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

偽・誤情報の拡散を抑制するためのSNSにおける シェア行動プロセス可視化と信頼性を評価する表示の検討

成果報告書

2026/3/19

研04_東京大学大学院工学系研究科

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	• 偽・誤情報の流通・拡散の防止として、コンテンツモデレーションの手法が用いられているが、表現の自由との兼ね合いで透明性確保が難しいという限界があり、十分な偽・誤情報対策として機能しない可能性がある。このため、偽・誤情報の特徴を的確に捉える精緻な拡散行動メカニズムの解明が求められている。 • 信頼性情報の可視化によって、利用者の自律的な判断を促すメカニズムを解明することにより、偽・誤情報対策の迅速化と質的向上を目指す。		
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）	東京大学　鳥海研究室、(株)Lightblue、(株)電通

研究および有効性等に関する検証の取組・成果

- Chrome拡張機能の開発: Webブラウザ上で動作する拡張機能を開発し、Webページの構造解析（DOM監視）により投稿者情報を特定および、X（旧Twitter）上で、投稿者の過去のコミュニティノート（CN）付与履歴を可視化するツールを実装した。
- 実証実験による効果検証:募集した一般ユーザー38名を対象に、実際のSNS環境を模したシステムを用いて比較実験を実施。介入群（ツール使用）は対照群（ツール使用なし）に比べ、全体のリポスト数を約59%、CN付与投稿へのリポストを約73%抑制した。
- 長期的な行動変容: 追跡調査により、実験直後は高い抑制効果が見られるものの、2週間経過後には「慣れ」により効果が減衰する傾向（習慣化）を確認した。これらが報告書の主張と整合することを検証済み。

指標	対照群（ツール使用なし）	介入群（ツール使用あり）	抑制率
全体リポスト	7.37回	3.00回	**59.3%**
CN付与投稿へのリポスト	2.74回	0.74回	**73.1%**

研究・調査にあたっての課題・展望

- 大規模実証の必要性: 本研究では中程度の実用的な効果量（Cohen's d ≈ 0.5）を確認したが、統計的有意差の確立にはより大きなサンプルサイズが必要である。
- 習慣化への対策: 長期利用における効果減衰を防ぐため、UIの動的変更や、信頼性スコアに応じた介入強度の調整など、持続的な効果維持のための機能改善が求められる。

代表者コメント

東京大学大学院工学系研究科教授
鳥海不二夫

本研究では過去の信頼性の低い投稿の可視化が情報拡散の判断にどのような影響を与えるのかを明らかにした．投稿者が過去にコミュニティノートが付与された事実の可視化が，情報拡散抑制に効果的であることが示された．当研究の結果は偽誤情報対策に大きな貢献が見込まれる．

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

背景：偽・誤情報拡散のメカニズムと認知的要因

1. 認知バイアスの影響

確証バイアス (Confirmation Bias): 利用者は自身の信念と一致する情報を選択的に受容し、反証を無視する傾向がある [1][2]。アルゴリズムによる推薦システムがこれを増幅させ、エコーチェンバー（Echo Chamber）を形成する。

感情的伝染 (Emotional Contagion): 恐怖、怒り、嫌悪などの強い感情を喚起する情報は、中立的な真実の情報よりも拡散速度が速く、より広範囲に伝播する（Vosoughi et al., Science 2018）。

社会的証明 (Social Proof): 多くの「いいね」や「リポスト」が付いていると、それだけで信頼性が高いと誤認してしまう傾向がある。

2. 信頼性判断の困難さ

SNS上の情報流は高速であり、利用者は個々の投稿の真偽を検証する時間的・認知的リソース（Cognitive Budget）を持たない。

発信者の匿名性が高く、「誰が言っているか」という出典の信頼性を判断する手がかりが乏しい。

[1] Nickerson, R. S. (1998). Confirmation bias : A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220.
[2] Del Vicario, M., et al. (2016). The spreading of misinformation online. Proc. Natl. Acad. Sci. U.S.A. 113 (3) 554-559.

