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 概要版

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・調査によりアプローチする課題・目指す姿

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
  - 誤情報研究は世界的に急増しているが、その多くは英語圏のデータや文脈に偏っており、非英語圏やグローバル・サウスの実態、および「何が研究されていないか」が可視化されていない。
  - その結果、対策投資が飽和領域に偏り、真にリスクが高い未解明領域への対応が遅れる恐れがある。
- 課題② 海外対策技術の「日本社会への適合性」の不透明さ
  - 欧米で有効とされる手法が、日本において同様に機能するか不明確であり、「逆効果（Backfire Effect）」を生むリスクがある。海外ベストプラクティスの無批判な導入は、安全性・実効性の両面で課題となり得る。
- 課題③ 実証実験における「倫理的・実務的制約」
  - 実SNSユーザーへの偽情報曝露実験は倫理的ハードルが高く、アンケート調査ではバイアスがかかりやすい。
  - 安全かつ再現性のある検証環境（シミュレーション）の確立が求められている。

上記課題を踏まえ目指す姿・ゴール

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。

- 中間ゴール①：研究動向の可視化
  - OpenAlexを用いた包括的分析により、地理・言語・プラットフォームによる研究の傾向を定量化・可視化し、現在の研究動向を理解する。
- 中間ゴール②：「シリコンサンプリング」による次世代検証モデルの確立
  - LLMエージェントを用いた社会シミュレーション手法を構築し、倫理的リスクを回避しながら、日本固有の文脈（ペルソナ）における介入効果を低コスト・短期間で検証できる手法を提案する。

1-2. 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

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。

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 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
  - 英米とグローバルサウスの関心の違いの把握
  - 高速かつプレシジョン（精密）な介入の実装

目次

2-1. 研究および有効性等に関する検証の全体像

研究および有効性等に関する検証の全体像

■ 研究動向分析：社会的危機が牽引する「イベント駆動型」研究の解明

- 規模と成長
  - OpenAlex収録の約5.2万件の文献を分析した結果、2010年から2025年で論文数は21.6倍に急増したことが判明
- イベント駆動型の進展
  - 「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。
- 地域差の可視化
  - 欧米は「政治・選挙」への関心が高い一方、グローバル・サウス（インド・インドネシア等）では「健康・ワクチン」への関心が突出しており、地域による課題認識の差異を明らかにした。

■ シミュレーション：シリコンサンプリングによる検証手法の提示

- WVS（世界価値観調査）の日本データを用い、LLMエージェントによる社会シミュレーションを実施した。「個別ペルソナ法（詳細設定）」vs「集約統計法（全体分布）」、「GPT-4 Turbo」vs「Gemini 1.5 Pro」、ペルソナとして与える情報を変化させシミュレーション結果を比較した 。
- 主要な成果モデルや手法によって効果推定が変動することを確認した。特に詳細なペルソナを用いた手法では、ナッジ介入における「逆効果（Backfire）」現象を示した。

