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

生成AI時代における偽誤情報流通と認知特性の解明に関する研究・調査

成果報告書

2026/3/19

研03_東京大学大学院情報学環

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	• 生成AI技術の急速な発展によって人間の判断を欺くコンテンツが容易に作成されるようなっている。膨大な情報の中で、すべての情報を正確に収集・理解・吟味することが不可能であり、不本意な意思決定をしてしまう場面が増加。情報の受け手である人間の認知特性自体が偽情報に対する脆弱性となっている。 • そこで現在及び将来の偽誤情報脅威の体系化と日本固有の生成AI偽誤情報受容性の解明を図る		
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）	東京大学大学院情報学環澁谷研究室

研究および有効性等に関する検証の取組・成果

【研究項目1】

- 偽誤情報関連キーワードで収集したX投稿（2025年10-11月、336万件、リツイートを除く）では、政治政党批判やメディア批判に関するトピックが最も多く、ついで、偽誤情報関連の言説を批判否定するもの、排外主義的な主張に関するトピックが多い
- 偽誤情報関連投稿で用いられている手法としては信用失墜型（Discrediting）が多い → 発信元を攻撃で正しい情報さえも受け付けない土壌が形成される懸念
- 偽誤情報対策は単なる真偽判定の問題ではない。情報をどのように受容し、どのように評価し、どのような行動へと接続するのかという認知過程そのものを対象化する設計が不可欠

【研究項目2】

- 日米独仏印の5カ国でオンライン実験を実施（N=5,443）し、強い警告（メッセージ型）と控えめな警告（ラベル型）の効果を検証
- 動画の正確性評価（７段階評価）に加え、ユーザーの迷い（不確実性）を測定し、個人属性や地域、AIの認識、性格、価値観などで制御の上比較
- 現在主流となっている控えめな生成AIラベルは、ユーザーの認知的警戒心を高める効果は限定的、誤情報受容抑制の十分条件とはなり得ない可能性が示唆
- 技術的対策に加え、ユーザー自身の認知バイアスへ働きかける教育が不可欠
- 現状把握や将来の脅威の理解を深めるためにも、SNS プラットフォームにおけるデータ公開・透明性を求める必要性

研究・調査にあたっての課題・展望

- 今後は、①プラットフォームとの連携によるデータアクセスと透明性の確保、②単なる注意喚起ではなく識別能力を高める介入設計の検証、③「どちらとも言えない」という不確実性の機能と拡散行動への影響の追跡、④利用者の異質性を前提とした適応的設計、⑤実環境での長期的・持続的効果の検証が求められる
- 誤情報を一律に抑制することではなく、生成AI時代に即した利用者の判断過程そのものを支えるインターフェース設計も重要

代表者コメント

東京大学
大学院情報学環
准教授
澁谷遊野

日々変化する情報空間を的確に捉え、日本の文脈に即した情報流通構造をより深く理解していく必要があると考えます。偽誤情報への対応には特効薬はなく多層的・多面的なアプローチが不可欠です。今後は、プラットフォームデータ透明性の向上を求めるとともに、異質性を前提とした介入設計、さらに実環境における長期的効果の検証を進めていきます。

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

1. 生成AI時代における偽誤情報の急速な高度化

- 生成AIの発展により、文章・画像・音声・動画を用いた偽情報の作成が容易化し、実在の人物の顔や声を模倣したディープフェイクなど、従来よりも精巧な情報操作が可能となっている
- 一般利用者が真偽を見分けることが極めて困難な状況が拡大

2. 情報過多による人間の認知的限界

- デジタル空間には膨大な情報が流通し、すべてを正確に確認・吟味することは不可能で、時間的・認知的制約の中で、不十分な根拠に基づく判断が増加→「人間の認知特性そのもの」が偽情報に対する脆弱性となっている

3. 既存対策の限界

- AIによるフェイク検知や制度的規制は一定の効果を持つ一方で、完全な防御は困難
- 技術対策だけでは、利用者の判断プロセスや心理的要因に対応できないため、人間の認知特性を踏まえた新たな対策アプローチが不可欠

4. 日本社会における固有の課題

- 国際比較調査では、日本はAIへの理解度や受容のあり方が他国と異なる傾向が示されているものの、日本特有のメディア環境・文化的背景を考慮した研究が十分に行われていない
- 国内文脈に適合した偽情報対策の設計が求められている

5. 将来リスクへの備えの不足

- 生成AI技術は今後さらに進化し、より巧妙な偽情報が出現することが予想される
- 現時点の対策だけでなく、将来の脅威を見据えた体系的理解が必要
- 長期的視点に立った社会実装可能な対策設計が求められる

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

- 偽誤情報の実態と拡散構造をデータに基づき体系化
- 人間の認知特性を考慮した実証的な対策評価
- 日本社会に適した効果的な介入手法の開発
- 将来の高度化する脅威に対応できる知見の創出

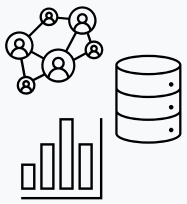
→ 生成AI時代における安全で信頼できる情報環境の実現に貢献

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

現状把握と実態解明

研究課題1

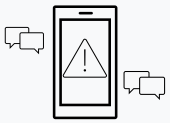


実態解明と可視化
偽誤情報の特徴・拡散パターンを体系化
誤判断を招く形式・手法をデータで解明

人間中心の対策設計と開発

研究課題2

認知特性・バイアスの理解
「なぜ人は誤った判断をするのか」を説明

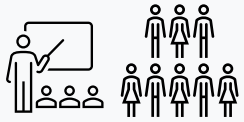


効果的な介入手法
警告ラベル等の介入の有効性を実証的に評価
ユーザー負担が少なく実効性の高い対策手法の設計

日本に適した対策
日本固有の認知傾向やメディア環境を考慮
国際比較で課題の明確化

社会実装と未来への貢献

普及啓発活動



社会的インパクト
市民の情報判断力の向上
政策立案・教育施策への科学的エビデンス提供

将来リスクへの長期的備え
進化する生成AI技術に対応した対策の基盤構築

研究基盤：データ駆動アプローチと学際的連携

目次

3-1. 研究の全体像

研究に係る取組・成果の全体像

本研究では、生成AIの普及に伴って高度化する偽・誤情報の脅威に対し、データに基づく体系的理解と、人間中心の対策モデルの構築を目指す。そのために、（1）現在および将来の偽誤情報脅威の体系化、（2）日本固有の生成AI偽誤情報受容性の実証分析、という2つの研究課題を相互に連関させながら推進する。

研究課題１：生成AIを活用した偽誤情報がデジタル空間でどのように生成・流通しているかを大規模データ分析により解明

Twitter（現X）、Meta、Google、TikTokなど主要プラットフォームからAPIや手作業を通じてデータを収集し、日本語を含むコンテンツを中心としたデータベースを構築する。機械学習やネットワーク解析を用いて、画像パターン、頻出フレーズ、文体、拡散構造などを分析し、健康デマや政治的虚偽主張といった具体的事例の拡散メカニズムを明らかにする。またXのコミュニティノートの分析を通じて、投稿の問題点や脅威パターンを補完的に評価する。さらに、海外事例や最新の生成AI技術動向を踏まえ、社会経済的・文化的背景と関連づけた将来脅威の検討を行い、問題パターン・脅威パターン・認知特性・ターゲット人口などの観点から包括的な「偽誤情報脅威マッピング」を完成させる。

研究課題2：体系化された脅威知見を基盤として、日本固有の生成AI偽誤情報受容の特徴を実証的に分析

先行研究で指摘されてきたアルゴリズム嫌悪やAI受容の文脈依存性を踏まえ、事前調査（Pre-study）により仮説を精緻化した上で、介入を組み込んだオンライン実験を設計する。具体的には、生成AI明示ラベルの提示など複数の介入手法を比較し、参加者を処置群・対照群にランダムに割り当てて生成AI作成動画の正確性評価を実施する。本実験は日本の他欧米・アジア圏を対象とした大規模国際比較として行い、AI理解度、デジタルリテラシー、社会経済背景などの共変量を含めた多変量解析により、誤判断に至る認知要因を定量的に解明する。