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

既存対策の現状と課題

事後的対策アプローチ

- ファクトチェック (Fact-checking): 専門家や認定機関による検証。質は高いが、拡散から検証記事の公開までに数時間～数日を要し、誤情報の拡散スピードに追いつけない。
- プラットフォームによる削除・凍結: 強制力があるが、判断基準の不透明性やオーバーブロッキング（過剰規制）のリスクがあり、表現の自由の観点から慎重な運用が求められる。

コミュニティ主導型対策（Crowdsourced Fact-checking）

- X コミュニティノート: ユーザー同士の相互監視と、「ブリッジングアルゴリズム（Bridging Algorithm）」による党派性を超えた合意形成に基づく評価システム。党派を超えた評価者の同意がないとノートが表示されないため、中立性が担保される。
- 課題: コミュニティノート自体は有効だが、「ノートが付くまでのタイムラグ（数時間～数日）」や「表示されないノート（False Negatives）」の課題がある。

本研究の意義: 本研究は、コミュニティノートの「過去の履歴」という即時利用可能なメタデータを活用することで、このギャップを埋めるものである。

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

目指す姿：「判断プロセス」への介入による自律的防衛

基本思想

- 本事業では、情報の受け手であるユーザー自身をエンパワーメントする技術の開発を目指す。
- 投稿者が過去にどれだけ「信頼できる情報」を提供してきたか、あるいは「ミスリーディングな情報」を発信してきたかを可視化することで、ユーザーのヒューリスティックな判断（直感・感情による判断）を、システマティックな判断（論理・根拠に基づく判断）へと誘導する。

技術的ゴール

1. 即時的な信頼性情報の提供: コンテンツ単位のファクトチェック結果を待たずとも、発信者のレピュテーション（評判スコア）から「注意すべき投稿」をその場で識別できるUIを実現する。

2. 既存UIへのシームレスな統合: ユーザーの通常の利用体験を阻害せず、必要な情報を自然に視野に入れられるデザインを追求する。

3. プライバシーへの配慮: ユーザーの閲覧履歴や行動を不必要に収集せず、匿名化されたデータのみを分析に用いる設計とする。

社会的ゴール

1. 悪意ある拡散への抑止効果 (Deterrence): 「不正確な情報を発信・拡散し続けると、信頼性スコア（可視化されたラベル）が低下する」という社会的インセンティブ構造を作り出し、偽情報拡散のエコシステム自体を健全化する。

2. メディアリテラシーの向上: ツールを日常的に使用することで、ユーザー自身が「発信者の信頼性を確認する」という習慣を身につけ、長期的なリテラシー向上に貢献する。

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

最終的に目指す姿：情報空間の健全化

短期的な目標（本事業期間内）

- 開発したChrome拡張機能の有効性を実証実験により定量的に検証する。
- ユーザーの行動変容（リポスト抑制、信頼性判断の精度向上）を確認する。
- プラットフォーマーや政策立案者に対し、具体的な技術要件と実装指針を提言する。

中長期的なビジョン（事業終了後～5年）

- プラットフォームへの標準搭載：本技術の考え方（発信者の信頼性履歴の可視化）が、X、YouTube、TikTok等の主要プラットフォームに標準機能として実装されることを目指す。
- 国際標準化：信頼性スコアの算出方法やUIガイドラインに関する国際的な標準規格の策定に貢献する。
- AIとの融合：生成AIの回答における信頼性評価にも本技術を応用し、「AIが生成した情報の信頼性」をユーザーが判断できる環境を構築する。

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

直接的な効果

1. 拡散行動の抑制

- 信頼性が低いと判定される発信者の投稿に対し、ユーザーが「一呼吸置く」ことで、衝動的なリポストやシェアが抑制される。
- 結果として、偽・誤情報のウイルス的拡散（Viral Spread）の初期段階を遅延または阻止できる可能性がある。

2. 信頼性判断の精度向上

- ユーザーは、コンテンツの内容だけでなく、「この情報を発信しているのは誰か」という出典の信頼性を意識するようになる。
- これにより、「もっともらしいが虚偽の情報」に騙されるリスクが低減する。

3. 発信者への抑止効果 (Deterrent Effect)

- 信頼性スコアが可視化されることで、意図的に偽情報を発信するアカウントのインセンティブ構造が変化する。
- 「評判が下がる」というリスクを認識することで、投稿前に内容を精査する行動が促される可能性がある。

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

間接的・長期的な効果

1. メディアリテラシーの向上

- 本ツールを日常的に使用することで、「情報源を確認する」「発信者の過去の実績を調べる」といった習慣が自然と身につく。
- 特に、デジタルネイティブ世代への教育効果が期待される。

