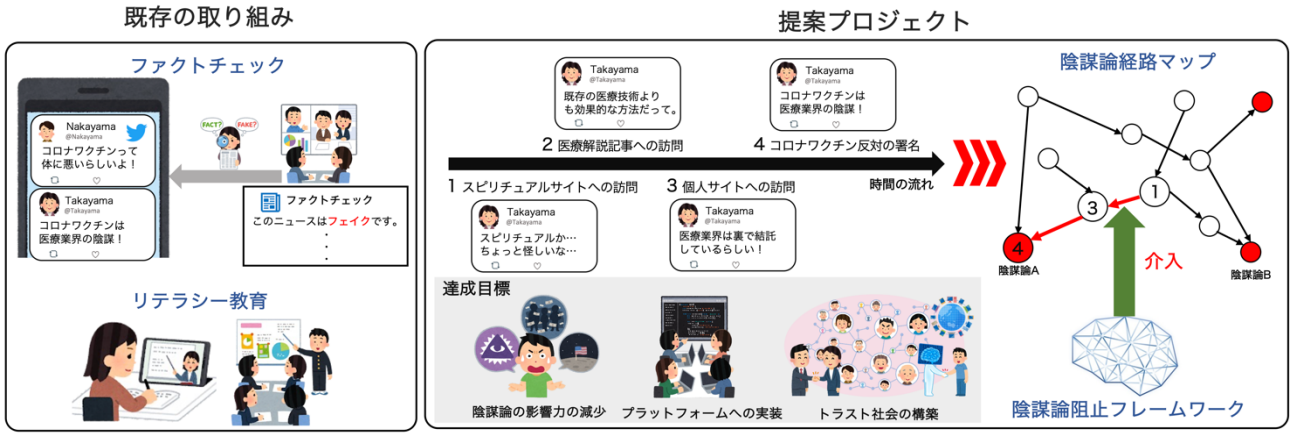
この現状の１つ目の理由として、陰謀論に対するファクトチェックの優先度が低いからである。ファクトチェックには検証作業に多くの労力と時間が必要となるため、緊急性が高く新しいニュースが対象となり、過去に検証されたニュースや多くの人が既に間違いだと考えるニュースは検証されることは多くない。そのため、似たニュースが何度も繰り返し流布する陰謀論に対して、ファクトチェックの優先順位は低くなるという問題がある[5]。

２つめの理由は、陰謀論はファクトチェックによる情報訂正の効果が限定的[6]であるからである。陰謀論は認知バイアスなどの影響により、信じている言説と矛盾する証拠に抵抗が生じ、陰謀の証拠がないこと自体ですら陰謀が存在する証拠として解釈される[7]。専門家や権威者からの情報を拒絶する作用や重大な出来事を陰謀の産物とみなすことで、反論を受け付けず信じてしまうことが知られている[4]。先に論じた、米国の陰謀グループである QAnon による米国議会議事堂襲撃事件は、ファクトチェ

ックで何度も否定されていたが。しかし、QAnon 信者は「2020 年のアメリカ合衆国大統領選挙での選挙不正があった」という強い信念を持ち続け、多くの被害がもたらされた[8]。このように、単なる事実の指摘では陰謀論対策として不十分であるという問題がある。

3 つめの理由は、陰謀論信者のクラスター化である。陰謀論信者はソーシャルメディア上でグループ形成する傾向がある。典型的な例として、QAnon のユーザは Gab や Parler といったメインストリームではないソーシャルメディアで他と孤立して集団を形成している[9]。日本でも同様の QAnon 信者によって、COVID-19 ワクチンや安倍元首相の銃撃事件に関する陰謀論を信じる人々が Twitter や Facebook でグループを形成している[10]。このように、陰謀論信者は積極的な交流を通じた強固なクラスターを形成しているため、グループの外部からの介入に頑健であるという問題がある。

以上の議論をまとめると、**単なる事実の指摘だけでは、陰謀論信者は強固な信念を捨てない**ため、陰謀論を信じた人に事後的な介入であるデバンキング(Debunking)の効果が限定的であるということである。そこで、本研究開発プロジェクトでは、陰謀論を信じる経路を明らかにするとともに、**陰謀論を信じる前の陰謀論信者予備軍への介入を行う技術を開発**する。これにより、**事前的な介入プレバンキング（Prebunking）を実現し、陰謀論信者を生み出すことを事前に抑制**する。




