令和7年度　インターネット上の偽・誤情報等への対策技術の開発・実証事業

グローバル・メタアナリシスと国内実証による対策技術の有効性の研究・調査

成果報告書

2026/3/19

研01_株式会社新領域安全保障研究所

目次

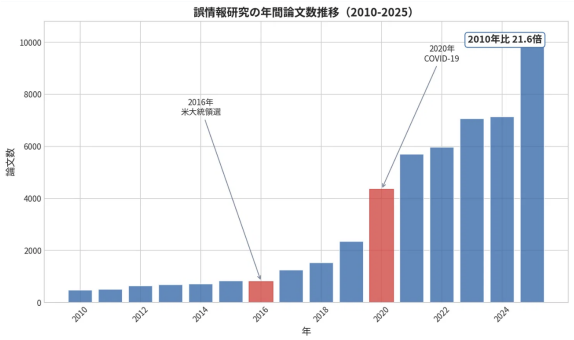
目次

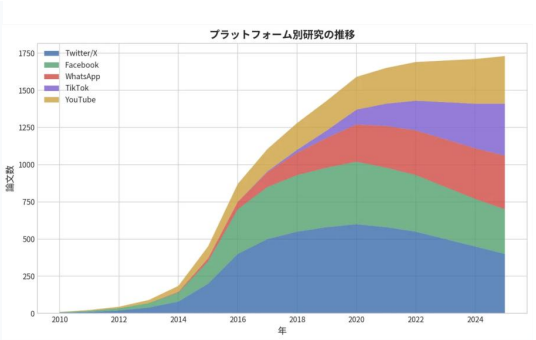
1-1. 研究・調査のサマリ

アプローチする課題・目指す姿	• 英語圏偏重の既存研究を相対化し、日本固有の文脈に即した対策を立案するため世界的な誤情報研究動向を明らかにする。その上で。日本の文脈を再現可能なLLMエージェントを用いた「シリコンサンプリング」による安全な検証環境の構築を実施し、コストの低い再現実験の手法を提示する。		
研究・調査区分	偽・誤情報対策技術に係る研究	実施体制（下線：研究・調査主体）	株式会社新領域安全保障研究所

研究および有効性等に関する検証の取組・成果

- 誤情報研究動向の分析：OpenAlexのAPIを用い、誤情報研究が過去15年で爆発的に拡大し、選挙・パンデミック・生成AI等の社会的危機が研究を牽引する「イベント駆動型」であることを整理。また、研究対象の地理・言語・プラットフォーム偏りなどを可視化し、誤情報研究における欧米の「政治」への関心に対しグローバルサウスでは「健康」への関心が高いことなどを示した。
- 国内実証（シリコンサンプリングの提示・比較検証）：WVS（世界価値観調査）日本データを用い、LLMエージェントによる社会シミュレーションを構築。個別ペルソナ法／集約統計法、GPT系／Gemini系、複数介入条件を組み合わせて比較した。モデル・手法・入力情報によって推定が揺れること、介入によっては平均では効いて見えても特定層で逆効果を取り得ることを示唆。





研究・調査にあたっての課題・展望

- 研究動向分析はOpenAlex＋タイトル中心の抽出のため、非英語圏・ローカル語／領域固有語の取りこぼしがあり得る。→多言語・多概念での研究ウォッチへ拡張。
- シリコンサンプリングは有望だが、現時点では妥当性（現実をどれだけ当てているか）検証のベンチマークが不足。→比較用の人間データ整備と評価枠組みの確立が必要。
- ペルソナ設計（どの情報を与えるか）や、ネットワーク効果・反復曝露・画像/動画など現実環境要因の取り込みが今後の課題。→より実環境に近いシミュレーションへ改善し、日本での誤情報対策に資するエビデンス創出へつなげる。

代表者コメント

新領域安全保障研究所
代表取締役　齋藤孝道
（明治大学
理工学部教授）

世界的な研究動向を理解することは重要ですが、一方海外で成功したとされる対策手法を無批判に日本へ導入することは、予期せぬ副作用を生むリスクを伴います。今回は世界の研究動向を追いかけながらも、海外の知見についてどのように日本に適用するのか低コストで検証する手法の提示を目指しました。

目次

2-1. 研究・調査によりアプローチする課題

研究・調査によりアプローチする課題

課題①：誤情報研究の全体像の欠如

社会課題化と学術的関心の高まり

- 2016年の米大統領選やBrexit（Post-Truth Era）、2020年のCOVID-19パンデミック（Infodemic）、そして近年の生成AIの台頭を経て、誤情報は民主主義や公衆衛生に対する重大な脅威として認識されている[1][2]。
- これに伴い、世界的に関連研究は急増していると推察されるが、その全体像やトレンドの変遷、主要な論点を定量的に把握した包括的な分析は不足している[3]。

知見の偏りの可能性

- 既存の社会科学研究は、欧米のデータや文脈に偏る傾向が指摘されている[4]。誤情報対策においても、英語圏を中心とした知見がそのままグローバルスタンダードとして扱われ、非英語圏やグローバル・サウスの実態が見過ごされている懸念がある[5][6]。
- このため、現状の研究動向を俯瞰し、「何が研究され、何が研究されていないのか（Missing Link）」を客観的に明らかにする必要がある。

参考文献

[1] Lazer, D. M., et al. (2018). The science of fake news. *Science*, 359(6380), 1094-1096.
[2] Cinelli, M., et al. (2020). The COVID-19 social media infodemic. *Scientific Reports*, 10, 16598.
[3] Ha, L., Andreu Perez, L., & Ray, R. (2021). Mapping Recent Development in Scholarship on Fake News and Misinformation, 2008 to 2017: Disciplinary Contribution, Topics, and Impact. American Behavioral Scientist, 65(2), 290-315.
[4] Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3), 61-83.
[5] Humprecht, E., Esser, F., & Van Aelst, P. (2020). Resilience to Online Disinformation: A framework for cross-national comparative research. *The International Journal of Press/Politics*, 25(3), 493-516.
[6] Ong, J. C., & Cabañes, J. V. A. (2018). *Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines* (Report). Newton Tech4Dev Network.

2-1. 研究・調査によりアプローチする課題

研究・調査によりアプローチする課題

課題②：海外対策技術の「日本社会への適合性」の不透明さ

- 欧米諸国では、Accuracy Nudge（正確さ想起）やPrebunking（情報の予防接種）といった介入手法の効果が先行研究で報告されている 。
- しかし、情報の受容態度やシェアの動機は、文化的背景（個人主義か集団主義か等）やプラットフォームの利用実態に強く依存する 。
- 日本固有の文脈において、これらの海外製手法が同様に機能するのか、あるいは文化的な摩擦により「逆効果（いわゆるBackfire Effect）」を生む可能性があるのかは、十分に検証されていない 。

課題③：実証実験における「倫理的・実務的制約」

人間を対象とした検証の困難さ

- 対策の有効性を検証するために、実際のSNSユーザーに対して意図的に偽情報を曝露することは、倫理的な観点から実施が極めて困難である 。
- また、アンケート調査ベースの実験では、回答者が社会的に望ましい回答を行うバイアスが生じやすく、実際の共有行動を正確に予測できない課題がある。

安全な検証環境の必要性

- 新たな偽情報のナラティブや対策手法が次々と現れる中、倫理的リスクを回避しつつ、日本人の行動特性を反映した迅速かつ再現性のある検証環境（シミュレーション技術）の確立が求められている。

2-2. 研究・調査により目指す姿・ゴール

研究・調査を通して目指す姿・ゴール

➢ 以下は、本研究・調査期間のみではなく、最終的に目指す姿やゴールを記述しており、本研究・調査期間で以下の全ての達成を目指すという意図で記したものではない。

最終ゴール：データと文脈に基づく「日本型」対策モデルの確立

- インターネット上の偽・誤情報対策において、経験則や海外事例の無批判な導入から脱却し、「客観的データ」と「日本固有の文脈」に基づいた証拠に基づく政策立案（EBPM）の基盤を確立が必要。そのために以下のようなゴールの達成が必要。
  - **グローバルな知見とローカルな実態の統合**
    - 近年、世界的に「偽・誤情報（Misinformation）」への関心が高まり、対策技術の研究が進展しているが、その多くは欧米の文脈で語られている。本事業は、グローバルな学術トレンドを定量的に把握した上で、日本社会特有の文化的・社会的背景（ローカルな文脈）を加味し、欧米の理論がそのまま適用できない領域を埋めることを最終的に目指す。
  - **「現象の記述」から「解決策の実装」への転換**
    - 単なる実態調査に留まらず、具体的な介入手法（ナッジ、リテラシー教育等）の有効性を検証することで、プラットフォーム事業者や政策立案者が即座に参照可能な実践的知見の提供が必要。そのためにはその有効性を手軽に検証できる環境の構築や手法の提案が重要となる。