3-2. 研究の個別詳細

研究項目1：現在及び将来の偽誤情報脅威の体系化

- 偽誤情報がどのように作られ、どのように流通し、どのような特徴を持つかを体系的に整理する

分析手法

X投稿データ（日本語投稿）

↓

テキスト前処理
- URL/メンション除去・短文除去/日本語表現の正規化

↓

文埋め込み（Sentence Embedding）
- 意味表現ベクトル化

文章が持つ意味をコンピュータが計算可能な数値（座標）に変換する技術で単なる単語の一致ではなく、文脈やニュアンスを考慮して変換するため、意味の近い文章同士を近くに、遠い文章を遠くに配置することが可能となる

↓

クラスタリング（MiniBatch K-means）

↓

類似度閾値による判定（ナラティブに割当/OTHER（未割当））

意味表現ベクトルにおいて、新しいデータが既存の言説群とどれくらい近いかを判定するプロセス。閾値を満たしたデータは、特定の話題グループ（ナラティブ）に分類割当される。どの話題とも類似度が低く、閾値を満たさなかったデータはOTHERとして扱われる。

↓

クラスタ解釈（TF-IDF /代表ツイート抽出）

自動分類されたデータ群（クラスタ）が、具体的にどのような話題で構成されているかを人間が理解するための補助分析。TF-IDFは、その話題の中で「特徴的に使われている重要語」を抽出する計算式。代表ツイート抽出：その話題グループの中心（重心）に位置するデータを特定する。

- X API Filtered Stream[1]エンドポイントでリツイートは除外・日本語投稿のみを対象にキーワード[2]を設定して収集
  - キーワードの選定には先行研究で用いられているキーワードとChatGPTで与えられたキーワードを組み合わせた
- 収集期間：2025年10月11日から11月26日まで（日本時間）

- 収集データ：3,367,289投稿
- テキスト前処理後：2,582,812投稿を対象として分析
- 収集データの中には、偽誤情報とは言えない投稿や偽誤情報を指摘・批判する投稿も含まれているが、それらも全て含めて分析を行う

- テキスト前処理後の2,582,812投稿を対象に100のトピックをクラスタリング手法を用いて抽出

[1] X（旧Twitter）社が提供する公式のデータ接続窓口の一つ。世界中で投稿される膨大な全データの中から、あらかじめ設定した条件（キーワードやルール）に合致する投稿のみを、リアルタイムで選別して受信する仕組み

[2] **本研究課題で用いたキーワード**
一般的な偽誤情報キーワード：マスゴミ OR 偽情報 OR フェイクニュース OR デマ OR 誤報 OR 虚偽 OR 流言 OR 噂話 OR 真偽不明 OR 怪情報 OR 報道しない自由 OR 人工地震 OR 地震予知 OR 気象兵器 OR HAARP OR 放射能デマ OR 水道水汚染 OR 人工ウイルス OR 不正選挙 OR 選挙操作 OR 投票操作 OR 選管陰謀 OR 人口削減 OR 移民陰謀 OR 5G危険 OR 電磁波攻撃 OR 監視社会 OR AI陰謀 OR マイクロチップ OR DNA改造
陰謀論キーワード：陰謀論 OR 隠蔽 OR 操作 OR 支配 OR ディープステート OR イルミナティ OR 秘密結社
コロナ関連キーワード：コロナワクチン OR ワクチン副反応 OR 反ワクチン OR ワクチン義務化 OR ワクチンパスポート OR ファイザー OR モデルナ OR ビッグファーマ
気候変動キーワード：気候変動 OR 地球温暖化 OR 気候危機 OR 海面上昇 OR 温室効果ガス OR 二酸化炭素排出 OR 再生可能エネルギー OR パリ協定 OR 温暖化 OR 気候変動
生成AI：生成AI OR GenAI

3-2. 研究の個別詳細

研究項目1：現在及び将来の偽誤情報脅威の体系化

- 収集データに見られるトピックとしては、政治・政党批判に関連する言説が最も多い、次いでメディア批判に関する言説などが続く
- 生成AIに関する言説は、創作に関わる生成AIに関する議論が最も多く見られ、偽誤情報そのものに関連する生成AI言説は多くはない

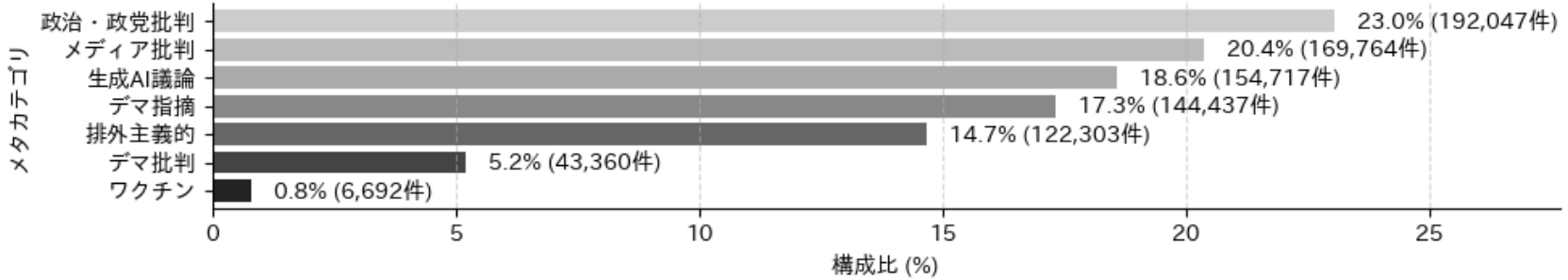


図. 100クラスタのメタカテゴリ別のと投稿数と構成比（その他を除く）
註：図中メタカテゴリは、下表の人手ラベルをさらにまとめたカテゴリを示す

表. 抽出した100クラスタのうち最も投稿数が多い10のクラスタと投稿数（その他を除く）

人手ラベル	投稿数	代表語
デマ指摘	44,920	デマ, マスゴミ, 笑い, 遣る, 過ぎる, 馬鹿, どう, 阿呆, 本当, やばい, 酷い, 操作
排外主義的・ナショナリズム	30,929	日本, 支配, 外国, 差別, 中国, デマ, 国民, 戦争, 来る, アメリカ-America, 植民, 移民
陰謀論批判	30,067	陰謀, 信ずる, 来る, 嵌まる, デマ, 過ぎる, 分かる, こう, 本当, 笑い, 馬鹿, 自分
メディア批判	29,464	デマ, 止める, 支配, 流す, 遣る, マスゴミ, もう, 行く, 操作, 下さる, 来る, 欲しい
政治・政党批判	29,329	政治, デマ, 議員, 民主, 参政, 支持, 立憲, 国民, 政党, 自民, 国会, 印象
排外主義的	29,144	中国, 日本, 共産, 支配, タカイチ, デマ, 中華, マスゴミ, 政府, 世界, 来る, 朝日
メディア批判	27,793	報道, メディア-media, マスコミ, 偏向, マスゴミ, 印象, 情報, 記者, 新聞, デマ, 操作, ニュース
生成AI関連議論	27,579	AI, 生成, 使う, 人間, 創作, 作る, 自分, 技術, 問題, 来る, 学習, 行く
生成AI関連議論（創作）	27,301	AI, 生成, 画像, イラスト-illustration, 描く, 使う, 作る, 絵師, 写真, 自分, 絵書き, 来る
メディア批判	26,268	デマ, マスゴミ, 馬鹿, 過ぎる, 笑い, 本当, 悪い, 操作, 支配, 全部, 印象, 遣る

3-2. 研究の個別詳細

研究項目1：現在及び将来の偽誤情報脅威の体系化

- 情報操作の手法枠組（DEPICT）に基づき、収集データで用いられている手法を整理
  - 信用失墜（Discrediting）・感情（Emotion）・分断/敵味方化（Polarization）・なりすまし（Impersonation）・陰謀化（Conspiracy）・あらし（Trolling）
- 収集データでは、**信用失墜型（Discrediting）**が多い
  → 発信元を攻撃することで正しい情報さえも受け付けない土壌が形成されている懸念
- メディア批判・政治政党批判において信用失墜に関する戦略がとられている
  - 特定の個人および・政府省庁に対する批判が多い