## 3．研究開発プロジェクトの意義や独創性

**【類似の取り組みや政策・施策、研究等では、なぜ問題が解決できないのか】**

項目 2 で議論したように、通常のフェイクニュース対策は陰謀論に対して不十分である。現状行われているメディア・情報リテラシー教育でも、自発的なファクトチェックや情報検索能力の養成が行われてはいる[11]が、その性質上即効性のある施策ではない。また、そもそも陰謀論信者は事実を指摘されても、簡単に自身の信念を否定しないので、リテラシー教育だけでは陰謀論対策として不十分である。

これに加え、陰謀論に陥る経路は多様性かつ複雑性なので、通常のリテラシー教育では陰謀論のプレバンキングにはなり得ない。たとえば、政治的陰謀論を信じている人であっても、そのきっかけは政治とは無関係なことがある。つまり、政治的陰謀論へ流入する様々な入り口が存在するということである。たとえば、日本においては EM 菌などのオーガニック食品やスピリチュアルへの関心から陰謀論へと結びつくケースが増加している。これは、 コンスピリチュアリティ （conspirituality）とよばれ、陰謀論とスピリチュアルが複雑に絡まった現象として知られている[12]。他にも、ワクチンに関する陰

謀論への入り口として、代替医療への関心が知られており、代替医療への関心がきっかけでワクチン関連の陰謀論を信じる経路が指摘されている[13]。つまり、ある陰謀論へ行き着く経路への入り口は流入先の陰謀論とは無関係だったり、関係が薄い場合が多い。

よく用いられるプレバンキング手法の一つとして、心理的予防接種があり、焦点を絞ったリテラシー教育としても捉えることができる [14]。心理的予防接種は、人々が誤った情報に触れる前に正しい情報をピンポイントで知らせ、陰謀論を予防する試みである。しかし、先に論じたように陰謀論を信じるに至る経路は複雑かつ多様なため、網羅的に情報を予防的に接種することは困難である。

**【プロジェクトの意義や独創性】**

そこで、申請グループは人々が陰謀論を信じてしまう経路を網羅的に把握することで、その経路への進行を未然に防止する手法の検討を行う。陰謀論を信じる人々の契機をできる限り網羅的に把握することは、陰謀論の蔓延を防ぐ上で非常に重要な情報になりこのプロジェクトの意義は高い。また、防止策を開発することで未然に陰謀論を抑制するので、社会の信頼性と安定性が実現できる。

本プロジェクトは、事後介入であるデバンキングではなく、事前介入であるプレバンキングを効果的に行うため、人々の陰謀論への流入の全体像を包括的に把握することに独創性がある。この全体像の把握のために、複数のデータソースに対しドメインアダプテーションを用いる点にも本プロジェクトの独創性がある。複数のデータソースを使う際、データソースやユーザによって特徴量の性質が異なるため単純にはこれらのデータセットを統合して分析することは難しい。本プロジェクトはドメインアダプテーションと呼ばれる機械学習手法で、データセット・ユーザ・コンテキストの差異を考慮する。**本プロジェクトの取り組みは世界で初めてのもので、独創性と新規性が非常に高いものとなっている。**

陰謀論の氾濫は QAnon による米国議会議事堂襲撃事件やワクチン忌避などといった、実社会に大きな影響を年々与えており、陰謀論の蔓延を防止することは喫緊の課題である。そのためこの問題の根本的な解決に取り組む本プロジェクトの意義は高い。本プロジェクトで開発するプレバンキングのためのフレームワークは陰謀論の被害者が発生することを未然に防ぐことを目的とする。つまり、政治や社会の混乱を未然に防ぐだけでなく、それによる被害者の発生を防ぐことができる。**陰謀論にとらわれると、その個人の人生全体に負の影響を与えてしまことからも、本プロジェクトの意義は非常に高い。**