2-2. 研究・調査により目指す姿・ゴール

研究・調査を通して目指す姿・ゴール

最終ゴールのゴール達成のため以下のような中間ゴールの達成が必要となる。

中間ゴール① 「対策技術に係る研究」におけるゴール：構造的バイアスの可視化と研究地図の策定

- **学術的「空白地帯」の特定と是正**
  - 現在の誤情報研究は、米国・英国などの英語圏や、データ取得が容易なTwitter/X等の特定プラットフォームに極端に集中している現状がある。本研究では、こうした「地理的バイアス」や「プラットフォーム・バイアス」を定量的に可視化することを目指す。これにより、さまざまなソーシャルメディア環境、あるいは日本を含む非英語圏で発生しているリスクが見過ごされている現状に警鐘を鳴らし、より包括的な対策の必要性を訴求する。
- **危機の変遷に即したアジェンダの再定義**
  - 誤情報のトピックが「政治的プロパガンダ」から「公衆衛生（パンデミック）」、さらには「AI生成偽情報」へと推移している実態を明らかにする。これにより、誤情報対策を単なるメディア論の問題としてではなく、民主主義の守護および国民の生命・健康を守るための安全保障上の課題として再定義し、分野横断的な連携を促すための共通認識を形成する。

中間ゴール② 「有効性等に関する検証」におけるゴール：倫理的かつ高精度な検証環境の構築

- **「シリコンサンプリング」による次世代検証モデルの確立**
  - 実際のSNSユーザーを対象とした偽情報実験は、倫理的な制約が大きく、実施が極めて困難である。本検証では、LLMエージェントを用いて仮想的な社会反応を再現する「シリコンサンプリング」の手法を確立することを目指す。これにより、倫理的リスクを回避しながら、多様なペルソナ（属性・価値観）に対する介入効果を、再現性を持って反復検証できる「サンドボックス（実験場）」を社会に提供する。
- **文化的摩擦（逆効果）の回避と最適解の特定**
  - 欧米の実証研究で有効とされる「正確さ想起（Accuracy Nudge）」等の手法が、日本人の心理特性（不安回避や同調志向等）においては「逆効果（Backfire Effect）」をもたらす可能性を検証する。その上で日本社会において副作用が少なく、かつ高い抑制効果が見込める介入手法を低コストで特定し、日本版の偽情報対策ガイドライン策定に資するエビデンスを創出する手法を生み出す。

2-3.研究・調査により期待される偽・誤情報対策への効果

研究・調査により期待される偽・誤情報対策への効果

① 政策レベルの効果

- 本事業における「学術動向の全容解明」と「日本固有の検証」は、国家レベルでの偽・誤情報対策戦略を策定するための羅針盤となり、以下の効果をもたらす。
  - 研究リソース配分の最適化
    - 過去の学術動向分析により、「英米以外での研究」や「クローズドなSNS（LINE等）」における研究はどのように行われているのか明らかにできる。これにより、政府や助成機関は、海外でどのような研究が行われているか把握することができ、既に飽和している領域ではなく、真にリスクが高い未解明領域へ集中的に研究資源を投下することが可能となり、対策の空白を埋めることができる。
  - 誤った「グローバル・スタンダード」導入による混乱の回避
    - 欧米で主流の対策（ナッジ等）が日本社会では「逆効果」となり得ることをシミュレーションで示した。グローバルな潮流（分析）とローカルな検証（実証）を組み合わせることで、海外のベストプラクティスを無批判に導入して失敗するリスクを回避し、日本の文化的土壌に適合した安全で実効性の高いガイドラインを策定できる。

2-3.研究・調査により期待される偽・誤情報対策への効果

研究・調査により期待される偽・誤情報対策への効果

② 実務・技術レベルの効果

- プラットフォーム事業者や技術開発者に対し、「どこにリスクがあるか（動向分析）」と「どう対処すべきか（検証結果）」の両面からエビデンスを提供することで、以下の効果が期待される。
  - 関心が集まっているプラットフォームの把握
    - 動向分析により、社会的影響力が大きいにもかかわらずこれまで対策研究が手薄であったプラットフォーム（YouTube, TikTok, クローズドチャット等）を特定しようとした。研究が対象としているプラットフォームを明らかにすることで、今何に世界的に注目が集まっているのか把握できる。
  - 英米とグローバルサウスの関心の違いの把握
    - 誤情報研究が、欧米では「政治（民主主義）」、グローバルサウスでは「健康（公衆衛生）」の2大テーマに収斂している現状を可視化した。これまで縦割りになりがちだった「情報通信」「医療」「AI技術」の各分野が共通の課題認識を持ち、相互に知見を共有する分野横断的な連携（クロス・セクター）体制の構築を促進するため前提となる地域による関心の違いがここからわかった。
  - 高速かつプレシジョン（精密）な介入の実装
    - シリコンサンプリング技術の確立により、人間を対象とした実験が困難な状況下でも、AIエージェントを用いて数日で対策効果を検証可能となる。これにより、生成AIによる新たな脅威に対しても即座に有効な「情報の予防接種（Prebunking）」等のシナリオを開発し、ターゲット層の特性に合わせて打ち手を使い分ける高度な対策の実装が実現する。

目次

3-1. 研究の全体像

研究に係る取組・成果の全体像

■「危機」が牽引する爆発的な論文数の量的拡大（2010年比 21.6倍）

- OpenAlex収録の約5.2万件の文献分析によると、誤情報研究は過去15年間で21.6倍に急増した。「2016年 米大統領選（政治）」「2020年 COVID-19（公衆衛生）」「2023年 生成AI（技術）」という3つの社会的危機が研究を牽引していると考えられる（イベント駆動型）。
- 政治的な意味を有するようになってしまった「Fake News」から、包括的な「Misinformation」という語がより使われるようになった。

■ 英語圏と新興国の関心の違い

- 英語圏（米英）：論文の絶対数で圧倒的シェアを持つ。「選挙」や「政治的分極化」への関心が高く、民主主義制度の守護が主要テーマ。
- グローバル・サウス（インド/インドネシア等）： 近年の成長率が著しい（インドネシアは2015年比32倍）。「ワクチン」や「健康」への関心が突出しており、実生活に直結する問題として扱われている。

■ プラットフォーム研究の変遷と今後の課題

- これまでデータ取得が容易なTwitter (X) 研究が主流であったが、近年のAPI有償化・高騰を受け、その数は減少傾向に転じている 。
- YouTubeやTikTokなど他の影響力あるメディアや、グローバル・サウス等でもよく使われているWhatsApp（暗号化アプリ）など、データ取得が困難なプラットフォームの研究も増加し続けている。

3-2. 研究の個別詳細

論文データの取得方法

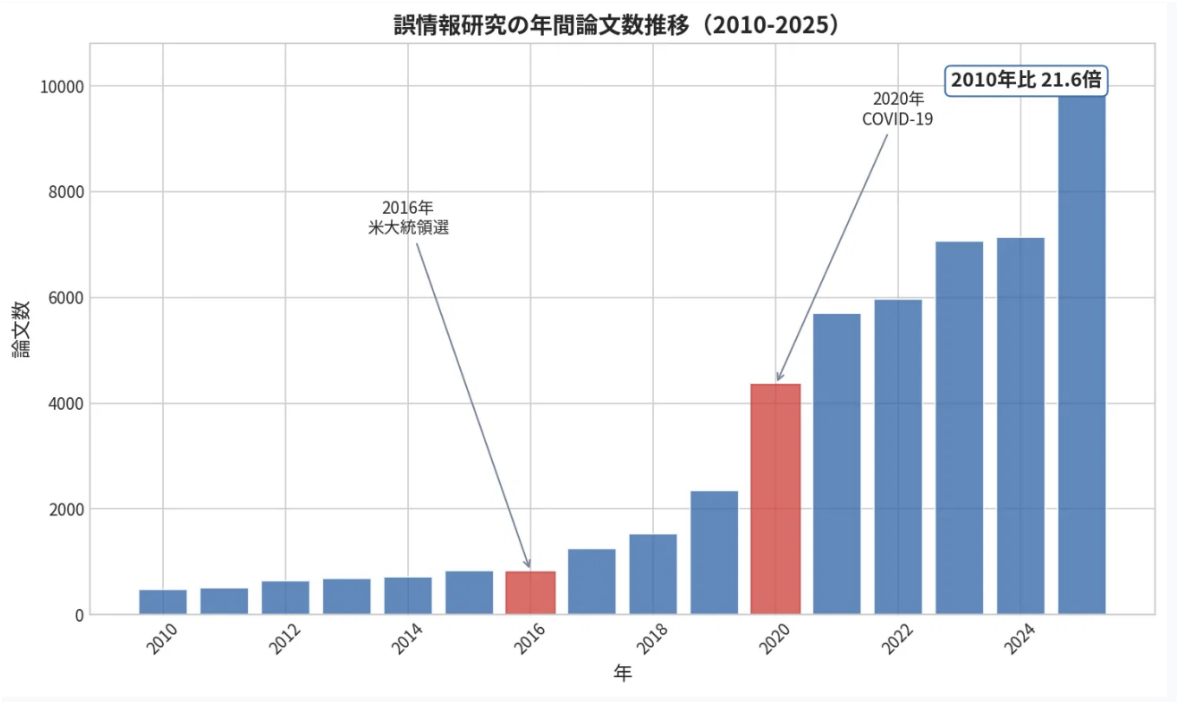
- データソース：OpenAlex
  - 選定理由： 全世界の学術文献約2.5億件を収録するオープンデータベース「OpenAlex」を採用。従来の有料データベース（Web of Science等）と比較して、文献やプレプリント（査読前論文）のカバレッジが広く、急成長し、領域横断的な誤情報研究の全体像を捉えるのに最適であるため。
  - OpenAlexの公式APIを利用して論文を取得した。
    （今後特に断りのない論文数データ・グラフはOpenAlexを利用して当社が算出したもの）