表. トピックごとに用いられている操作手法戦略等の例（偽誤情報脅威マッピング）

分類	戦略	手法	対象	具体的フレーズの例
メディア批判	信用失墜	なし	マスコミ	「テレビ局は不祥事を報道せず隠蔽した」
	信用失墜	なし	なし	「いじめ隠蔽に奪われていく」
	信用失墜	疑問・疑念	マスコミ	「偏向報道があまりにも酷すぎた（問いかけ）」
	なし	なし	なし	（特定の戦略を伴わない不満の表出）
	挑発・煽り	なし	なし	（感情的な攻撃や嘲笑）
政治・政党批判	信用失墜	なし	特定の個人・団体	「クソデマ」「卑怯ですよ？」
	信用失墜	なし	政府・省庁	「独裁的な政治が社会を崩壊させた」
	信用失墜	蔑称・他者化	なし	「デマ流すの辞めろよ」
	分断・敵味方化	なし	なし	「共産党なんて反社でしょw（属性排除）」

註：背景色がある列は主要パタン

3-2. 研究の個別詳細

研究項目1：現在及び将来の偽誤情報脅威の体系化

- 言説の構造を体系的に分析するため、対象・側面・評価の3軸分類体系を構築
- この3軸分類により、「誰に対して」「何について」「どのように評価しているか」という言説構造を体系的に分析
- メタカテゴリごとに100件の投稿をランダムサンプリングし、人手による3軸分類ラベリングを実施

表. 言説分析のための分類体系（対象・側面・評価）

（A）対象（Target）

カテゴリ	概念的定義	判定基準（予先の例）
政府・政治	国家・自治体の意思決定および制度運用（選挙を含む）	政府、政党、政治家、行政、司法、警察、選挙制度など
専門知	医療・科学・専門家による知識生産およびその正当性	医師、研究者、統計、科学的知見、専門家の信頼性など
メディア	報道機関・プラットフォーム等による情報流通および可視性制御	テレビ、新聞、SNS、プラットフォーム、検閲、アルゴリズムなど
経済	企業・産業・市場による利害および資源配分	企業、業界、市場、利権、利益構造など
社会集団	属性集団や社会的カテゴリー（分断の境界）	特定の民族、国民、支持者、集団など
国際・安全保障	国家間関係、外国政府、国際秩序および安全保障	外国政府、国際機関、外交、紛争など
環境	気候変動、災害、エネルギー問題など環境領域そのもの	気候変動、原発、再生可能エネルギーなど
科学技術	AI、通信、監視技術等の技術そのもの	AI、監視技術、通信技術、サイバー技術など
個人	特定の個人に対する言及	個人名、固有名詞など
対象不明	文脈不足により対象が特定できない発話	主語不明、引用のみの投稿など

（B）側面（Aspect）

カテゴリ	概念的定義	判定基準（例）
行為・対応	対象の発言、行動、政策、対応、制度運用	政策決定、対応の適否、行動評価など
人となり・内的特性	対象の能力、意図、誠実性、道徳性などの人格的評価	能力、誠実性、意図、信頼性など
主張の真偽	情報や発言の正確性、真偽、信頼性	デマ、虚偽、根拠、捏造、正確性など

（C）評価（Evaluation）

カテゴリ	概念的定義	判定基準（例）
肯定	対象を良い・正しい・望ましいと評価する発話	称賛、支持、擁護、正当化など
否定	対象を悪い・誤り・望ましくないと評価する発話	批判、非難、侮辱、能力否定など
両義的	同一対象に対し肯定と否定の両方の評価を含む発話	「良いが問題もある」など
非評価	対象への価値判断を含まない発話	事実報告、引用、リンク共有など

3-2. 研究の個別詳細

研究項目1：現在及び将来の偽誤情報脅威の体系化

- 言説の多くは、政府・政治およびメディアが主な対象で、評価は否定的なものが多く、制度や情報源に対する批判的言説が中心
- 評価対象は主に行為・対応や主張の真偽に集中し、制度運用や情報の信頼性が主要な争点

- 政治・政党に関する言説は、約76%が否定評価であり、政策や政治行動への批判が中心
- メディアに関する言説も約73%が否定評価であり、報道姿勢や情報操作への不信が多い
- 生成AIに関する言説は非評価が最多（40%）であり、評価が分化した過渡的な議論段階か

表. ランダムサンプリング100件の3軸分類結果の集計

メタカテゴリ	対象	側面	評価	N	割合
メディア批判	メディア	行為・対応	否定	21	21.00%
メディア批判	メディア	人となり・内的特性	否定	10	10.00%
メディア批判	メディア	主張の真偽	否定	8	8.00%
メディア批判	対象不明	行為・対応	非評価	7	7.00%
メディア批判	個人	人となり・内的特性	否定	5	5.00%
政治・政党批判	政府・政治	行為・対応	否定	30	30.00%
政治・政党批判	政府・政治	人となり・内的特性	否定	11	11.00%
政治・政党批判	メディア	行為・対応	否定	5	5.00%
政治・政党批判	個人	行為・対応	否定	4	4.00%
政治・政党批判	政府・政治	行為・対応	非評価	3	3.00%
生成AI議論	科学技術	主張の真偽	非評価	16	16.00%
生成AI議論	科学技術	行為・対応	肯定	9	9.00%
生成AI議論	科学技術	行為・対応	否定	9	9.00%
生成AI議論	科学技術	主張の真偽	否定	7	7.00%
生成AI議論	社会集団	人となり・内的特性	否定	6	6.00%

表. 各トピックごとの多重クロス集計

メタカテゴリ	最頻評価	上位3対象	上位3側面	最も顕著な組合せ（上位1）
政治・政党批判	否定 (76.0%)	政府・政治 (48) / 個人 (13) / 社会集団 (8)	行為・対応 (59) / 人となり・内的特性 (21) / 主張の真偽 (19)	政府・政治に対する行為・対応の否定 (30件)
メディア批判	否定 (73.0%)	メディア (51) / 個人 (16) / 対象不明 (16)	行為・対応 (42) / 人となり・内的特性 (26) / 主張の真偽 (20)	メディアに対する行為・対応の否定 (21件)
生成AI議論	非評価 (40.0%)	科学技術 (47) / 社会集団 (16) / 経済 (13)	行為・対応 (49) / 主張の真偽 (38) / 人となり・内的特性 (12)	科学技術に対する主張の真偽の非評価 (16件)

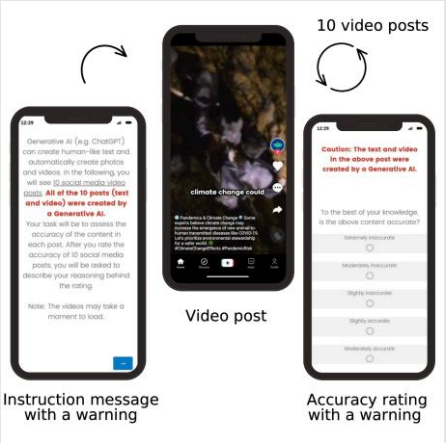
3-2. 研究の個別詳細

研究項目2：日本固有の生成AI偽誤情報受容性の実証分析

- 人間の認知的脆弱性の実証的解明と介入手法の有効性を評価

分析手法

本研究では、誤情報を含む動画と含まない動画を、商用生成AI動画作成サービス（Lumen5 / InVideo AI）を用いて作成し、5カ国（日本・ドイツ・フランス・アメリカ・インドネシア）を対象としたオンライン実験を実施する。被験者はランダムに3つの条件群へ割り当てられ、提示された動画の正確性を評価する。
介入条件として、以下の3群を設定する。
（1）動画の前後に「生成AIが作成した」旨を示す警告メッセージを提示する群（Message）
（2）動画内左下に既存プラットフォームで用いられている「生成AI作成」というラベル表示を提示する群（Label）
（3）警告やラベルを一切表示しない統制群
これらの条件を比較することで、生成AI関連の偽誤情報に対する介入手法の効果を実証的に検証する。さらに、被験者のAI理解度、デジタルリテラシー、認知傾向などの個人特性を測定し、誤った判断に至る要因との関連を多角的に分析する。