## ４．プロジェクトの目指す社会像、将来ビジョン

これまでのトラスト形成は、インターネット上の取引やフェイク拡散、AI の正しい扱い方などに着目している[15]。トラスト形成のための施策の多くは、フェイクニュースやインターネット上の暴言などの被害に合わないための取り組みなど、いかに無辜の人々が被害者にならないか、被害者になってもすぐに回復できるための施策に重点的に取り組まれてきた。それに対し、陰謀論を信じてしまう人々は、コロナワクチンへの忌避感や米国議会議事堂襲撃事件などの例が示すように、ファクトチェックなどの既存のトラスト形成のための施策の効果は示されていないだけでなく、被害者であるとともに加害者になってしまうという特殊な性質がある。申請グループの提案するプロジェクトでは、彼らが加害者にな

る可能性のある経路を発見し、その経路に流入することを未然に防ぐというアプローチをとることで、誰もがコミュニケーションや情報収集が安心して行えるというトラストの形成を目指す。さらに、このプロジェクトは**被害者になることを防ぐだけでなく、加害者になることも防ぎ、情報発信やコミュニケーションを気軽に安心して行える社会を目指すという、これまでのトラスト形成の取り組みから見落とされていた対象に注目する**画期的な取り組みである。

提案プロジェクトでは、上記の社会を構築するために、１：日本における陰謀論検出のデータセット構築、２：謀論の入り口となるトピックの発見、３：陰謀論に陥るユーザの前兆の発見アルゴリズムの開発、そして４：陰謀論にはまる前兆行動の阻止フレームワークの提案に取り組む。これらの取り組みの**研究成果を得るには、グループ内で協力によりこれまでの共同研究の経験から 2-3 年以内に達成**できると考える。しかし、得られた研究成果を実際の社会システムやプラットフォームに落とし込むことは慎重に行う必要がある。申請グループのプロジェクトは、陰謀論を信じてしまう経路を発見し向かうことを防止するものであることから、「誤った情報」への接触を防ぐだけではなく、「誤っていない情報」への接触を防いでしまう可能性がある。このことから、プロジェクトの成果がプラットフォームのモデレーションやポリシーに適合するか慎重に議論を重ね調整していく必要がある。申請グループは、**ソーシャルメディアプラットフォームなどのサービス事業者も納得いく形で、誰もがインターネット上でのコミュニケーションや情報収集が安心して行え、陰謀論に陥りづらくなるというビジョン**を目指す。

**５．プロジェクトの目標達成のために解決すべき課題、ボトルネック**

提案プロジェクトを進める上での課題・障壁として以下のものが挙げられる。

**■多様なデータソースへの対応**：陰謀論は様々なプラットフォームで広まるため、それぞれのプラットフォームからデータを取得することが必要である。しかし、各プラットフォームから得られるデータは形式や性質がそれぞれ異なるため、分析やモデリングの際に同時に処理することが困難である。従来のアプローチではプラットフォームごとにモデル構築や分析を行うのが一般的だが、陰謀論やそれに着目するユーザの特性はプラットフォームに依存しない側面も多いため、個別の分析は非効率である。

**▶解決策**：自然言語処理や数理モデリングによる前処理技術やドメインアダプテーションを活用し、統一的な形式に変換するプロセスを導入することで、異なるプラットフォームのデータを同時に解析するための汎用的データ処理フレームワークを構築する。これにより、陰謀論の広がりや特徴を総合的に理解し、社会への影響や対策についてより具体的な洞察を得ることが期待される。

**■データのドメインシフト**：インターネット上のユーザの行動やコンテンツのトレンドは時間とともに変化するため、開発したモデルやアリゴリズムが常に有効である保証は得られない。

**▶解決策**：新たなプラットフォームやトピックの出現に対応するために、構築したモデルやアルゴリズムが更新可能であることが求められる。具体的には、統計的モデル更新手法である Streaming アルゴリズム[16]や Continuous Learning[17]の枠組みを取り入れることで更新可能なモデル構築を達成する。

**■モデリング結果を補強する理論的背景**：陰謀論に傾く人々の行動パターンや前兆行動を、データの側

面から把握するデータ駆動型アプローチは主観的なバイアスの影響を最小限に抑えるという利点がある一方で、データの観察から得られる理論的な背景による担保が不十分であるという課題がある。