- 検索クエリと抽出基準
  - キーワード設定：タイトルに misinformation、disinformation、fake news、infodemic （いずれも英語の単語だが、これらの単語と同じスペルの単語を有する英語以外の論文が含まれることもある）のいずれかを含む文献を対象とした。抄録・サマリーのみに当該単語を含む文献はノイズが多くなってしまうために除外し、タイトルにこれらの単語を含む文献に限定している（論文数の過小評価になっている可能性に注意）。
  - 対象期間： 2010年1月1日から2025年12月31日まで
  - 文献タイプ： 査読付きジャーナル論文に加え、速報性が重視される分野特性を考慮し、プレプリント（arXiv等）や国際会議録（Proceedings）も分析対象に含めた

3-2. 研究の個別詳細

論文データの急増

- 誤情報研究の年間論文数は、2010年の368件から2024年には約8,000件規模へと劇的に増加しており、この15年間で21.6倍という驚異的な拡大を遂げている。年平均成長率は24.5%に達しており、これは一般的な学術分野の成長速度を遥かに上回る。



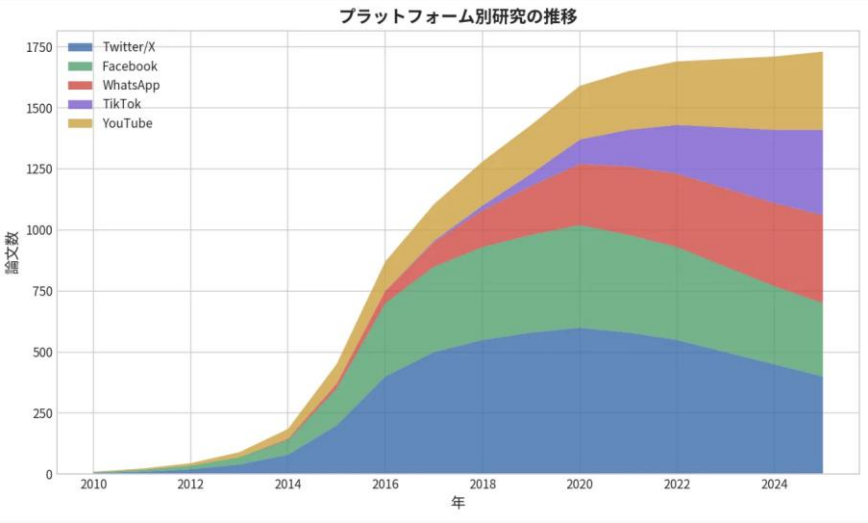
「危機」が研究を牽引する「イベント駆動型」の論文増加

- 2017年の急伸（政治的危機）：2016年の米大統領選とBrexitにおける「フェイクニュース」の氾濫を受け、翌2017年には論文数が前年比1.5倍に急増した 。これは学術界が「民主主義の危機」に対して即座に反応したことを示している。
- 2020年の加速（公衆衛生の危機）：COVID-19パンデミックの発生により、誤情報は人命に関わる「インフォデミック」として再定義され、成長曲線はさらに急角度で上昇した 。
- 2023年以降の持続（技術的危機）：生成AI（ChatGPT等）の普及により、高品質な誤情報の大量生成という新たな脅威が出現し、研究ブームは沈静化することなく続いている 。

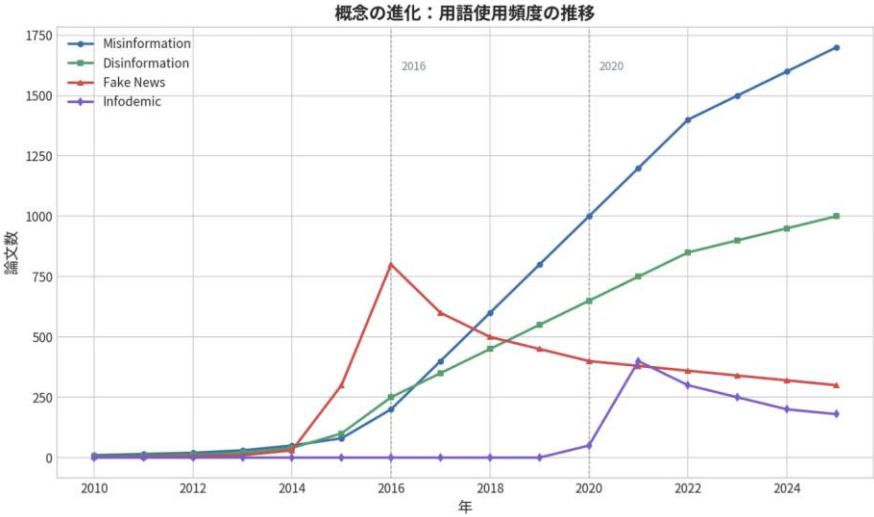
3-2. 研究の個別詳細

対象プラットフォーム別の論文数・用語の変遷

- 論文のタイトルや抄録に各プラットフォーム名が含まれるものをキーワード検索で抽出し、年ごとに集計
- プラットフォーム別の研究数を見ると、APIによるデータアクセスの容易さから、X（旧Twitter）を対象とした研究が他のプラットフォームより大きい傾向にあったが、API料金高騰化から近年は低下傾向。



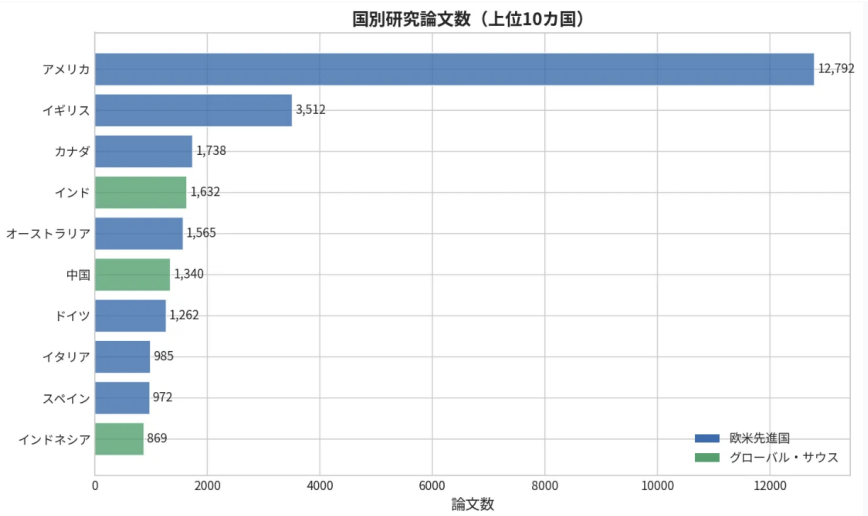
- 対象論文群のタイトルと抄録テキストに対し、各キーワードの年間出現回数をカウント。
- 用語の変遷を見ると、2016年米大統領選で急増した「Fake News」は、政治的な乱用を背景に学術的な使用が減少傾向にある。代わって、意図の有無を問わない包括的な「Misinformation」が標準的な用語として定着し、2020年にはCOVID-19の影響で「Infodemic」が特異的に急増した。