2. プラットフォーム設計への影響

- 本事業で得られたエビデンスは、プラットフォーマーに対し、ユーザーインターフェース設計の改善を促す政策提言の根拠となる。
- プラットフォームが削除・凍結の件数や政府からの要請内容等を定期的に公開する「透明性レポート」の標準化に加え、個々のアカウント単位での信頼性を可視化する「レピュテーションスコア」の導入が、ユーザーの判断支援に不可欠であることを提言する。

3. 健全な言論空間の形成

- 信頼性の高い情報発信者が評価され、信頼性の低い発信者が可視化されることで、「良質な情報が流通しやすい」エコシステムの構築に寄与する。
- これは、民主主義社会における「熟議」の質を高めることにもつながる。

目次

3-1. 研究の全体像

研究に係る取組・成果の全体像

研究の方法

X（旧Twitter）を対象とし、Chrome拡張機能を用いた実験を実施。参加者には、指定された検索ワードに基づいて抽出された投稿（約200件）を閲覧してもらい、各投稿について「リポストすべきか」「信頼できる情報か」を判断してもらった。

フェーズ	内容	所要時間
① 事前説明・同意取得	研究概要説明、同意書取得	10分
② 実験（Chrome拡張使用）	指定ワードで抽出された投稿を閲覧・評価	30分
③ アンケート調査	属性、SNS利用、政治的関心、実験内容	10分
④ Chrome拡張1ヶ月利用（該当者のみ）	日常利用中の行動ログ取得	1ヶ月
⑤ 事後アンケート・インタビュー（任意）	利用感想、行動変化の自己評価など	15分

期待される成果

- CN履歴のバックグラウンド取得: Webページの構成要素から投稿を検知し、CNデータベースと照合。
- 信頼性指標の提示: Helpful / Not Helpful / Status Badge を表示。
- UIへのシームレスな統合: ユーザー名上にインジケータを追加するデザインを採用。

3-2. 研究の個別詳細

研究の方法

実験の概要

本研究では、X（旧Twitter）を対象とし、Chrome拡張機能を用いた実験を実施した。参加者には、指定された検索ワードに基づいて抽出された投稿を閲覧してもらい、それぞれの投稿について「リポストすべきか」「信頼できる情報か」を判断してもらった。

実験中の行動記録

- 実験中は画面遷移やボタンのログの収集を行い、利用者がどのような情報（例：投稿内容、利用者名、過去の投稿履歴など）を参照して判断しているかを分析。
- また、実験後にはアンケート調査を実施し、SNS利用傾向や政治的関心、情報リテラシーとの関連を明らかにした。
- さらに、一部の参加者にはChrome拡張を1ヶ月間日常的に使用してもらい、自然な利用環境下での行動変容も観察した。

3-2. 研究の個別詳細

実験システムの構成

実験用ロギング環境（Sandbox）

実証実験のために、通常のX利用環境を模しつつ、詳細な行動ログを取得可能なシステムを構築した。

- フロントエンド (Wrapper Application):
  - 実際のXのWebインターフェースと同等の操作感を提供。
  - 信頼性の評価のボタンを追加。
- データ収集バックエンド:
  - 行動ログ収集: クリック、スクロール等のインタラクションを記録。
  - 属性データ管理: 参加者ID（匿名化済み）と、事前アンケートによる属性情報（リテラシー・SNS利用頻度）の紐付け管理。
  - X API連携: 追跡調査用に、参加者のアクティビティ（リポスト、いいね）を収集するパイプライン。

図: 実際の実験の画面

3-2. 研究の個別詳細

実証実験のデザイン

実験参加者

- 募集数: 40名（最終有効データ数: 38名）
- スクリーニング条件:
  - X（Twitter）のアカウントを保有し、日常的に利用していること。
  - 過去3ヶ月以内に10回以上のリポスト（引用リポスト含む）を行っていること（＝情報拡散のアクティブユーザー）。
- 属性: 20代～60代の男女。

実験デザイン：群間比較（Between-Subjects Design）

同一の投稿セットに対し、以下の2条件で閲覧・評価を行わせた。

1. 介入群 (Experimental Group, n=19): Chrome拡張機能が有効。各投稿に投稿者のCN付与履歴（信頼性指標）が表示される。
2. 対照群 (Control Group, n=19): Chrome拡張機能が無効。通常のXと同じ表示（ベースライン）。