実験用の動画を生成AIで作成

誤情報を含む見出しと含まない見出しをそれぞれ4件ファクトチェック団体*の記事から抽出

生成AIビデオメーカー(lumen5とinvideo AI)でテキストからソーシャルメディア投稿用動画を生成

実験デザイン

被験者は各動画を視聴後、内容の正確さを7段階のリッカート尺度で選択（あなたの知る限り、上記の内容は正確ですか？）

全被験者
- 処置群（Label）：動画内に生成AIが動画を作成したことを示す警告ラベルを表示
- 処置群（Message）：動画視聴前後に生成AIが動画を作成したことを示す強調する警告メッセージを表示
- 対照群：動画が生成AIによって作成されたことを知らされない

註：Politifact、Full Fact、Science Feedback、リトマス、日本ファクトチェックセンター、ファクトチェック・ナビの記事から選定し、PreStudyを行い本実験で用いる動画を選定

3-2. 研究の個別詳細

研究項目2：日本固有の生成AI偽誤情報受容性の実証分析

- 人間の認知的脆弱性の実証的解明と介入手法の有効性を評価

分析手法（続）

被験者

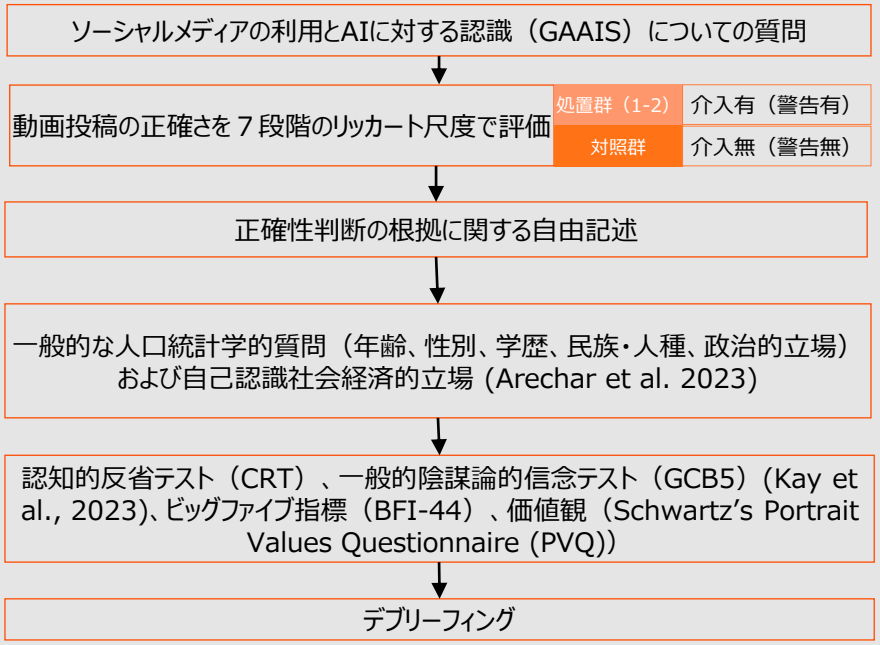
- 日本・米国・ドイツ・フランス・インドネシアのソーシャルメディアユーザー（N=6,722）をPureSpectrum経由で募集
- 回答した被験者には謝金（0.5～0.8USD）を支払う
- 18歳以上のを募集し、性別と年齢の分布はセンサスに均等割合となるようにする
- 質問紙にアテンションチェック問題を複数用意し、基準に満たなかった被験者1,297を分析から除外し、分析対象の被験者は5,443名

表. オンライン実験の対象地域と回答者数

国	実施言語	回答者数	分析対象の回答者数
米国	英語	1,306	1,054
ドイツ	ドイツ語	1,284	1,050
フランス	フランス語	1,252	1,063
日本	日本語	1,681	1,285
インドネシア	インドネシア語	1,199	991
合計		6,722	5,443

註：分析対象の回答者数
- 質問紙の前半・後半にアテンションチェック問題を3問配置
- うち2問以上に適切に回答できなかった被験者は除外
- すべての質問に回答した被験者を有効回答者とする

オンライン実験の流れ（質問紙）



3-2. 研究の個別詳細

研究項目2：日本固有の生成AI偽誤情報受容性の実証分析

- 人間の認知的脆弱性の実証的解明と介入手法の有効性を評価

分析手法（続）

Identification Strategies

事後層化（poststratification）法としてEntropy Balancingをデータに適用（Miratrix et al. 2013; Allcott et al., 2020）
- 先行研究に基づき、年齢、性別、人種／民族、政党支持率に応じて観測値に重み付けを行う（hainmueller et al., 2012）

その後、以下の仮説を検証する

仮説 1	ソーシャルメディア投稿の動画とテキストがAIによって作成されたという情報に触れることは、その後のソーシャルメディア投稿の正確性評価に影響を与える

次式に基づき介入の効果を推定

$$Y_{ij} = \beta_0 + \beta_1 C_{1i} + \beta_2 C_{2i} + \beta_3 T_{ij} + \beta_4 (C_{1i} \times T_{ij}) + \beta_5 (C_{2i} \times T_{ij}) + \mathbf{X}_i' \boldsymbol{\gamma} + \varepsilon_{ij}$$

- 被験者uによる投稿i の正確性スコア（1-7）
- 被験者i がメッセージ警告介入を受けた場合のバイナリ指標
- 被験者i がラベル警告介入を受けた場合のバイナリ指標
- 動画投稿tがTrueかどうかのバイナリ指標
- 真偽判別に関するAET（ユーザーが真の投稿と偽の投稿を相対的に見分ける能力）
- Covariates (GAASI、GCT-5、CRT、ソーシャルメディア／国民メディアへの信頼度、COVIDワクチン接種状況、自己申告による社会経済的レベル、性格、民族、年齢、性別、ソーシャルメディア利用状況、動画形式)

仮説2	介入効果は、被験者の正確性判断への迷いへ影響を与える

さらに、上の仮説に基づき、被験者の正確性判断の不確実性（U_{ij}）を線形確率モデルとロジスティック回帰モデルの両方を用いてモデル化

$$U_{ij} = \begin{cases} 1 & \text{if } Y_{ij} = 4 \text{ (Neither)} \\ 0 & \text{otherwise} \end{cases}$$

3-2. 研究の個別詳細

研究項目2：日本固有の生成AI偽誤情報受容性の実証分析

- 5カ国（日本・ドイツ・フランス・アメリカ・インドネシア）の被験者5,443名を3グループにランダム割り当て
  - 介入パタン1：ラベル警告（弱）
  - 介入パタン2：メッセージ警告（強）
  - 介入なし
- 生成AI動画を視聴後、動画の正確性を7段階尺度で評価
- どの被験者グループでも誤った動画で「どちらとも言えない」の選択割合が最も高い（25.2-27.2%）
  → 多くのユーザーが真偽判断に迷い、判断の不確実性が大きいか