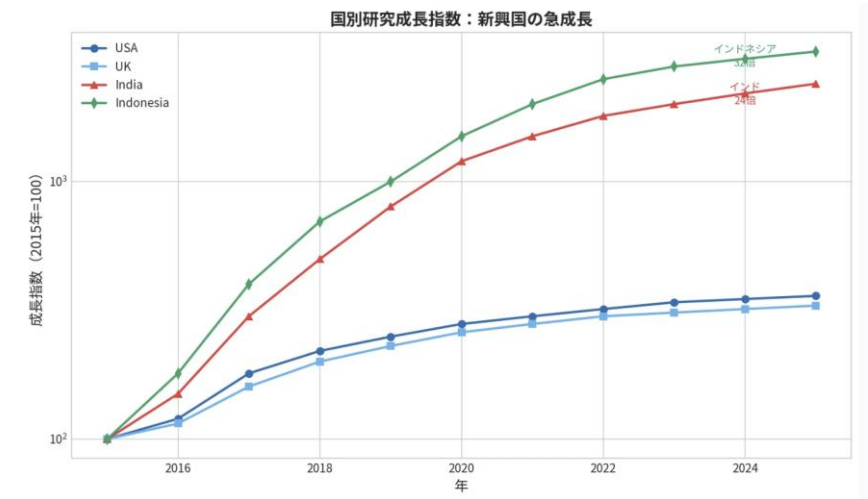


3-2. 研究の個別詳細

国別研究論文数

- 論文執筆者が所属する機関の国別の論文数を算出した。（注：複数著者が含まれる論文については、それぞれの著者についてカウントされている。）
- 欧米圏の圧倒的なシェア（右上図）
  - 論文総数では米国（12,792件）と英国（3,512件）が突出しており、依然として英語圏が研究の中心地となっている 。
- グローバル・サウスの爆発的成長（右下図）
  - 2015年比の成長率では、インドネシアが32倍、インドが24倍と急増。ソーシャルメディアの普及や大規模選挙などを背景に、アジア新興国が誤情報研究の「最前線」として台頭している。
- 成長曲線の二極化
  - 欧米（米3.6倍・英3.3倍）が成熟期に入り成長が鈍化する一方、新興国は指数関数的な伸びを示しており、研究の地理的重心が拡散しつつある 。

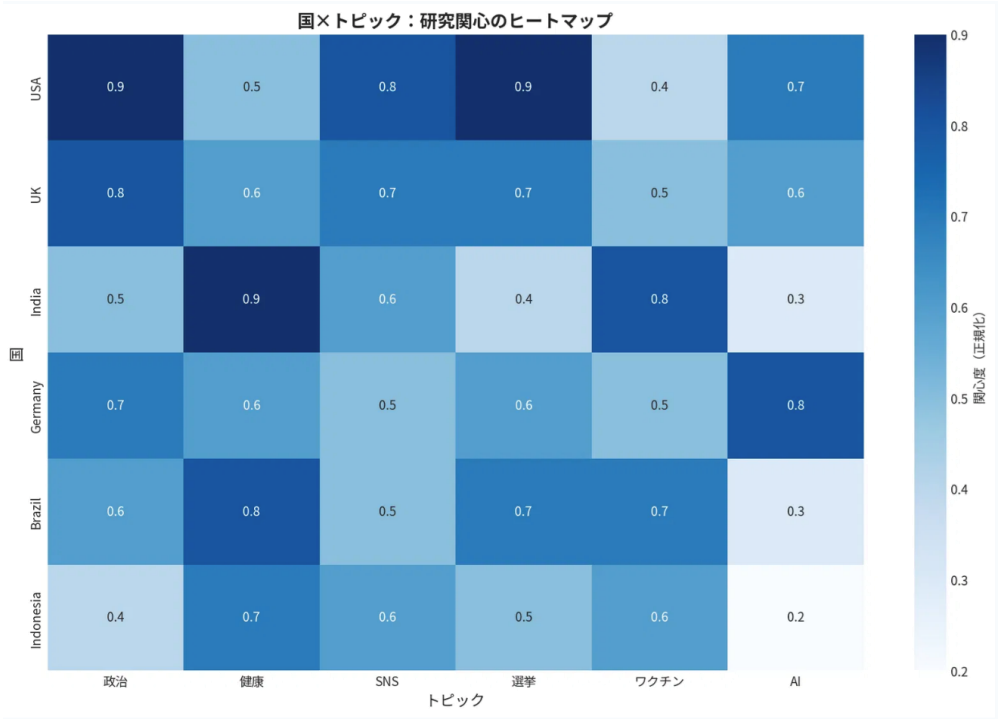




3-2. 研究の個別詳細

国別の研究関心の違い

- 論文執筆者が所属する機関の国と扱っているテーマのヒートマップを作成した。「相対的な強弱」を見るため、割合ではなくMin-Max正規化をして作成した（二重カウントあり）。
- 先進国（USA/UK）の場合
  - 米国や英国のヒートマップは、「政治」や「選挙」で極めて高いスコア（0.8-0.9）を示している。
  - 政治的分極化への関心の集中
    - これらの国々では、ソーシャルメディアがいかに政治的分極化を加速させ、民主主義制度を揺るがすかが主要な問いとなっている。



- 新興国（インド/インドネシア）
  - インドやブラジルのヒートマップは対照的で、「健康」や「ワクチン」への関心が突出（0.8-0.9）。
- 国や地域によって研究関心は大きく異なる。
  - 新興国ではより健康にダイレクトに関わる問題として偽情報・誤情報が研究上扱われており、欧米と扱われているテーマ・文脈が違う部分があると推測される。様々な国・地域の論文を参照することで、英語圏の論文を相対化できる可能性が示唆される。

目次

4-1. 有効性等に関する検証の全体像

有効性等に関する検証に係る取組・成果の全体像

■ 検証のアプローチ：シリコンサンプリングによる多角的な比較検証

- 従来の実験に伴う倫理的課題やバイアスを回避するため、World Values Survey Wave7データを用いたLLMエージェントによる社会シミュレーション（シリコンサンプリング）を採用。
- 手法の特性を確認するため、性質の異なる「2つのシミュレーション手法（個別ペルソナ法 vs 集約統計法）」と「2つのLLMモデル（GPT-4 vs Gemini）」を設定し、計96条件（2つのシミュレーション手法×2つのLLMモデル×6つの介入方法×4種類のペルソナ入力情報）にて差異を検証した。

■ モデルおよび手法の違いによるシミュレーション結果の差異

- モデル間の差異：Gemini 1.5 Proは介入に対して鋭敏かつ極端に反応する傾向が見られた一方、GPT-4 Turboは比較的安定的で緩やかな変化を示し、使用モデルによって結果の感度が異なることが確認された 。
- 手法間の差異：集約統計を用いる手法（C）では効果が平準化された一方、詳細な個別ペルソナを用いる手法（B）では、「正確さ想起」における「逆効果（Backfire）」のような複雑な心理変容を明確に検出できた。

■ ペルソナ情報の違いによる反応の差異

- 「社会規範」介入は入力されたペルソナ情報に依らず一定の抑制効果を示したが、「正確さ想起」介入はペルソナによって結果が大きく異なり、与えるペルソナ情報がシミュレーション結果に強い影響を与えることが示された。

4-2. 有効性等に関する検証の個別詳細

手法紹介：シリコンサンプリングとは

従来の課題

- 実際のSNSユーザーに対し、意図的に偽ニュースを流して反応を見る実験は、倫理的に必ずしも望ましいこととは限らない
- アンケート調査では、「自分は騙されない」というバイアスがかかり、実際の共有行動（拡散）を正確に予測できない可能性がある

解決策：シリコンサンプリング（Silicon Sampling）

- 大規模言語モデル（LLM）によって構築されたAIエージェントに、人間の社会属性や人格（ペルソナ）を付与し、彼らが社会的事象に対してどう反応するかをシミュレーションする計算社会科学の手法である。
- メリット
  - 倫理的安全性:実在の人間を危険に晒すことなく、偽情報の曝露実験が可能
  - 再現性と制御性:全く同じ条件下で何度でも実験を繰り返すことができ、介入効果の微細な差異を厳密に測定できる
  - 「日本社会」の再現: 日本人の人口統計データや価値観を反映したエージェントを生成することで、日本独自の文脈での反応を予測する。今回は、世界価値観調査（WVS） Wave 7の日本データを利用（N=1353）
    - 「日本独自の文脈」とは：欧米と比較して相対的に高い伝統メディアや制度への信頼、集団規範への同調性の高さ、そして経済や安全保障に対する特有の不安・安全志向など、日本人の情報受容や意思決定に影響を与える内面的な価値観特性

今回の実験の位置付け

- 今回の実験においては正解データがないため、この手法がどの程度妥当なのかどうか検証することはできないので、あくまで手法の提案と実行例を提示することを目的とする。
- 異なるLLMモデルやデータ入力方法で実行し、それらの違いを見ることも目的とする。