3-2. 研究の個別詳細

実験タスクの詳細

閲覧タスク

- 投稿セット: 4つのテーマ（政治、災害、科学、エンタメ）に関連する投稿群。
  - 選定プロセス: 指定された検索ワードに基づいて抽出された投稿群に対し、CN付与歴のあるユーザー（高/低評価含む）とないユーザーを一定比率（約20%以上）で混合。
  - 順序効果の排除 提示順序は参加者ごとにランダム化。
  - 特に偽誤情報が広まった際に世の中に影響が強いと考えられるテーマを選択。
- 評価アクション: 各投稿について以下の判断を求めた。
  i. 拡散判断:「この投稿をリポストしますか？」（はい / いいえ / 迷った）
  ii. 信頼性評価:「この投稿の内容を信頼しますか？」（5段階リッカート尺度）
  iii. 自由記述: 判断の理由（任意）。

実施環境

- 株式会社Lightblueオフィスにて実施。
- 所要時間：約90分（事前説明・同意取得10分、実験30分、アンケート10分、事後インタビュー等）。

3-2. 研究の個別詳細

アンケート設計

事前アンケート

- 個人属性: 年齢、性別、職業、SNS利用頻度。
- 情報リテラシー: 既存の尺度（Hargittai, 2005等）を参考に、メディアリテラシーに関する設問を作成。
- 偽情報への認知: コミュニティノートの存在を知っているか、過去に偽情報に接触した経験があるか等。
- 政治的関心: 左右の政治的スタンスに関する設問（6項目）。

事後アンケート

- 拡張機能のユーザビリティ (SUS: System Usability Scale): 10項目の標準化された尺度を使用。
- ツールの有用性:「信頼性判断に役立ったか」「日常的に使いたいか」等（7段階リッカート尺度）。
- 自由記述: ツールの良かった点、改善点、悪用の可能性についての懸念等。

3-2. 研究の個別詳細

追跡調査（1ヶ月間のChrome拡張利用）

長期的効果と習慣化（Habituation）の検証

実験室環境での一時的な効果（プライミング効果）だけでなく、実生活での行動変容を検証するため、追跡調査を実施した。

実施内容

- 期間: 本実験終了後、1ヶ月間。
- 対象者: 同意を得た参加者の一部（該当者のみ）。
- 方法: 参加者の同意に基づき、X API (v2) を使用して対象アカウントの公開アクティビティを取得。日常利用中の行動ログを取得。
- 分析対象:
  - リポスト頻度: 日次の平均リポスト数。
  - リポスト対象の質: リポストした投稿の元ツイートにCNが付与されているか、またそのCNの評価（Helpful/Not Helpful）。
  - 比較分析: 介入群と対照群における、誤情報拡散リスクの高い投稿へのエンゲージメント率の差分。
- 事後アンケート: 利用感想、行動変化の自己評価など（15分程度）。

※なお、本実験および追跡調査の結果については、「目次4.研究・調査における「有効性等に関する検証」」にまとめて記載。

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

検証の観点

1. 行動変容の有無: 信頼性情報の可視化が、ユーザーのリポスト（拡散）行動に影響を与えるか。
2. 認知への影響: 信頼性情報の可視化が、ユーザーの投稿に対する信頼性評価に影響を与えるか。
3. ユーザー受容性: ツールの使いやすさ（ユーザビリティ）と、日常的な利用意向を評価する。
4. 持続性: 実験室環境ではなく、日常的な利用環境においても効果が持続するか。

主要な定量的成果（分析結果）

分析対象	対照群	介入群	抑制効果	統計的有意性
全体のリポスト数	7.37回	3.00回	**59.3%**	p=0.0952（有意差なし）
CN付与ユーザの投稿へのリポスト	52回	14回	**73.1%**	p=0.1520（有意差なし）
CN付与投稿への信頼性評価	3.12	3.00	–	**p=0.0031（5%有意）**