▶**解決策**：社会学や心理学などの人文科学の理論的背景を活用することで、陰謀論に関連する行動や思考の特徴を解釈し、モデルを補強することを目指す。人文科学における専門的な知識は協力者の知見を統合し、モデルの解釈の向上に活用する。

■**提案フレームワークの評価**：提案プロジェクトで構築したフレームワークが実際に陰謀論の拡散や信じ込みを本当に抑制できるかどうかの評価することは、既存の指標では困難である。

▶**解決策**：構築フレームワークを社会実装する前に、事前に試験運用やクラウドソーシングを活用して効果の事前調査を行うことで、フレームワークの機能や特性、陰謀論拡散への抑制効果の事前評価が期待できる。社会実装した後でも、A/B テストや介入を受けたユーザの追跡調査を行うことでフレームワークの有効性の評価が達成できる。１つの評価指標を用いるのではなく、これらの評価手法を組み合わせ様々な観点でフレームワークを検証することで、有効性を客観的かつ綿密に評価につながる。

■ **SNS プラットフォームなどのサービス事業者との連携**：構築フレームワークを社会実装し多くの人々の情報収集プロセスに介入するためには、既存の SNS プラットフォームとの密な連携が必要である。

▶**解決策**：この課題は申請グループによる研究成果や交渉の努力に大きく依存する。提案フレームワークの有効性を提示できる研究成果やデモの作成を行うことで、プラットフォームとの協力体制をとり、陰謀論への関心や信じ込みを軽減するための取り組みを支援することを目指す。

## ６．プロジェクトの創出する成果の活用・展開

### （１）プロジェクトの創出するアウトプット、想定するアウトカム

提案プロジェクトでは１：日本における陰謀論検出のデータセット構築・検出アルゴリズムの構築、２：陰謀論の入り口となるトピックの発見、３：陰謀論に陥るユーザの前兆の発見と予測アルゴリズムの開発、４：陰謀論にはまる前兆行動の阻止フレームワークの提案に取り組み、以下の成果を創出する。

・**陰謀論検出データセットと検出アルゴリズム**：日本語のソーシャルメディア上で陰謀論に関連する投稿を行うユーザを包括的に収集し、基盤となるデータセットを構築する。また、このデータセットを活用して、陰謀論を信じるユーザを検出するための機械学習ベースのアルゴリズムを開発する。データセット、アルゴリズムともに日本語ベースのものは未だ存在せず、本成果は日本における陰謀論研究の重要なリソースとなることが期待される。これらは日本における陰謀論の特徴や傾向を理解する手がかりとなるとともに、陰謀論の拡散を早期に把握し、社会への悪影響を最小限に抑えることが期待される。

・**陰謀論経路マップ**：陰謀論を発信するユーザの行動を分析・モデリングすることで、陰謀論への流入経路を可視化したマップを作成する。これは、陰謀論への対策や予防策の立案に貢献するもので、インターネット上のユーザが陰謀論に接触する可能性のある状況を未然に防ぐための重要なリソースとなる。

・**陰謀論防止フレームワーク**：陰謀論に陥りやすいユーザの行動パターンと前兆行動発見アルゴリズムに基づき、事前に陰謀論に陥る経路を阻止し傾倒を防止するフレームワークを提案する。ユーザの行動に介入する仕組みを設計し、公開または実装することで、社会全体で陰謀論に陥る人数を減らす効果が

期待される。陰謀論に対する関心や傾倒が低下し、より健全なトラスト社会が構築されることを目指す。

**（２）プロジェクトの成果の波及効果、インパクト**

**【学術的・公共的価値の創出】**提案プロジェクトは、陰謀論に陥ることを事前に防止する新規性の高いアプローチを提案し、計算社会科学の分野において新たな学術的貢献をもたらすことが期待される。それだけでなく、日本におけるウェブ・ソーシャルメディア上での陰謀論の拡散に特に注目し、トラスト社会の構築と安定化に寄与するという公共的価値を生み出すことが期待できる。