4-2. 有効性等に関する検証の個別詳細

今回実行した２つのシリコンサンプリング手法と使用したLLMモデル

本実験では、結果の比較のため、性質の異なる2つのシミュレーション手法を比較検証した。

- **手法B：個別ペルソナ法 (Individual Persona)**
  - WVSのデータをもとに「40代男性、都市部在住、メディア不信傾向あり」といった詳細なペルソナを持つエージェントを1体ずつ生成し、個別に偽情報を見せて反応を観察する 。
  - 強み: エージェントが自身の文脈（「自分は以前メディアに騙されたから、今回も疑う」等）に基づいて判断するため、人間の複雑な心理メカニズムや、「逆効果（Backfire）」のような予期せぬ反応を検出するのに優れている 。個人の心理変容を捉えるための深層分析に使いやすい。
  - 1つのペルソナに対して1つの回答だけをさせるというわけではなく、どれぐらいの確率で当該ペルソナがそれぞれの選択肢について選択するのか確率を提示させ、それを全てのペルソナ（＝回答個票）に対して1回ずつ実行し最後に合算して全体の分布を出した（注：手法Aは1つのペルソナに対して1つの回答だけをさせるものだが、正解データがある予備実験において手法Bや手法Cより誤差が大幅に大きかったため今回は採用しなかった）。
- **手法C：集約統計法 (Aggregate Statistics)**
  - WVSのデータをもとに、個別のエージェントを作らず、集団全体の統計分布（例：「この集団の30%は保守的である」）をLLMに入力し、集団全体の反応分布を一括で予測させる 。
  - 計算コストが低く、マクロな傾向（全体として増えるか減るか）を素早く把握できる。一方、個人の文脈が捨象されるため、今回のような「個人の不安に起因する逆効果」を見落とすリスクがある（実際に本実験でも逆効果を検出できなかった）。
- **使用したLLMモデル（いずれもAPIで呼び出し）**
  - GPT-4 Turbo
  - Gemini 1.5 Pro

4-2. 有効性等に関する検証の個別詳細

介入条件、ペルソナ条件、今回の実験条件の構成のまとめ

介入条件

- Control（統制群）：介入なし。ベースライン。
- Accuracy Nudge（正確さ想起）：投稿をシェアする前に「この情報の正確さを判断してください」と一言尋ねる手法。英語圏ではしばしば有効とされる手法。
- Norm（社会規範）：「多くのユーザーが、情報の正確さを重視しています」というメッセージを表示。「みんながそうしている」という同調心理（バンドワゴン効果）に働きかける 。
- Prebunking（情報の予防接種）：偽情報を見る前に、「感情を煽る投稿には注意しましょう」と、偽情報の手口そのものを警告・教育し、精神的な免疫をつけさせる 。
- Friction（摩擦）：シェアボタンを押した後に「本当にシェアしますか？」と確認画面を出し、物理的な手間（摩擦）を増やして再考を促す 。
- Fact Check（事実確認）：投稿の下に、専門家による訂正情報を付記する。

ペルソナ条件

- WVSのデータ全てを入力することはコンテクスト量の制限やLLMがどの情報を参照すればいいか分からなくなる可能性などを考えると望ましくない。
- ４つのパターンで異なる情報のセット（手法Bの場合はペルソナ情報、手法Cの場合は全体の分布の情報）をLLMに与えて比較した。

実験条件の構成

要因	レベル数	具体例
シミュレーション手法	2	Method B, C
LLMモデル	2	GPT-4, Gemini
介入条件	6	Control, Accuracy Nudge, Norm, Prebunking, Friction, Fact Check
ペルソナ	4	Demographic, Trust Media, Ideology Values, Anxiety Security
総条件数	**2 × 2 × 6 × 4 = 96条件**	

4-2. 有効性等に関する検証の個別詳細

ペルソナ条件の詳細

WVS Wave 7の日本データ（N=1,353）には多数の設問項目が含まれるが、全てを一度にLLMへ入力することはコンテクスト長の制約やモデルの注意分散を招くため望ましくない。そこで、以下の4つの観点から設問項目を選択的に抽出し、それぞれ異なる情報セットとしてLLMに入力した。

① 人口統計（Demographic）
年齢、性別、学歴、居住地域（都市部／地方）、職業、世帯収入など、回答者の基本的な社会属性に関する項目。ペルソナの「どのような人物か」を規定するベースライン情報として使用。

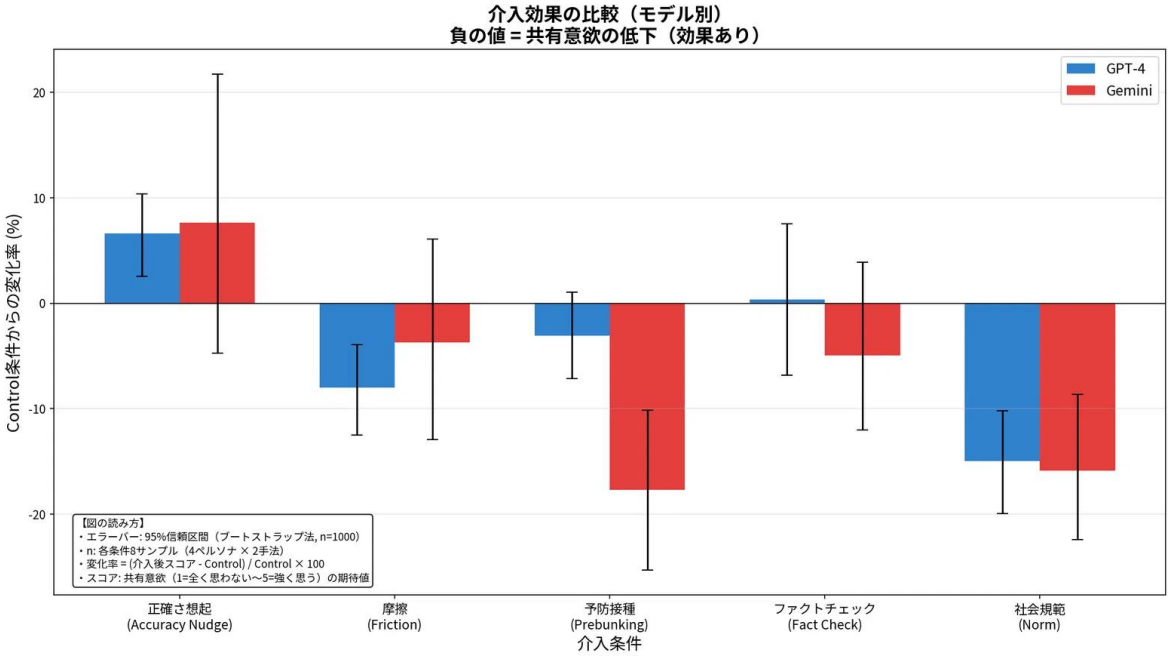
② メディア信頼（Trust_Media）
新聞・テレビ・SNS等の各メディアに対する信頼度、情報入手時に主に利用するメディアの種類に関する項目。メディアへの信頼が高い層と低い層で偽情報への反応が異なるかを検証するためのペルソナのセット。

③ 価値観（Ideology_Values）
政治的立場（リベラル／保守の自己認識）、民主主義への満足度、政治参加の意欲、伝統的価値観と世俗的価値観のバランスなどに関する項目。政治的イデオロギーや価値観が偽情報の共有意欲に与える影響を分析するためのペルソナのセット。

④ 不安・安全（Anxiety_Security）
経済的不安、雇用への懸念、治安に対する認識、社会全般への信頼度（対人信頼）などに関する項目。不安感や安全志向が強い層ほど偽情報に脆弱であるという仮説を検証するためのペルソナのセット。