- リポスト数については、全体・CN付与投稿限定いずれも有意水準5%で統計的に有意な差は確認されなかった。これはサンプルサイズ（n=38）の制約による検出力の不足も一因と考えられる。
- ただし、CN付与ユーザの投稿への信頼性評価においては有意な差（p=0.0031）が確認され、ツールがユーザーの認知に影響を与えていることが示された。
- 効果量（Cohen's d）は約0.5であり 実用的には中程度の効果が観察されている

4-2. 有効性等に関する検証の個別詳細

定量分析結果：拡散行動への影響

仮説検証：信頼性情報の可視化はリポスト（拡散）を抑制するか？

- 分析結果: 実験課題における平均リポスト数を比較。
  - 対照群 (Control): 平均 7.37回 (Std: 12.10)
  - 介入群 (Experimental): 平均 3.00回 (Std: 5.79)
- 統計的検定結果:
  - Mann-Whitney U検定: U = 235.00, p = 0.0952
    （ノンパラメトリック検定の一種で、2つのグループ間に差があるかを統計的に検証する手法）
  - 効果量（Cohen's d）：0.46（小～中程度）
    （統計的な有意差とは別に、その効果の実質的な大きさを示す指標。0.5程度は中程度の効果とされる）

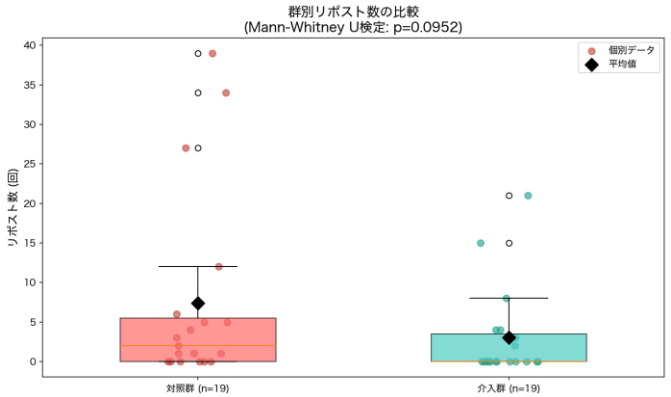


図: 群ごとの実験におけるリポスト数の比較

- 結果の解釈:
  - 介入群のリポスト数は対照群の約41%に留まり（約59%の抑制効果）、実用的には大きな効果が観察された。ただし、有意水準5%では統計的に有意な差は確認されなかった（p=0.0952）。
- ベースラインとの比較:
  - 実験前の日常利用における平均リポスト数は、介入群 (30.8回) の方が対照群 (21.2回) よりも多かった。
  - すなわち、普段はリポスト頻度が高いユーザー群」であるにも関わらず、信頼性情報が可視化された環境下では、リポストを抑制したことが確認された。

4-2. 有効性等に関する検証の個別詳細

定量分析結果：コミュニティノート付与投稿への行動変容

CN付与ユーザー投稿へのリポスト抑制効果

対象投稿400件のうち、121件（30.2%）にコミュニティノートが付与されていた。
CN付与ユーザの投稿に限定した分析を実施。

群	CN付与ユーザの投稿へのリポスト	CN未付与ユーザの投稿へのリポスト
対照群	52回（平均2.74回）	97回
介入群	14回（平均0.74回）	41回
抑制率	**73.1%**	**57.7%**

統計的検定（CN付与ユーザの投稿へのリポスト）

- Mann-Whitney U検定: U = 142.00, p = 0.1520（有意水準5%で有意差なし）
- 効果量（Cohen's d）：0.48（中程度）
- 考察: 統計的に有意な差は確認されなかったが、これはサンプルサイズ（各群n=19）の制約による検出力の不足も一因と考えられる。実用的には約73%という大きな抑制効果が観察されており、効果量も中程度（0.48）であった。より大規模な実験での検証が望まれる。

4-2. 有効性等に関する検証の個別詳細

定量分析結果：信頼性評価への影響

CN付与ユーザの投稿への信頼性評価（5段階）

群 × CN有無	平均評価	標準偏差	n
対照群・CN未付与	3.10	1.09	2,834
対照群・CN付与	3.12	1.07	1,366
介入群・CN未付与	3.06	1.05	2,978
介入群・CN付与	**3.00**	**1.07**	**1,423**