**【現在及び将来の社会・産業ニーズへの貢献、国内外の他の分野・地域への波及・展開】**提案プロジェクトによる、陰謀論の早期検出や流入経路の可視化、陰謀論に陥るユーザの防止フレームワークは、政策策定者やプラットフォーム事業者などの様々なステークホルダーに対し、陰謀論対策の基盤を提供し情報セキュリティや意思決定のプロセスにおいて重要な役割を担う。将来的にも陰謀論の拡散や影響の予防に向けたニーズは高まることから、本プロジェクトから得られる知見は重要である。また、陰謀論は国境を超えて存在することから、提案プロジェクトの取り組みは、国際的な陰謀論研究コミュニティやステークホルダーに対しても重要な知見やモデルを提供することが期待される。

**【SDGs への貢献】**提案プロジェクトの SDGs への貢献は以下のようになる。

**・SDG 4（質の高い教育をみんなに）**：陰謀論に陥ることを防ぐ提案プロジェクトは、人々の受け取る情報の信頼性を向上させ、情報教育やメディアリテラシーの強化に寄与する。

**・SDG 9 （産業と技術革新の基盤をつくろう）**：陰謀論の早期検出や経路マップの開発は情報セキュリティやデジタルトラストの向上に寄与する。これは、イノベーションとテクノロジーの発展を促進し、社会的なリスクや混乱を軽減する基盤構築に寄与する。

**・SDG 16（平和と公正をすべての人に）**： 陰謀論は社会に対する悪影響をもたらす可能性があり、社会の平和と公正に対する脅威となり得る。提案プロジェクトは、陰謀論の拡散を抑えることで社会的な不安や分断を軽減し、トラスト社会の構築に貢献する。

**（３）プロジェクトの限界**

**・オフラインの影響の特定**：人々のオフラインでの行動や意見形成はデータとして残らないため、オンラインと比較して観察や分析が困難である。特に、家庭や親しい友人との会話、教育環境、地域コミュニティなどのプライベートな空間で人々が陰謀論に影響を受ける可能性は高いことから、このような場所にも直接アプローチしていくことが今後の課題となる。この課題に対処するためには、オフライン環境に向けたアンケート調査やインタビューなどの定性的な研究方法を活用することが考えられる。

**・社会全体の協力**：このプロジェクトの成果を最大限に活用するためには、政策や教育やステークホルダーとの協力が必要である。しかし、それらの変革をもたらすには時間とコストがかかり、プロジェクト自体の枠組みを超えた課題である。プロジェクトの成果を広く共有するなど、社会全体での理解と受け入れを促進するためアウトリーチ活動を計画することが望まれる。

# 参考文献

[1] Jolley, D., & Douglas, K. M. (2017). Journal of Applied Social Psychology, 47.
[2] Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). Science, 359.
[3] 日本ファクトチェックセンター（JFC）. https://factcheckcenter.jp/ (Accessed: 2023/06/27)
[4] 総務省. https://www.soumu.go.jp/use_the_internet_wisely/special/fakenews/ (Accessed: 2023/06/27) [5] Byford, J. (2011). Springer.
[6] Carey, J. M., Guess, A. M., Loewen, P. J., Merkley, E., Nyhan, B., Phillips, J. B., & Reifler, J. (2022). Nature Human Behaviour, 6.
[7] Byford, J. (2011). Springer.
[8] Amarasingam, A., & Argentino, M. A. (2020). CTC Sentinel, 13.
[9] Bär, D., Pröllochs, N., & Feuerriegel, S. (2023). Proceedings of the International AAAI Conference on Web and Social Media, 17.
[10] Business Insider. https://www.businessinsider.jp/post-228586 (Accessed: 2023/06/27)
[11] McGrew, S., Breakstone, J., Ortega, T., Smith, M., & Wineburg, S. (2018). Theory & Research in Social Education, 46.
[12] Ward, C., & Voas, D. (2011). Journal of Contemporary Religion, 26.
[13] Paytubi, S., Benavente, Y., Montoliu, A., Binefa, G., Brotons, M., Ibáñez, R., ... & Costas, L. (2022). BMJ, 379.
[14] Roozenbeek, J., Van Der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Science Advances, 8.
[15] CRDS 研究開発センター. https://www.jst.go.jp/crds/report/CRDS-FY2022-WR-05.html (Accessed: 2023/06/27)
[16] Gong, S., Zhang, Y., & Yu, G. (2017). Proceedings of the VLDB Endowment, 11.
[17] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Proceedings of the National Academy of Sciences, 114.