4-2. 有効性等に関する検証の個別詳細

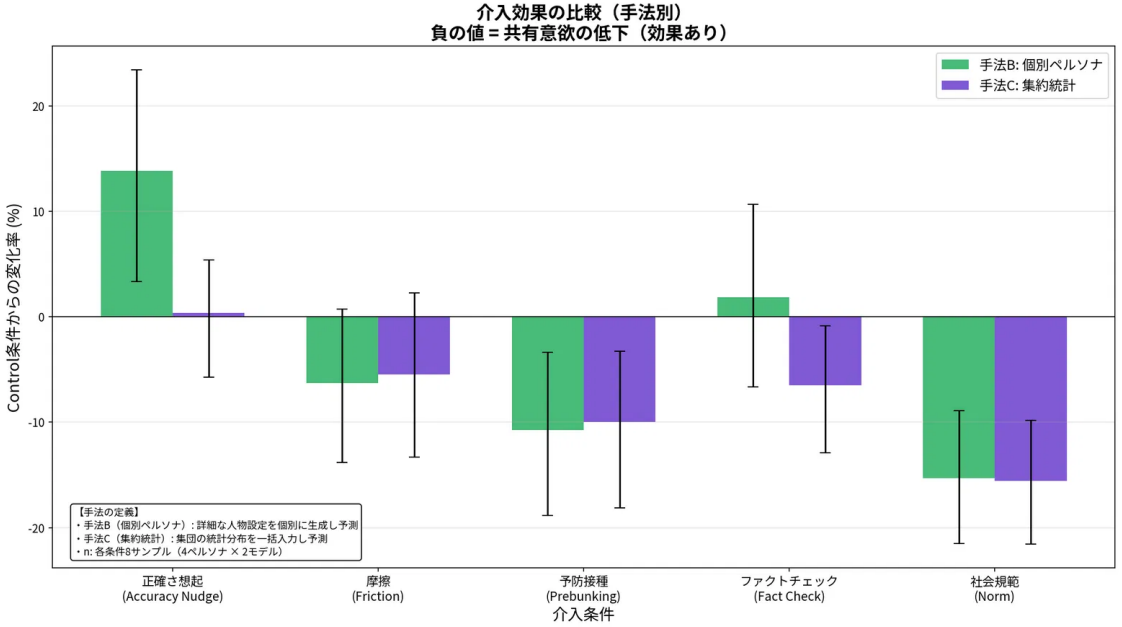
実験結果



■ グラフの見方（縦軸：介入による共有意欲の変化率）
ゼロより下（マイナス）： 共有意欲が下がり「対策として効果あり」を示す
ゼロより上（プラス）： 共有意欲が上がり「対策が逆効果」になったことを示す
■ この結果から言えること
使用するAIモデルによって、対策効果の感度にばらつきあり
Gemini 1.5 Proは介入に敏感に反応し、極端な値が出やすい傾向がある（例：予防接種の効果 -17.7%）
GPT-4 Turboは変化が比較的緩やかで、安定した数値を示す（例：同 -3.1%）。
単一モデルでの検証は特有のバイアスリスクを伴うため、
確実な傾向把握には複数モデルを用いた「クロスバリデーション（交差検証）」が不可欠

4-2. 有効性等に関する検証の個別詳細

実験結果

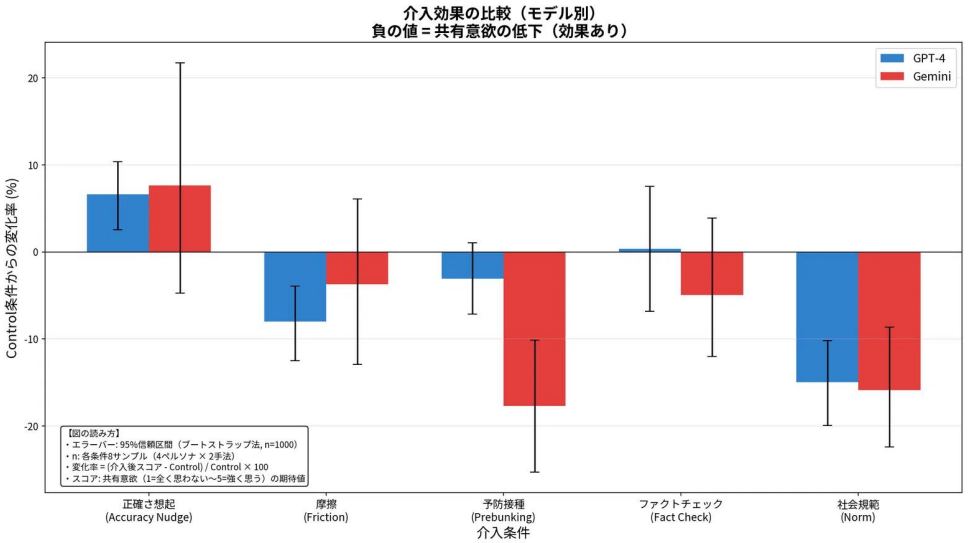


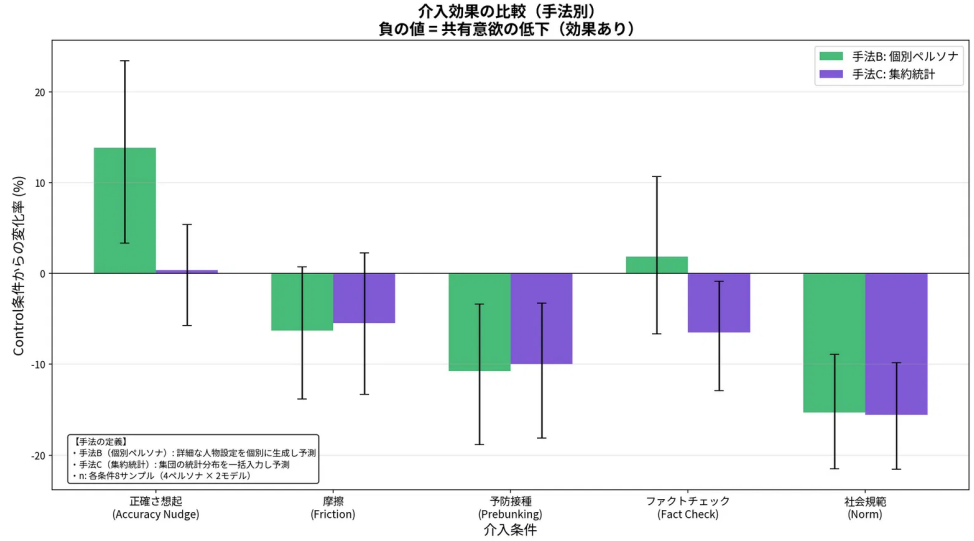
■ グラフの見方（縦軸：介入による共有意欲の変化率）
ゼロより下（マイナス）：「効果あり」、**上（プラス）：**「逆効果」を示す
緑色（手法B）： 詳細な人物設定を持つ「個別ペルソナ手法」
紫色（手法C）： 集団分布を一括処理する「集約統計手法」で
■ この結果から言えること「正確さを考えて」と促す介入において、2つの手法間で予測結果に大きな差が発生
集約統計（手法C/紫）では、共有意欲にほぼ変化なし（+0.4%）。
一方、個別ペルソナ（手法B/緑）では、介入の文脈を深く解釈し、強い逆効果（+13.9%）を明確に検出
逆効果のような複雑な心理メカニズムを的確に捉えるには、
計算コストをかけてでも「個別ペルソナ手法（手法B）」を採用する必要性があることを示唆

4-2. 有効性等に関する検証の個別詳細

モデルと手法による比較

- モデル間の感度差とバイアス（上図）
  - Gemini 1.5 Proは介入に対して敏感に反応する傾向があり（例：Prebunking効果は -17.7%）、GPT-4 Turboは比較的安定的で緩やかな変化（同 -3.1%）を示した。
  - この結果から、単一モデルでの検証には特有のバイアスリスクが伴うことが判明し、確実な傾向把握には複数モデルを用いたクロスバリデーションが不可欠であると言える。
- 手法による検出精度の差（下図）
  - 詳細な人物設定を持つ「手法B（個別ペルソナ）」では、「正確さ想起（Accuracy Nudge）」における強い逆効果（+13.9%）を示した。
  - 一方、集団分布を一括処理する「手法C（集約統計）」では逆効果が示されなかった。
  - 逆効果のような複雑な心理メカニズムを捉えるには、計算コストをかけてでも個別ペルソナ手法（手法B）の採用必要性があることを示唆。