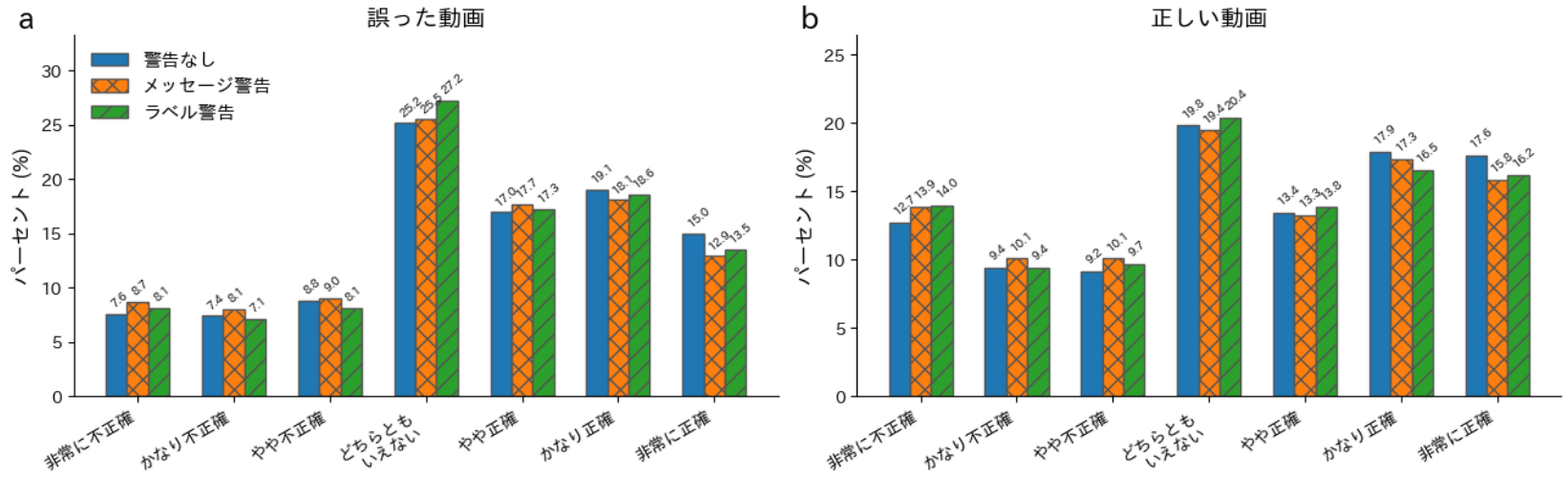


図. 生成AI動画に対する正確性評価の分布（5カ国・N=5,443）

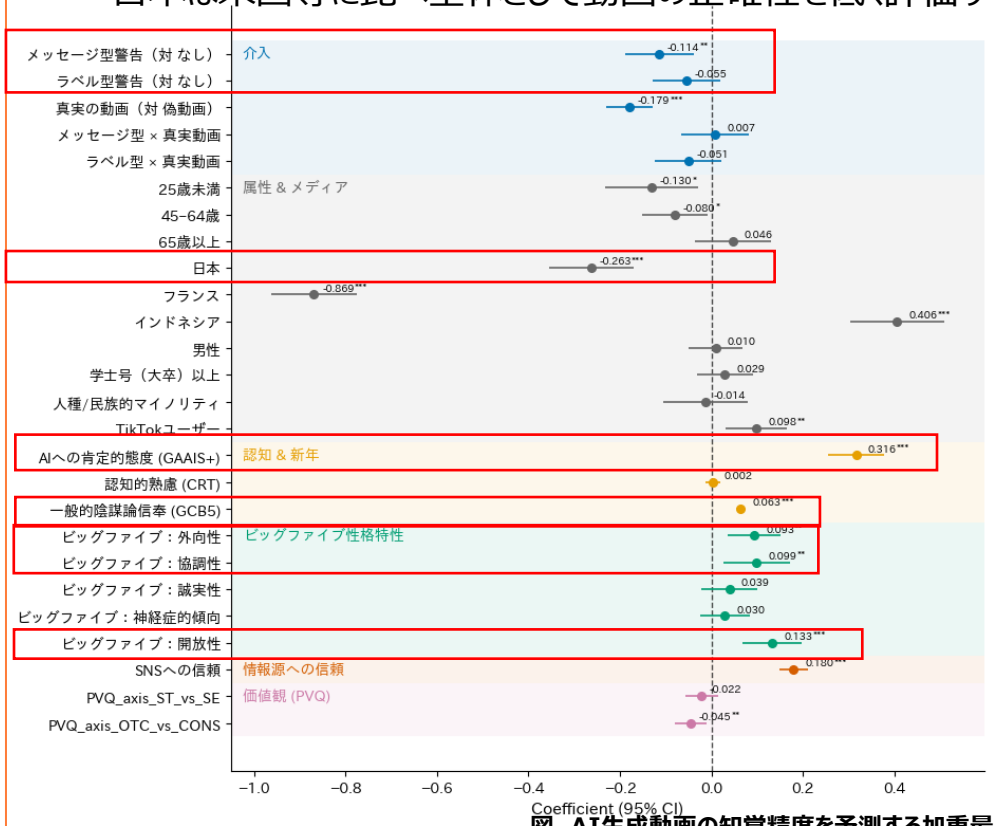
(a) 誤情報を含む動画（False Video Posts）に対する評価(b) 正しい情報を含む動画（True Video Posts）に対する評価。各バーは介入条件別の回答分布を示す（No warning / Label warning / Message warning）

3-2. 研究の個別詳細

研究項目2：日本固有の生成AI偽誤情報受容性の実証分析

介入が動画の正確性判断へ与える効果の検証

- メッセージ型介入（強）は 動画の正確性評価を下げるが、ラベル型介入（弱）は 影響は与えない
- 警告は全体の警戒心を高めるが、本物と偽物を正しく見分ける力（識別力）を向上させる効果はない
- 警告のデザイン以上に、本人のAIへの態度、性格特性、陰謀論への親和性が判断を左右している
- 日本は米国等に比べ全体として動画の正確性を低く評価する傾向がある



図の読み方

横軸
- 点は影響の大きさを示し、横の線はその推定の幅（95%信頼区間）を示す
- 点が中央の縦の点線（0）をまたいでいない場合は、統計的に意味のある影響と考える
- 右にあるほど動画を正確だと評価しやすい
- 左にあるほど動画を正確だと評価しにくい

縦軸（項目を背景色でにグループ分け）
- 介入：実験で操作した警告の種類の影響、動画が正しい内容かなど
- 属性・メディア：年齢、国、性別、TikTok利用など
- 認知・信念：AIへの態度、陰謀論傾向、認知的熟慮（CRT）など
- 性格特性：外向性・協調性などの性格要因
- 価値観：シュワルツの価値観尺度に基づく項目（伝統、権力、普遍主義など）
- 信頼：ニュースやSNSなどへの信頼度

図. AI生成動画の知覚精度を予測する加重最小二乗法（WLS）結果

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ 点は回帰係数を、水平線は参加者レベルでクラスタリングされた95%信頼区間を示す

3-2. 研究の個別詳細

研究項目2：日本固有の生成AI偽誤情報受容性の実証分析

介入によって人はどれくらい『どちらともいえない』と答えやすくなるかに関する効果の検証

- メッセージ型（強）は被験者が動画の正確性について「どちらとも言えない」を選択するかどうかに影響はほぼないが、ラベル型（弱）は「どちらとも言えない」を選びやすくなる（+2.18pp 有意）