- 統計的検定（CN付与ユーザの投稿、介入群 vs 対照群）：Mann-Whitney U = 911,476, p = 0.0031（有意水準5%で有意）
- 考察：介入群ではCN付与ユーザの投稿に対する信頼性評価が有意に低下しており、ツールが信頼性判断に影響を与えていることを示す。

テーマ別の差異

- 災害・科学：客観的事実が重視されるトピックでは、CN情報の有無が判断に強く影響した（抑制効果が大）。
- 政治・エンタメ：個人の支持・選好が強く働くトピックでは、CN情報が表示されていても、確証バイアスによりリポストを選択するケースが散見されたが、全体としては抑制傾向にあった。

4-2. 有効性等に関する検証の個別詳細

定性分析：ユーザー受容性 (System Usability Scale)

ユーザビリティ評価 (SUS)

- 事後アンケートにおけるSUS（System Usability Scale）スコアを分析（n=19）。
- 結果: 平均スコアは 53.3点（標準偏差: 15.1、範囲: 27.5～75.0点）

指標	値
平均スコア	**53.3点**
標準偏差	15.1
中央値	55.0点
最小～最大	27.5～75.0点

- 解釈: SUSの標準的な基準では、68点以上が「平均以上」、80点以上が「優秀」とされる。本ツールのスコアは68点を下回り、ユーザビリティに改善の余地があることが示された。
- 考察: 本ツールはChrome拡張機能として既存のXインターフェースに情報を追加する形式であるが、情報の表示方法や視認性についてさらなる改善が求められる。

自由記述コメントからの洞察

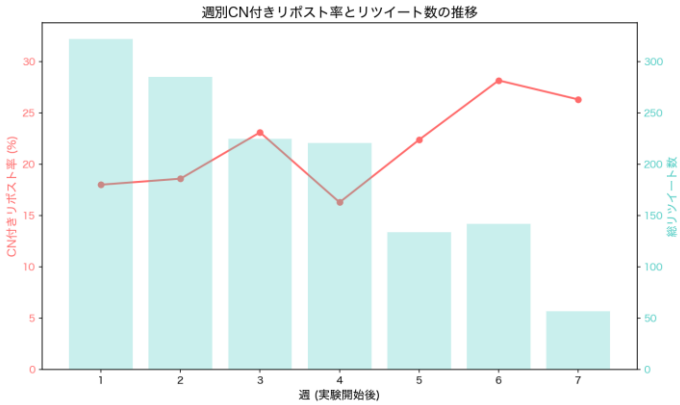
- ポジティブ:
  - 「普段は何気なくリポストしていたが、バッジが見えることで『本当に拡散していいのか？』と自問するようになった」「知らないアカウントの信頼性を測る指標として非常に役立つ」
- ネガティブ
  - 「画面が情報過多になる」「推しの活動者にバツ印がついているようで不快だった」

4-2. 有効性等に関する検証の個別詳細

追跡調査結果：長期的な行動変容

1ヶ月後の行動分析（追跡調査データ: 35名分）

- リポストの質の変化: 追跡調査の結果、実験直後の1週間（Week 1）におけるCN付き投稿のリポスト率は18.0%と低水準に抑えられていた。
- 習慣化の壁 (Habituation): しかし、2週間経過後（Week 3以降）はリポスト率が23.1%～28.2%へと上昇する傾向が見られた（図）。
  - 初期の抑制効果が時間経過とともに薄れる「慣れ（Habituation）」が生じている可能性が示唆される。
  - 今後の課題: 警告デザインの動的な変更など、持続的な効果を維持するためのUI改善が必要である。



図：週別CN付きリポスト率の推移（実験実施日基準）

信頼性表示の限界

- 技術的な介入による行動変容の可能性と課題が明らかになった。特に、ユーザーが強い関心や支持を持つトピック（政治やエンタメ）に関しては、単純な警告表示だけでは効果が限定的である傾向が見られた。

目次

5-1. 普及啓発活動の全体像

普及啓発活動に係る取組・成果の全体像

デジタル・ポジティブアクション事務局への報告

- 総務省やSNSを運営するデジタル・プラットフォーマー各社、関連団体等が推進する「デジタル・ポジティブアクション」事務局に、本研究の成果（信頼性評価表示の有効性、拡散行動の可視化結果）を報告。