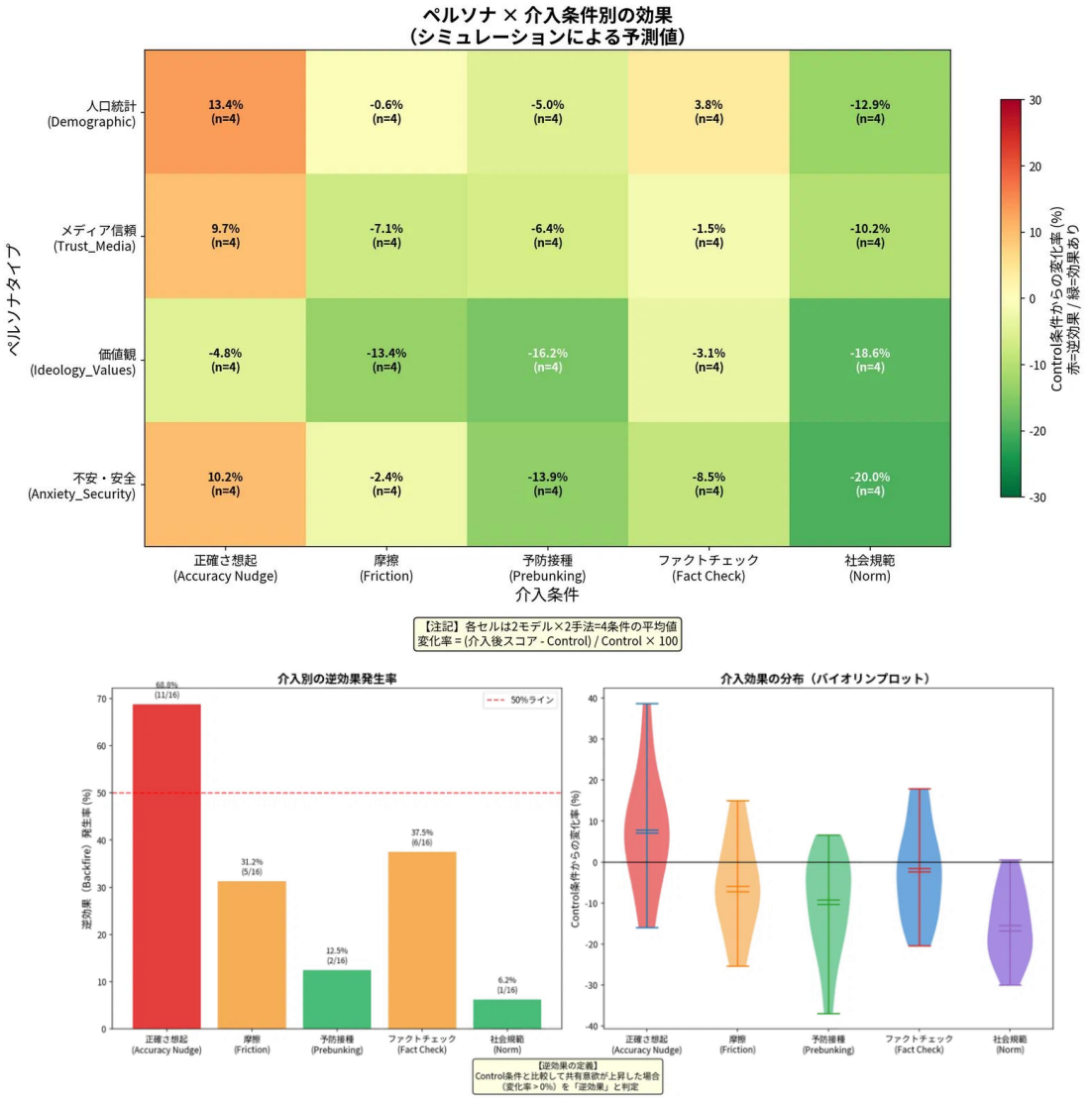




4-2. 有効性等に関する検証の個別詳細

ペルソナ情報による比較

- どのようなペルソナ情報(今回は4種類)を与えるかによってもシミュレーション結果は異なる。
- 社会規範の効果が高いという方向性は一致していた（入力情報に依らず日本らしさが一定程度再現できている可能性）。
- 一方、正確さ想起のシミュレーション結果はペルソナによって異なる結果となった（ただし右下図のように逆効果を示せている条件は多い。



4-2. 有効性等に関する検証の個別詳細

（参考）シミュレーション結果に基づく介入効果の詳細

■ 効果的と考えられる介入（モデル・手法を問わず一貫した抑制効果）

① 社会規範（Norm）：平均効果 -15.4%「多くのユーザーが情報の正確さを重視している」という規範的メッセージの提示。全4種類のペルソナに対して一貫して最も高い抑制効果を示し、モデル間・手法間のばらつきも小さかった。日本社会における同調傾向の強さと整合的な結果と解釈できる。

② 予防接種（Prebunking）：平均効果 -10.4% 偽情報を目にする前に、感情を煽る手口等を事前に警告・教育する手法。特に価値観（Ideology_Values）ペルソナで -16.2%、不安・安全（Anxiety_Security）ペルソナで -13.9% と、イデオロギーや不安に基づく拡散動機を持つ層に対して高い効果を示した。

■ 一定の効果が見られる介入

③ 摩擦（Friction）：平均効果 -5.9% 共有前に確認画面を挟む手法。大きな効果ではないが、安定して共有意欲を低下させた。

④ ファクトチェック（Fact Check）：平均効果 -2.3% 専門家による訂正情報の付記。抑制効果は確認されたが、他の介入と比較して効果量は小さかった。

■ 逆効果のリスクがある介入

⑤ 正確さ想起（Accuracy Nudge）：平均効果 +7.1%（逆効果）「この情報の正確さを判断してください」と促す手法。海外研究では -5～-15% の抑制効果が報告されているが、本シミュレーションでは全16条件中11条件（68.8%）で共有意欲がむしろ上昇した。特に人口統計（Demographic）ペルソナで +13.4%、不安・安全（Anxiety_Security）ペルソナで +10.2% と、逆効果が顕著であった。「正確さを確認せよ」という指示が「重要な情報である」というメタメッセージとして作用した可能性が考えられる。

■ 結果の信頼度に関する留意事項

社会規範（Norm）と予防接種（Prebunking）は、モデル・手法・ペルソナの違いを超えて一貫した方向性を示しており、シミュレーション上の結論としては相対的に信頼度が高い。正確さ想起（Accuracy Nudge）の逆効果についても、多くの条件で再現されたが、LLMの自己説得（ハルシネーションを含む根拠生成）による技術的アーティファクトの可能性も排除できないため、人間を対象とした実証実験による検証が不可欠である。

目次

5-1. 普及啓発活動の全体像

イベントの実施

- 概要：2026年2月17日12:00～13:00
- 実施方法：Zoomウェビナーを利用したオンラインセミナー
- 内容
  - 偽情報・誤情報研究の現在と今後のあるべき姿について、以下を踏まえて議論
    - 偽情報・誤情報研究のこれまでの世界的動向
    - 代表的な介入方法（対策）のシリコン・サンプリングによる検証結果
- 登壇者：
  - 石井大智（東京科学大学技術経営専門職学位課程）
  - 藤代裕之（法政大学社会学部メディア社会学科教授）
- 参加者：
  - 当社ウェブサイトやメールマガジンなどが募集。50人ほどが参加した。

5-2. 普及啓発活動の個別詳細

イベントでのポスター発表

- 展示イベント：「インターネット上の偽・誤情報等への対策技術の開発・実証事業成果発信イベント」
- 開催日時：令和8年3月16日(月) 13:00～17:00
- 主催：総務省
- 開催場所：大手町サンケイプラザ 4階ホール
- 展示内容：今回の分析結果をまとめたポスターを展示

目次

6-1. 研究・調査の総合的な考察

研究・調査の総合的な考察

全体のまとめ

- 本研究・調査は、①グローバルな誤情報研究動向の俯瞰と②日本文脈での有効性検証に対する新たな手法の提示をしたものである。
- ①については、OpenAlex収録文献（約5.2万件）を用いて、誤情報研究が過去15年で爆発的に拡大し、社会的危機（選挙・パンデミック・生成AI）が研究を牽引する「イベント駆動型」であることを示した。これにより、対策技術の評価軸は固定的ではなく、脅威の形（政治・公衆衛生・AI生成等）に応じて更新されるべきである、という前提が明確になった。
- ②についてはは、人間被験者を偽情報に曝露しづらいという倫理的制約や、アンケート回答の社会的望ましさバイアスを踏まえ、LLMエージェントによる社会シミュレーション（シリコンサンプリング）を採用した。これにより「日本社会の反応」を安全に、かつ反復可能に比較検証する実験場の方向性を提示できた。
- 総括すると、（1）誤情報研究知見の偏り（地理・言語・プラットフォーム）を可視化し、（2）誤情報対策手法の日本への適合性を検証する枠組みを用意し、（3）誤情報への関心の持ち方や介入手法の効果が地域・文脈・モデルで変動しうることを示した点に、本研究の中核的意義がある。