→ 小さな動画内ラベル（弱いラベル）は、偽動画に対して判断を保留する傾向を有意に増やしている（不確実性を広げている）

→ しかし真偽を区別させてはいない

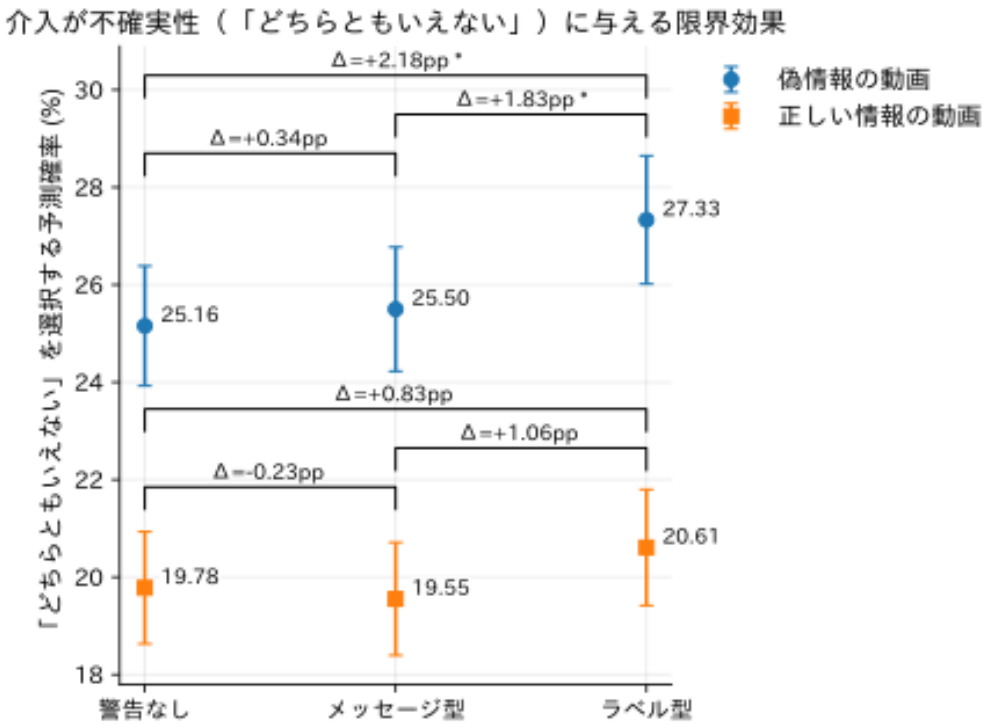


図. 介入が不確実性（「どちらとも言えない」）に与える限界効果

図の読み方

この図は、動画に警告をつけると、人はどれくらい『どちらともいえない』と答えやすくなるかを示す
ここでいう「どちらともいえない」は、動画が本当か誤っているのかかをはっきり判断しない＝不確実さ（迷い）の表れとも捉えられる

横軸
警告なし・メッセージ型（強めの警告文）・ラベル型（動画内に小さく表示される警告）のどの種類の警告かを示す

縦軸 （左側）
「どちらともいえない」と答える人の割合（%）
→ 数字が高いほど、判断を保留する人が多いことを意味

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

- 研究課題1（偽誤情報脅威の体系化）および研究課題2（受容性の実証分析）で得られた知見を基盤として、偽・誤情報対策に資する技術・システムの実効性を検証する

手法

（１）**専門家・実務家との協働検証**
- 心理学研究者、ヒューマンインターフェースデザイン研究者との議論を通じて、研究成果に基づくインターフェイスデザインの検討
- ファクトチェック団体関係者やユーザーとの警告表示・介入手法の設計に関する意見交換
- プラットフォーム運用者・メディア関係者との意見交換

（２）**オンライン実験を拡張したフィールドテストの実施**
- オンライン実験環境下での試験運用
- SNSプラットフォームTikTokを対象に実証評価（右図）

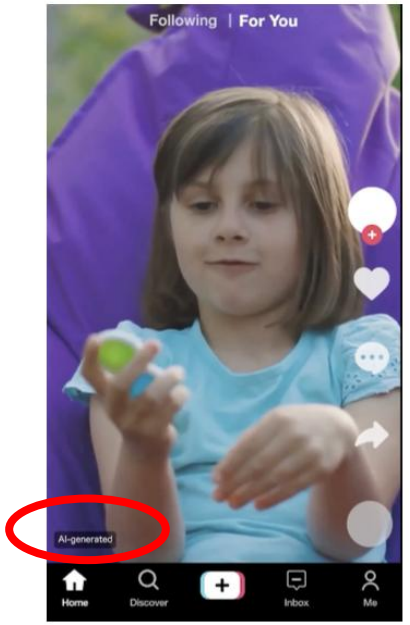


図. 図中赤丸内の「AI生成（AI-generated）」というTikTokラベルに対するユーザーの反応をフィールドテスト

評価

（１）**指標**
- ユーザーの反応・行動変化
- 情報信頼性の向上度
- システムの操作性・利便性に関する調査

（２）**ユーザー視点の評価**
- ユーザーインタビューの実施
- 介入表示の分かりやすさ・受容性の評価

4-2. 有効性等に関する検証の個別詳細

オンライン実験を拡張したフィールドテストの実施

- 弱いラベル警告を受けたユーザーの特徴は、クリック数が最も多く（35.65）、滞在時間も最長（281秒）で、First Clickが最も早い（18.9秒）
  →ラベル表示はユーザーの操作行動を増加させ、より多くの探索・確認行動を促している可能性
- 強いメッセージ警告を受けたユーザーの特徴は、滞在時間はやや短く、クリック数は介入なしとほぼ同水準
  →Message警告は認知的注意を喚起するが、積極的な操作行動にはつながりにくいか
- 国別ではインドネシアで最も活発な操作行動が観察される
- 日本では滞在時間は長いもののクリック数が少なく、慎重だが受動的な判断スタイルが示唆された

表. 介入条件別のユーザー行動指標平均値

条件	平均クリック数	平均ページ滞在時間（秒）	平均First Click（秒）	平均Last Click（秒）
	関心・探索行動の指標	情報処理に要した時間	迷い・判断の早さ	熟考の度合い
介入なし（対照群）	**29.17**	**261.17**	**25.85**	**39.12**
弱いラベル警告	**35.65**	**281.55**	**18.87**	**40.49**
強いメッセージ警告	**29.79**	**250.47**	**24.32**	**36.69**

表. 国別のユーザー行動指標平均値

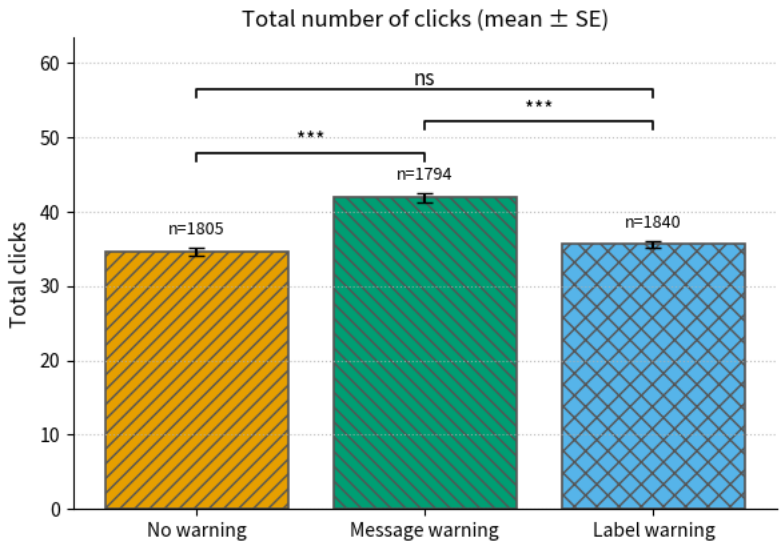
国	平均クリック数	平均ページ滞在時間	First Click	Last Click
ドイツ	27.18	232.91	22.24	35.60
アメリカ	28.59	274.02	23.69	41.38
フランス	32.22	225.39	20.99	35.19
インドネシア	**45.96**	**297.40**	**18.96**	**41.78**
日本	26.70	**290.75**	**27.57**	39.93

4-2. 有効性等に関する検証の個別詳細

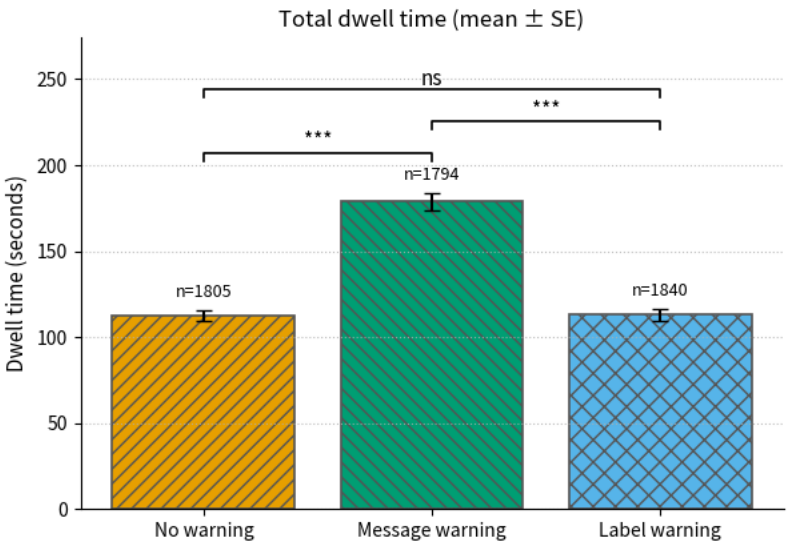
オンライン実験を拡張したフィールドテストの実施

操作性・利便性への影響

- 行動ログ（クリック数・滞在時間・First/Last Click）により、介入が操作行動に与える影響を評価
- 強いメッセージ警告ではクリック数と滞在時間が増加し、より多くの探索行動を促す傾向
- 弱いラベル警告では操作行動への影響は比較的小さく、ユーザー負担は限定的