政策提言の一環としての改善案提出

- 偽・誤情報対策に関する政策提言の一環として、研究成果を活用したインターフェース設計の改善案を提出。

主要SNSプラットフォーマーへの提言

- X（旧Twitter）をはじめとする主要SNSプラットフォーマーに対し、Chrome拡張で得られた利用者行動データをもとに、信頼性表示の改善提案を実施。
- コミュニティノート等の信頼性指標の表示方法に関する利用者視点のフィードバックを必要に応じて、公表資料として活用することについて協力。

5-2. 普及啓発活動の個別詳細

普及啓発活動の個別詳細

デジタル・ポジティブアクション事務局への報告

- 総務省やSNSを運営するデジタル・プラットフォーマー各社、関連団体等が推進する「デジタル・ポジティブアクション」事務局に、本研究の成果（信頼性評価表示の有効性、拡散行動の可視化結果）を報告していく予定。

政策提言への活用（今後の予定）

- 偽・誤情報対策に関する政策提言の一環として、研究成果を活用したインターフェース設計の改善案を提出していく。なお、今年度事業内では、事務局への報告に向けた資料準備等に着手した。

SNSプラットフォーマーへの改善提案（今後の予定）

- X（旧Twitter）をはじめとする主要SNSプラットフォーマーに対し、Chrome拡張で得られた利用者行動データをもとに、信頼性表示の改善提案を行っていく。
- コミュニティノート等の信頼性指標の表示方法に関する利用者視点のフィードバックを必要に応じて、公表資料として活用することについて協力していく。

目次

6-1. 研究・調査の総合的な考察

結論：ナッジとしての「信頼性可視化」の可能性と課題

主要な知見

1. リポスト抑制効果の観察: 全体で約59%、CN付与ユーザの投稿に限定すると約73%のリポスト抑制効果が観察された。ただし、サンプルサイズ（n=38）の制約により、リポスト数における統計的有意差は確認されなかった。

2. 信頼性評価への有意な影響: CN付与ユーザの投稿に対する信頼性評価においては、介入群と対照群の間に統計的に有意な差（p=0.0031）が確認された。これは、ツールがユーザーの認知に影響を与えていることを示す重要な知見である。

3. ユーザーエンパワーメントの実践: 本ツールは「拡散するな」と命令するのではなく、「判断材料」を提供するにとどまる。最終的な意思決定はユーザーに委ねられており、表現の自由を尊重するアプローチといえる。

6-1. 研究・調査の総合的な考察

発見された課題

1. サンプルサイズの制約

- 本実験は各群19名（計38名）で実施したが、リポスト行動における統計的有意差を検出するには検出力が不足していた可能性がある。効果量（Cohen's d）は約0.5と中程度であり、より大規模な実験での検証が必要である。

2. ユーザビリティの課題

- SUSスコアは平均53.3点と、標準的な基準（68点）を下回った。情報の表示方法や視認性についてさらなる改善が求められる。「情報過多」「不快感」といったネガティブなフィードバックも得られた。

3. 習慣化（Habituation）の問題

- 追跡調査において、実験直後（Week 1：18.0%）は高かった抑制効果が、2週間経過後のWeek 3以降（23%台後半）には減衰する傾向が定量的に確認された。警告表示への「慣れ」への対策が課題である。

6-1. 研究・調査の総合的な考察

政策的示唆

本研究から得られた示唆

- 発信者単位の信頼性表示の有効性: コンテンツ単位の評価（コミュニティノート）に加え、「発信者単位」の信頼性情報の可視化が、ユーザーの認知に影響を与えることが示された。
- サンプルサイズの重要性: 本研究の規模では行動変容（リポスト数）における統計的有意差は確認されなかったが、中程度の効果量が観察された。社会実装に向けては、より大規模な実証実験が必要である。

プラットフォーマーへの提言

- 信頼性指標の標準機能化: 発信者の過去のCN付与履歴等を可視化する機能の標準搭載を検討すべきである。
- UIデザインの改善: SUSスコアが低かった点を踏まえ、情報過多にならず、かつ効果的に情報を伝えるUIデザインの検討が必要。

政策立案者への提言

- 大規模実証研究への支援: 本実験で得られた知見を深化させるため、より大規模かつ長期的な実証研究への支援が必要。

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

今後の課題

1. サンプルサイズの拡大

- 本実験（n=38）では、リポスト行動における統計的有意差は確認されなかった。効果量（Cohen's d ≈ 0.5）から推定すると、80%の検出力を得るには各群約64名（計128名）程度のサンプルが必要と考えられる。
- 今後は、より大規模な実験により効果の検証を行うことが重要である。