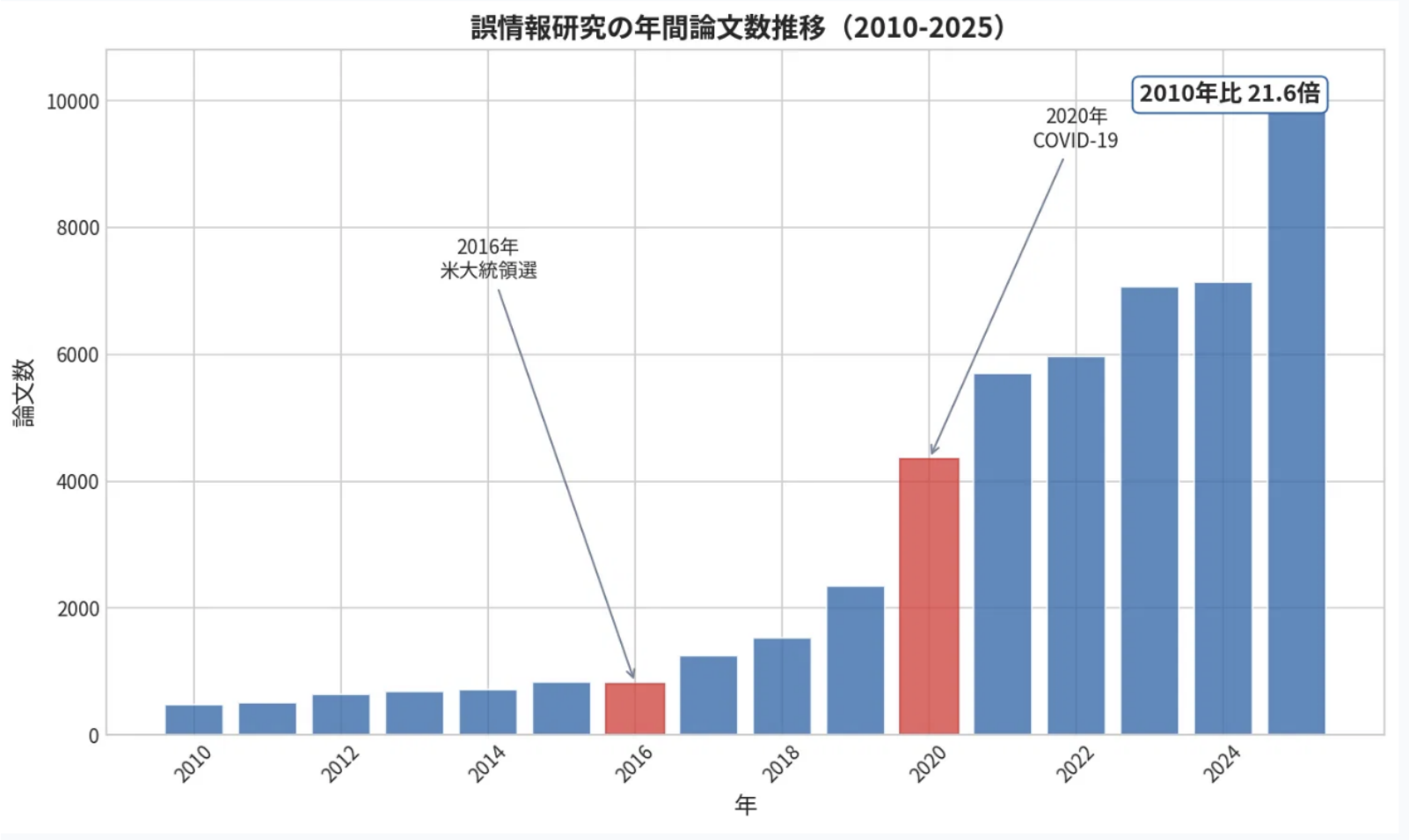
2-2. 研究および有効性等に関する検証の個別詳細

研究動向分析

- 調査方法
  - データ：OpenAlex（学術文献データベース）を使用し、公式APIから収録文献を取得
  - 抽出：misinformation / disinformation / fake news / infodemic をタイトルに含む文献を対象にスクリーニング
  - 対象規模：抽出文献を集計し、年次推移・地域差・テーマ差を把握可能な形に整理（約5.2万件規模）
- 拡大する誤情報研究
  - 過去15年で誤情報研究は急増（2025年は2010年比で約21.6倍、分析対象：約5.2万件）
  - 研究増加は平準的ではなく、社会的危機（イベント）に同期して跳ね上がる「イベント駆動型」の構造を持つ
    - - 2016年：米大統領選（政治・選挙）
    - - 2020年：COVID-19（公衆衛生・ワクチン）
    - - 2023年：生成AI（技術・合成メディア）
- 用語も変化：政治的な意味を有するようになった Fake News を入口にしつつ、より包括的な Misinformation / Disinformation へと広がり
- 地域差（国×トピックの違い）
  - 英米は「政治・選挙・分極化」中心の研究が厚い一方、グローバルサウス（例：インド・ブラジル等）は「健康・ワクチン」等の生活直結型テーマが相対的に主要関心となる。欧米知見だけでは、生活・医療に直結する誤情報リスクを過小評価する恐れがある。
- プラットフォーム偏り（X中心からの変化）
  - 研究対象は長らくTwitter/Xに偏重してきたが、API制限・有償化などの環境変化で、X研究の伸びは鈍化する傾向が見られる。一方で、影響力が大きいにも関わらずデータ取得が比較的困難な動画PF（YouTube/TikTok等）やクローズドチャット（WhatsApp等）への関心が増加。