図：介入条件別のクリック数



図：介入条件別の滞在時間

4-2. 有効性等に関する検証の個別詳細

オンライン実験を拡張したフィールドテストの実施

- 研究課題２オンライン実験を拡張し、正確性どのように判断をしたのか自由記述を求めた結果（1）証拠の有無（2）既有知識との整合性（3）表現スタイルへの信頼感、(4) 個人的感覚・直感、(5) 不確実性の表明が抽出された
- 特に誤情報動画では「根拠が不十分」「判断できない」という回答が多数を占め、多くのユーザーが判断に迷っている状態にあることが確認された

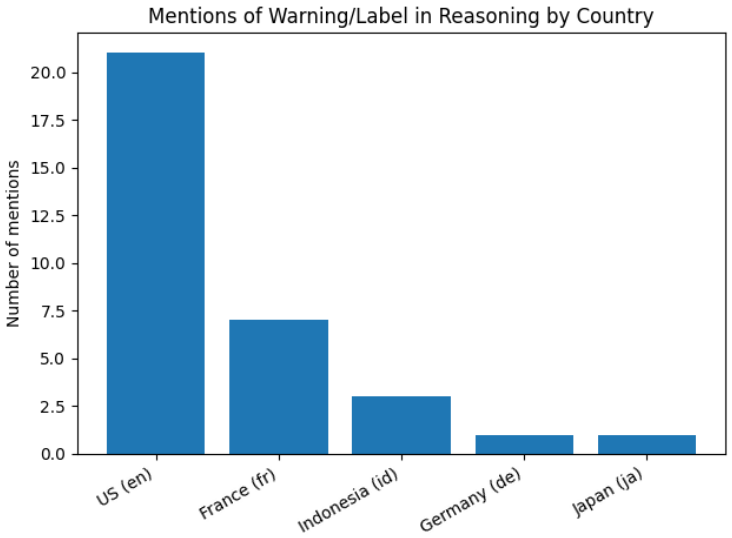
表. 自由記述コメントのカテゴリと代表コメント例

判断理由カテゴリ	言語	実際の声（抜粋）
証拠不足による判断	EN	“There is no source or evidence to support the claim.”
	FR	“Aucune preuve concrète n’est présentée.”
	DE	“Es fehlen verlässliche Belege.”
	JA	「具体的な根拠が示されていないため信用できない」
	ID	“Tidak ada bukti yang jelas untuk mendukung informasi ini.”
既有知識との整合性	EN	“This matches what I already know.”
	FR	“Cela correspond à mes connaissances.”
	DE	“Das stimmt mit meinem Wissen überein.”
	JA	「自分の知識と合致している」
	ID	“Informasi ini sesuai dengan yang saya ketahui.”
表現スタイルへの不信	EN	“The tone seems exaggerated and unrealistic.”
	FR	“Le ton est trop sensationnaliste pour être crédible.”
	DE	“Es wirkt übertrieben und manipulativ.”
	JA	「表現が大げさで信用できない」
	ID	“Cara penyampaiannya terlalu berlebihan.”
個人的経験に基づく判断	EN	“From my personal experience, this doesn’t seem true.”
	FR	“D’après mon expérience, cela paraît peu probable.”
	DE	“Aus eigener Erfahrung halte ich das für unwahrscheinlich.”
	JA	「自分の経験から考えると信じにくい」
	ID	“Berdasarkan pengalaman pribadi saya, ini kurang masuk akal.”
不確実性の表明	EN	“Not enough information to decide.”
	FR	“Je ne suis pas sûr, difficile à dire.”
	DE	“Ich bin mir unsicher.”
	JA	「判断材料が足りず、どちらとも言えない」
	ID	“Sulit menentukan apakah ini benar atau salah.”
直感的判断	EN	“It just doesn’t feel right.”
	FR	“Quelque chose semble suspect.”
	DE	“Es fühlt sich nicht glaubwürdig an.”
	JA	「なんとなく怪しいと感じた」

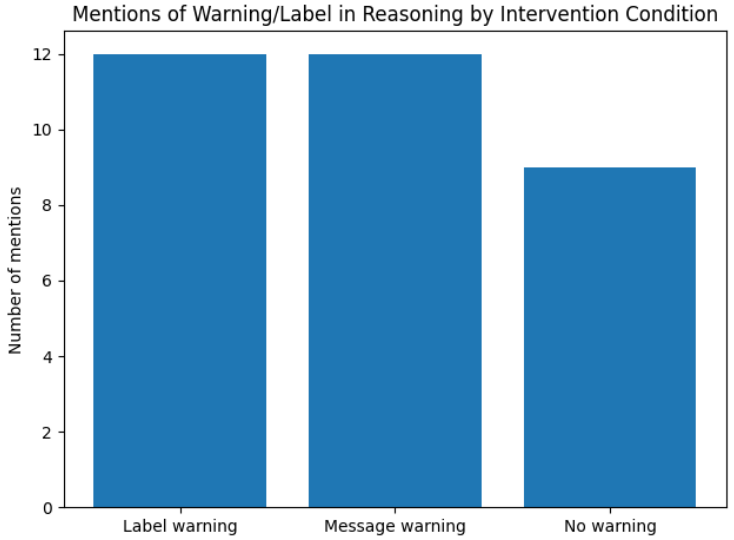
4-2. 有効性等に関する検証の個別詳細

オンライン実験を拡張したフィールドテストの実施（続）

- 警告ラベルや警告メッセージに明示的に言及して判断理由とした回答は全体の約0.4%にとどまり（7,580件のうち警告ラベル・警告メッセージ等に言及している回答：33件）、多くの参加者は介入表示そのものではなく、動画内容の論理性・証拠・既有知識に基づいて正確性を判断していると考えられる
- 言及の約 2/3 は英語圏（21/33）で、日本語では 1件のみと非英語圏ではかなり限定的
  → ラベルや警告を「判断理由として言語化する」傾向は英語圏で相対的にやや高いか
- 介入条件別では 弱いラベル群で12件、強いメッセージ群で12件、介入なし群9件と大きな差は見られず、多くの参加者が警告表示そのものを判断理由として明示的には用いていないことが示唆された



図：判断理由の自由記述における警告ラベル・警告メッセージへの言及数（国別、N=33）



図：判断理由の自由記述における警告表示への言及数（介入条件別、N=33）

介入なし条件でも9件の言及あるが、多くの場合、“message” を「動画の内容」という一般的意味で使用しているケースが含まれるため

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

- 本研究では、得られた知見を社会に広く還元し、偽・誤情報への対応力向上に寄与することを目的として、多角的な普及啓発活動を展開した。研究成果の発信にあたっては、専門家コミュニティへの学術的発信と、一般市民や実務者を対象とした社会的発信の双方を重視し、幅広い対象層に向けた取組を実施した。

5-2. 普及啓発活動の個別詳細

普及啓発活動に係る取組・成果

本研究プロジェクトでは、学術的成果の創出にとどまらず、その知見を広く社会へ還元し、偽誤情報対策の実効性を高めるための普及啓発活動を推進している。これまでの期間において、国内外の専門家・実務家との連携体制を構築（取組・成果）しており、今後はこれを基盤とした社会実装および情報発信（今後の展開）を以下の通り実施する