6-1. 研究・調査の総合的な考察

研究・調査の総合的な考察

① グローバルな研究動向の俯瞰について

- 文献数は2010年から2024年にかけて大幅に増加し、研究関心が「政治的危機」「公衆衛生危機」「生成AIの台頭」といった社会イベントに同期して跳ね上がる構造が確認された。これは、偽・誤情報対策が平時の啓発だけでなく、危機局面での迅速な技術更新・実装を要する領域であることを示唆する。
- 用語面でも、「Fake News」から、より包括的な「Misinformation」へ比重が移るなど、概念枠組みが変化している。対策技術の比較評価は、用語の流行や政治的含意にも影響され得るため、研究評価の前提（定義・分類）を定期的に点検する必要がある。
- 研究対象プラットフォームは歴史的にTwitter/X偏重だったが、API環境変化等により研究比重が変化している。YouTube/TikTok、暗号化チャット等、影響力が大きい一方でデータ取得が難しい領域が研究の盲点になりやすい点が、研究地図から浮き彫りになった。
- 国別の関心も一様ではなく、英語圏（英米）は政治・選挙、グローバル・サウスでは健康・ワクチン等に強い関心が見られる。従来の英語圏中心知見をそのまま標準解として適用すると、英語圏にはない「日本らしい」リスクを取り落とす可能性がある。今回の研究を参照すると例えば以下の2点が挙げられる。
  - 第一に、プラットフォーム利用実態の違いである。前半の文献分析で示した通り、グローバルな誤情報研究はTwitter/Xに偏重している。一方、日本ではLINEの利用率が突出して高く（総務省「情報通信メディアの利用時間と情報行動に関する調査」）、WhatsApp同様のクローズドなメッセージングアプリが情報流通の主要経路となっている。こうした閉鎖的プラットフォームは、文献分析で指摘した「データ取得が困難で研究が手薄な領域」に該当し、日本における偽・誤情報拡散の実態が十分に把握されていない恐れがある。
  - 第二に、社会心理的特性の違いである。後半のシミュレーション結果では、社会規範（Norm）介入が日本の文脈で特に高い抑制効果を示した一方、欧米で有効とされる正確さ想起（Accuracy Nudge）は逆効果となった。これらは、集団の規範に同調しやすい傾向や、指示を真摯に受け止める傾向といった、日本社会に特徴的な社会心理的要因が介入効果に影響している可能性を示唆するものである。欧米で開発・検証された介入手法を日本に導入する際は、こうした社会心理的差異を考慮した調整が不可欠である。
- ゆえに、何が研究され、何が研究されていないか（Missing Link）を特定し、国内課題に直結する研究へとつなぐことが重要である。

6-1. 研究・調査の総合的な考察

研究・調査の総合的な考察

② 日本文脈での有効性検証に対する新たな手法の提示について

- シリコンサンプリングでは、WVS（世界価値観調査）日本データ等を用い、複数のシミュレーション手法（個別ペルソナ法／集約統計法）と複数LLM（GPT-4 Turbo／Gemini 1.5 Pro）を掛け合わせ、介入手法を比較した。ここで重要なのは、介入効果が「手法」「モデル」「与えるペルソナ情報」によって変動し、単一条件の結果を一般化することにはリスクがある点である。
- 実際に、モデル間で介入への感度が異なり、特定モデルでは効果が過大・過敏に出る可能性が示された。したがって、実務で活用する際は、複数モデルでのクロスバリデーションを前提に「頑健に効く介入」と「条件依存で揺れる介入」を切り分ける設計が求められる。
- 介入別には、「社会規範（Norm）」がペルソナ情報の違いに対して比較的一貫した抑制効果を示す一方、「正確さ想起（Accuracy Nudge）」は条件によって逆効果（Backfire）を取り得ることが示唆された。さらに、この逆効果は個別ペルソナを丁寧に扱う手法でより明確に検出され、集約統計的な手法では見落とされ得る。平均すると効いて見えるが特定層では悪化するという、現実の政策実装で最も避けたい失敗パターンを、検証手法の設計次第で拾える／拾えない可能性が示唆される。
- この新たな手法の提示は「日本社会にそのまま輸入してよい介入」と「副作用を評価しながら調整すべき介入」を峻別するための出発点を提供したと言える。

6-2. 研究・調査にあたっての課題・展望

研究・調査にあたっての今後の課題およびそれらを踏まえた今後の展望

研究動向をウォッチする上での課題

- データソース・抽出条件の限界
  - OpenAlex＋タイトル中キーワード中心の抽出は、網羅性と再現性を担保しやすい一方、
  - ①非英語圏・ローカル語研究の取りこぼし、②領域固有語（例：プロパガンダ/詐欺/陰謀論等）で語られる研究の取りこぼしが起こり得る。
- 概念・用語の変遷が分析結果に影響
  - Fake News→Misinformationのように用語が動くため、年次比較は「実態の変化」だけでなく「呼称の流行」も混在する。
  - 定義（分類体系）の定期アップデートが必要。
- 「プラットフォーム研究」の測り方の限界
  - タイトル/抄録のプラットフォーム名出現だけでは、実際に当該PFデータを使った研究か、政策/評論的論文かを区別しきれない。
  - また、クローズドSNS（LINE等）や暗号化チャットは研究上もデータ取得上もになりやすい。

→多言語で網羅的な研究ウォッチが次のステップ。以下の手法が考えられる。

- **LLMによる自動翻訳・要約パイプライン:** 英語以外の論文（中国語、スペイン語等）をLLMで自動翻訳・要約し、言語の壁を超えたスクリーニングを低コストで実施する。
- **ローカルデータベースとの連携:** OpenAlexだけでなく、各地域の固有DB（例：南米のSciELO、中国のCNKI等）やプレプリントサーバー（arXiv等）を統合的に検索対象とする。
- **意味検索（Embedding）の導入:** 単なるキーワード一致ではなく、ベクトル検索を用いて「表現は異なるが意味が近い研究（概念）」を漏らさず抽出する。

6-2. 研究・調査にあたっての課題・展望

研究・調査にあたっての今後の課題およびそれらを踏まえた今後の展望

LLMを活用したシリコンサンプリングの技術的課題

- 妥当性（正解データ）検証が未解決
  - シリコンサンプリングは倫理面・再現性で強い一方、現時点では「どの程度現実を当てているか」を評価しづらい。
  - 比較のための人間データ （ベンチマーク）が必要
- モデル依存・手法依存（感度差）の大きさ
  - LLM（例：GPT系とGemini系）や、個別ペルソナ法 vs 集約統計法で効果推定が変動し、単一条件の結果を一般化すると誤るリスク（特に逆効果の見落とし）。
- ペルソナ設計の難しさ（どの情報を与えるか問題）
  - WVS等の情報をどこまで与えるかで推定が揺れる。「日本らしさ」を再現できているのか、どの変数が効いているのかを分解する必要。
- アウトカムの現実性
  - シェア意図/判断だけだと、実環境の ネットワーク効果（同調・炎上）、反復曝露、動画/画像の影響を十分に表現しきれない。

→現実との比較によって正しさがある程度保証された手法のもと、
以下のような手法でより現実の動きに合わせたシミュレーションを実現させることが次のステップ

- **マルチエージェント・シミュレーション（MAS）への拡張:** エージェント単体の反応だけでなく、エージェント同士が相互作用（シェア、リプライ、議論）するSNS空間を構築し、ネットワーク効果（エコーチェンバー等）を再現。
- **人間データによるパラメータ校正（Calibration）:** 小規模な人間対象実験（アンケートや行動実験）の結果を「正解データ」として用い、シミュレーションの予測精度が高まるようモデルのパラメータを調整。
- **過去事例の再現（Historical Validation）:** 実際に過去に起きた偽情報拡散事例をシミュレーション上で再現できるかテストし、妥当性を検証。

実社会での利活用を視野に入れた出口戦略（社会実装の展望）

- 上記で確立した高精度なシミュレーション技術は、学術的な研究にとどまらず、実社会の課題を解決するソフトウェアやシステムへの組み込みなど、具体的な出口を見据えている。
- 偽情報拡散予測システムへの組み込み: プラットフォーム事業者や行政・公的機関向けに、特定の情報が「どの属性の層に」「どのような速度と規模で」拡散するかをリアルタイムに予測するエンジンとして提供する。これにより、ファクトチェック情報の発信や注意喚起を行う際の「最適なタイミング」や「効果的なターゲット層」を事前に検証・策定することが可能となる。
- 情報防災・メディアリテラシー教育ツールとしての展開: 現実と同等の反応を示すAIエージェントの集団を活用し、一般ユーザーや企業のSNS担当者が、安全な環境（サンドボックス）で情報拡散のメカニズムや炎上発生時の対応プロセスを体験的に学ぶためのトレーニングツールとして活用する。

目次

7-1. 実施体制及び役割分担

本事業の実施体制図

新領域安全保障研究所
（研究・調査主体）

各団体の役割・業務範囲

- 全ての役務を新領域安全保障研究所のみで実施した。
- 再委託先なし

7-2. 全体スケジュール

主な実施事項	令和7年						令和8年	
	8月	9月	10月	11月	12月	1月	2月	3月
（1）インターネット上の偽・誤情報等への対策技術に係る研究の実施								
・論文データの収集	→	→	→					
・分析の実施		→	→	→				
（2）インターネット上の偽・誤情報等への有効性等に関する検証								
・LLMエージェントの構築		→	→	→				
・シミュレーション実験の実施			→	→	→			
・実験データ分析				→	→	→		
（3）成果報告書の作成								
・中間ドラフト・内部レビュー					→	→		
・最終版提出・修正対応						→	→	→
（4）普及啓発活動への協力								
・ウェビナー企画・開催						→	→	→
・記事企画・発信						→	→	→