2-2. 研究および有効性等に関する検証の個別詳細

増加してきた誤情報研究



注：OpenAlexのデータより当社作成

2-2. 研究および有効性等に関する検証の個別詳細

今回採用したシリコンサンプリングの手法について

- 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。

２つの手法

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。

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。

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。

2-2. 研究および有効性等に関する検証の個別詳細

今回採用したシリコンサンプリングの手法について

ペルソナ条件

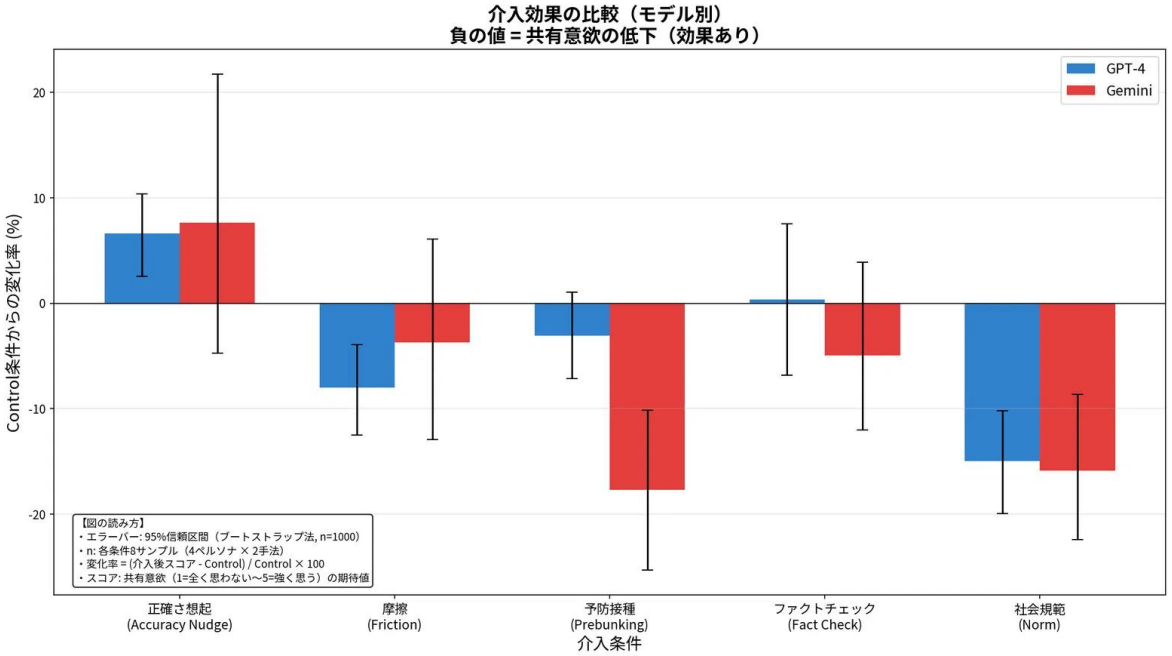
- WVSのデータ全てを入力することはコンテクスト量の制限やLLMがどの情報を参照すればいいか分からなくなる可能性などを考えると望ましくない。
- ４つのパターンで異なる情報のセット（手法Bの場合はペルソナ情報、手法Cの場合は全体の分布の情報）をLLMに与えて比較した。

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**	

2-2. 研究および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%）

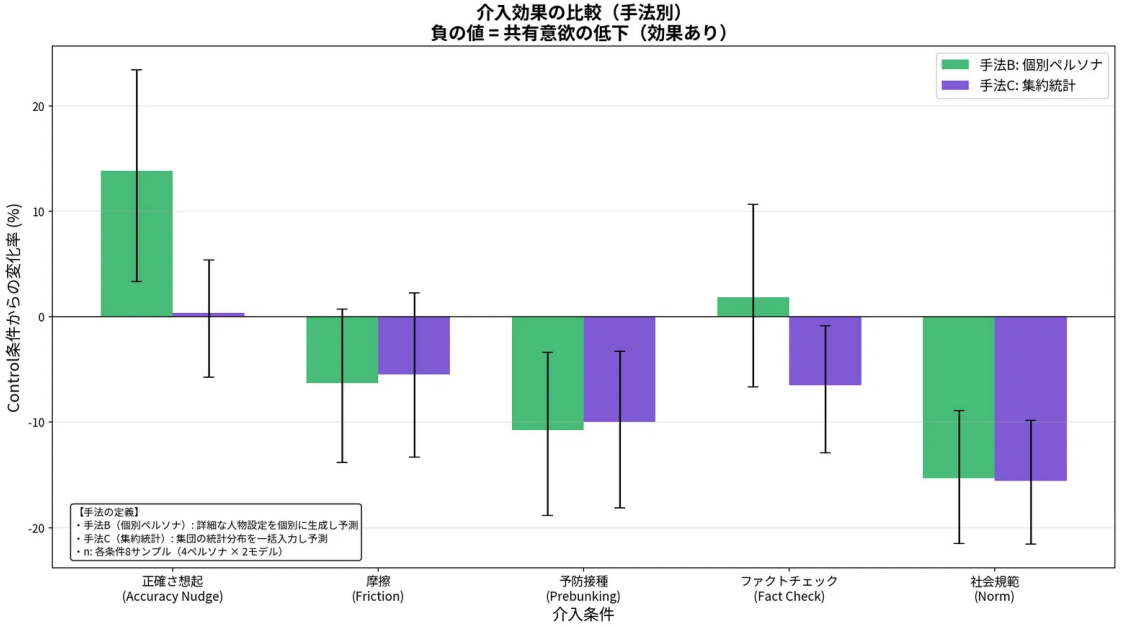
GPT-4 Turboは変化が比較的緩やかで、安定した数値を示す（例：同 -3.1%）。