今後の普及啓発活動計画

上記で得られた知見とネットワークを活用し、論文投稿（2026年2月予定）以降、段階的に以下の活動を展開する。

社会への発信とリテラシー向上

情報公開: ウェブサイトおよびメディアを通じ、生成AI偽誤情報の実態や対策に関する解説資料を一般公開する

教育・啓発: 市民団体と協働したイベント支援や、真偽判断のポイントを解説するワークショップ・セミナーを開催する

教育・実務分野との連携

教材開発: 開発企業との連携を検討し、研究成果を反映したリテラシー教育用コンテンツの制作・提供を目指す

実務への橋渡し: デジタルプラットフォームやマーケティング業界に対し、研究に基づくガイドラインや知見の共有を継続する

学術的成果の還元

国内外の学会・シンポジウム等での発表を通じ、学術コミュニティへの知見共有と、エビデンスに基づく政策提言の基礎資料を提供する

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

1. **既存アプローチの限界と「信用失墜型」攻撃の脅威の検討**
   - 現在主流となっている控えめな生成AIラベルでは、ユーザーの認知的警戒心を高める効果は限定的であり、偽誤情報抑制の十分条件とはなり得ないことが明らかになった
   - 政治・メディア批判においては信用失墜型（Discrediting）の手法が多用されており、正しい情報さえも受け付けない土壌が形成されている。単なる「真偽判定」ではなく、情報の受容・評価・行動という認知過程そのものを対象化した設計が不可欠

2. **技術的介入とプレバンキング・教育的介入等多様な介入手法の融合（ハイブリッド・アプローチ）**
   - 警告表示の効果は一律ではない。性格特性や個人の価値観に応じたインターフェイスデザイン（UI）の最適化が必要
   - 強い警告（メッセージ型）の有効性を含め、ユーザーの属性に合わせた柔軟な介入設計が求められる。
   - 技術的な警告表示に加え、ユーザー自身の認知バイアスへ働きかける「教育・リテラシー」の両輪が必要
   - プラットフォームのデータ透明性を確保しつつ、技術と教育の相乗効果を狙う多層的な防御策を構築すべきである

3. **今後の展望：国際連携による「知見の社会実装」へ**
   - 国際共同研究（性格特性と耐性）、および国内ファクトチェック団体・事業者との連携により、学術的知見を実効性のある介入手法の実装に関してさらなる検討を行う
   - 本研究で得られた知見（国別差異、介入効果測定法など）を、論文発表やレポート発表等を通じて、プラットフォーム事業者や政策立案者へ還元

→ デジタル空間のウェルビーイングとレジリエンス向上へ寄与する

6-1. 研究・調査の総合的な考察

研究・調査の総合的な考察

研究ゴールとの対応整理

研究ゴール	明らかになったこと	未解明・課題
① 偽誤情報の実態と拡散構造をデータに基づき体系化	• 日本の言説空間では 信用失墜型（Discrediting）言説が多用される傾向を確認 • 偽誤情報は単なる虚偽情報ではなく 制度的信頼（メディア・政治・専門家）への攻撃として機能	• SNSアルゴリズムが拡散に与える因果影響 • 実際の拡散ネットワーク構造・ボットや組織的拡散の関与度 ※主因：プラットフォームデータへのアクセス制約
② 人間の認知特性を考慮した実証的な対策評価	• 生成AI警告表示は 動画の正確性評価を低下させる効果を確認 • しかし介入効果は 個人の特性（AI信頼・信念・性格特性）に依存 • 人は警告よりも 既存信念を根拠に判断する傾向 → 単純な警告表示は偽誤情報対策の十分条件ではない	• 「どちらとも言えない」の増加が熟慮なのか判断回避なのか • 警告表示が拡散行動に与える影響
③ 日本社会に適した効果的な介入手法の開発	• 一律の警告表示では効果が限定的 • 異質性を前提とした設計が必要示唆された方向性： • UI最適化・プレバンキング（事前接種） • 教育的介入との併用 → ハイブリッド型対策の必要性	• 個人特性に応じた パーソナライズ介入の倫理設計 • 実環境での 長期効果
④ 将来の高度化する脅威に対応できる知見の創出	• 偽誤情報問題は 内容ではなく構造の問題 • 技術的対策だけでは不十分で認知・制度・プラットフォームを含む多層的対策が必要	• 長期的な介入効果・警告疲労・慣れ • AI生成コンテンツ増加による 情報環境の構造変化

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

- データアクセスと透明性の確保：実証研究では実際の拡散構造やアルゴリズム影響は未解明、データへのアクセス制約があり全体像の把握に限界
- 識別力を高める介入の設計
  - 警告は確信度を下げるが、真偽識別力は高めない → どの介入設計が判断制度を上げるのか検証が必要
- 不確実性は望ましいのか？
  - どちらとも言えないの増加は熟慮？回避的な行動？拡散抑制につながる？ → 行動レベルでの追跡が必要
- 一律介入ではなくて異質性を前提とした設計が必要
  - 偽誤情報に脆弱な層や介入効果がある層などに分けて介入の設計や効果検証を行うことが必要
- 実環境での持続効果
  - 警告疲労・慣れ・長期的影響は未検証

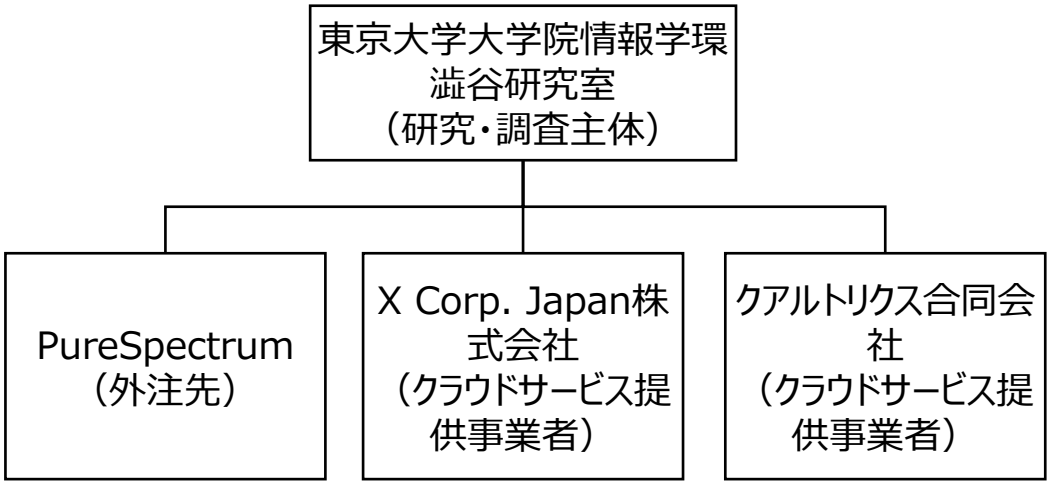
今後の展望

- プラットフォーム連携による実証基盤の構築
  - データ連携枠組み検討/アルゴリズム影響・拡散経路まで含めた因果検証
- 「識別力」を高める介入設計への転換：単なる警告表示から判断プロセスを支援する設計へ
- 不確実性の再評価：「適切な疑い」を支えるインターフェース設計
  - 拡散行動・共有意図・態度変容への長期影響を追跡
- 異質性を前提とした適応的設計：パーソナライズ警告の倫理的設計と脆弱層支援と公平性の両立
- 長期的・実環境での評価
  - 警告疲労・慣れ・逆効果の検証と実プラットフォームでのフィールド中長期実験

目次

7-1. 実施体制及び役割分担

本事業の実施体制図



各団体の役割・業務範囲

- 東京大学大学院情報学環澁谷研究室・本事業における統括責任者であり、本事業におけるプロジェクトマージメンと全般を行う
- PureSpecturm：実験被験者パネルの提供
- X Corp. Japan株式会社：XへのAPI接続
- クアルトリクス合同会社：研究・調査プラットフォーム（Qualtrics）の提供

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								
1.現在及び将来の偽誤情報の体系化(データ収集・解析・分類・整理)	→	→	→					
(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								
1.国内外の生成AI偽誤情報受容性の実証分析	→	→	→					
(3)成果報告書の作成								
1.成果報告書の作成				→	→	→	→	→
(4)普及啓発活動への協力								
1.論文執筆・学会／シンポジウム等での成果発表			→	→	→	→	→	→
2.各種メディア・業界団体等との連携による研究知見等の普及活動								→