2. ユーザビリティの改善

- SUSスコア（53.3点）が標準基準（68点）を下回っており、UI/UXの改善が必要である。
- 改善の方向性: 情報量を適切に制御し、必要な場合のみ詳細を表示するなど、ユーザー体験を損なわない設計が必要。

3. 習慣化（Habituation）への対策

- 同じUI表示が続くと警告効果が薄れる傾向が見られた。
- 改善の方向性: 警告のデザインを動的に変更する、信頼性スコアが著しく低い場合のみ強調表示するなど、状況に応じた最適化が必要。

4. モバイル環境への対応

- 本研究はPC版Chrome拡張機能で検証したが、SNS利用の大部分はスマートフォンアプリで行われている。
- 改善の方向性: モバイルアプリへの統合や、専用ブラウザアプリの開発が必要となる。

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

今後の展望

大規模実証実験の実施

- 本研究で得られた知見（効果量、必要サンプルサイズの推定等）を基に、より大規模な実験を計画・実施し、リポスト行動への効果を統計的に検証する。

UIデザインの改善と再検証

- SUSスコアの結果を踏まえ、ユーザビリティを向上させた新バージョンを開発し、再度効果検証を行う。

プラットフォーマーとの連携

- 本研究成果をSNSプラットフォーマーに共有し、発信者単位の信頼性表示機能の標準搭載について検討を促す。

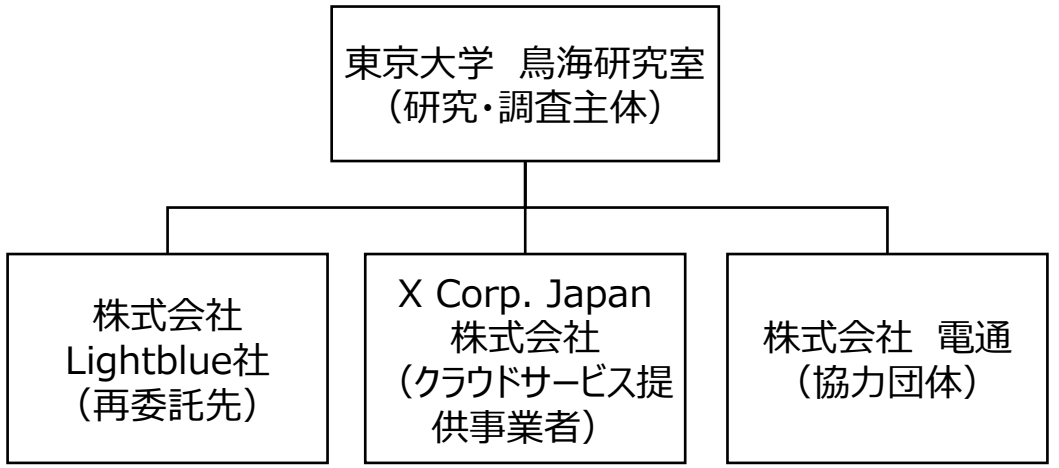
国・SNSを運営するデジタル・プラットフォーマー各社と討議すべき事項

- 信頼性指標の表示方法に関する検討
- 研究目的でのAPIアクセス拡充
- ユーザーエンパワーメントを促進するUI/UXガイドラインの策定

目次

7-1. 実施体制及び役割分担

本事業の実施体制図



※支払い発生無し（協力のみ）

各団体の役割・業務範囲

- 東京大学　鳥海研究室（研究・調査主体）
  - 調査研究の企画・統括
- 株式会社Lightblue
  - 被験者募集・実験
- 株式会社　電通
  - 調査テーマ・他調査データの共有、広報協力等
- X Corp. Japan株式会社
  - APIによるデータ取得

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）研究内容の詳細検討と確定	→							
（2）研究実施方法の確定	→	→						
（3）利用者の情報共有行動の可視化		→	→	→	→			
（4）被験者の募集		→	→	→	→			
（5）情報共有行動時の情報の信頼性評価				→	→	→		
（６）普及啓発活動への協力						→	→	→