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

2-2. 研究および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

目次

3-1. 研究・調査の総合的な考察

研究・調査の総合的な考察

- 本研究・調査は、
  ①グローバルな誤情報研究動向の俯瞰
  ②日本文脈での有効性検証に向けた新手法の提示
  を両輪として実施した。
- ①では、誤情報研究が社会的危機（選挙・パンデミック・生成AI等）により加速する「イベント駆動型」であり、対策技術の評価軸は固定ではなく、脅威の形に応じて更新されるべきという前提を明確化した。
- ②では、倫理面・調査バイアスの制約を踏まえ、LLMエージェントを用いた社会シミュレーション（シリコンサンプリング）により、日本社会の反応を安全かつ反復可能に比較検証する実験場の方向性を提示した。
- 総括すると、(1) 研究知見の偏りの可視化、(2) 日本への適合性検証枠組みの整備の方向性を示したことが、
- 本研究の意義である。

3-1. 研究・調査の総合的な考察

研究・調査の総合的な考察

＜研究動向について＞

- OpenAlex収録文献（約5.2万件）に基づき、誤情報研究は過去15年で急拡大し、危機事象が研究を牽引する構造が確認された。
- この構造は、誤情報対策の有効性評価が「単一の正解」になりにくく、政治・公衆衛生・AI生成など脅威タイプ別に、評価軸と優先順位を更新する必要を示唆する。
- 研究知見には、地理・言語・プラットフォームの偏りがあり、歴史的にはTwitter/X偏重だが、API環境変化等で研究比重も変化している。
- 国・地域で関心領域も異なるため、英語圏中心知見の「標準解」をそのまま適用すると、日本で重要なリスクを取り落とす可能性がある。ゆえに、未研究領域を特定し、その上で国内課題に直結する研究へ接続させることがが重要となる。

＜シミュレーション（シリコンサンプリング）について＞

- シリコンサンプリングでは、WVS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）×複数LLM（GPT-4 Turbo／Gemini 1.5 Pro）×異なるペルソナ情報で介入手法を比較した。
- 重要なのは、介入効果が「手法」「モデル」「与えるペルソナ情報」で変動しうるため、単一条件の結果を一般化すると判断リスクが生じる点である。
- 実務適用では、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。
- 今回の場合、介入方法別には、社会規範（Norm）は比較的一貫して抑制効果を示す一方、正確さ想起（Accuracy Nudge）は条件次第で逆効果（Backfire）を取り得ることがシミュレーションで示された。
- また逆効果は、個別ペルソナを丁寧に扱う手法で検出されやすく、集約統計的手法では見落とされ得るため、目的（逆効果検知／平均効果把握）に応じた手法選択が必要である。

3-2. 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

■ 研究動向分析の課題

- 英語・タイトル中心の抽出であるため、非英語圏やローカル言語、領域固有語、タイトルに現れない概念の取りこぼしが残る。
- 実利用プラットフォームの実態（クローズド空間／動画PF等）と、文献上の研究対象のギャップを埋める深層分析が必要。
- 抽出した論文が「議論されているだけ」なのか「実際に観測・介入まで行っているか」など、研究の成熟度を区別する整理も今後の課題。

■ シリコンサンプリングの課題

- 現実の人間データ（ベンチマーク）との突合・校正が不足しており、どれだけ現実を当てているかの直接評価ができていない。
- 推定結果がLLMモデルやプロンプト設計、ペルソナの与え方に依存しうるため、安定的な評価枠組み（再現性・頑健性）の整備が必要。
- 現実SNSのダイナミクス（ネットワーク効果、反復曝露、アルゴリズム、動画影響等）を十分に取り込めていない。

上記課題を踏まえた今後の展望

- 多言語・多概念への拡張
  - 非英語圏・ローカルPFを含む網羅的な研究ウォッチへ発展させ、地理・言語・媒体の盲点を継続監視する。
- 人間データとの比較・校正（ベンチマーク整備）
  - 比較用の人間データ（調査・実験・観測）を整備し、シミュレーションの予測精度を評価・補正できる枠組みを確立する。
  - モデル変更や条件変更時にも品質が追えるよう、共通指標・再現手順を整備する。
- 実環境シミュレーションへの進化
  - ネットワーク上の拡散・反復曝露・動画の影響・コミュニティ分断など、より現実に近いダイナミクスを組み込む。
  - 日本での偽・誤情報対策ガイドライン策定に直結するエビデンスを継続的に創出できる環境を構築する。