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 概要版

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

生成AIによる情報操作の高度化と人間の認知的限界の交錯

- 文章・画像・動画の生成が容易化し、ディープフェイクなど精巧な情報操作が拡大
- 一般利用者が真偽を見分けることが極めて困難になっている。
- 情報量の爆発的増加により、人々は常に不完全情報下での判断を迫られている
- 人間の認知特性自体が偽情報への脆弱性となっている。

既存対策の限界と日本固有の課題

- AI検知や規制には効果があるが万能ではなく、利用者の判断プロセスを踏まえた対策が不可欠
- AI受容やメディア環境の特徴が他国と異なり、日本文脈に適合した対策研究が十分でない

上記課題を踏まえ目指す姿・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

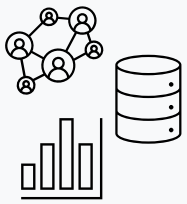
→ 生成AI時代における安全で信頼できる情報環境の実現に貢献

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

現状把握と実態解明

研究課題1

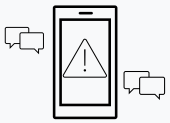


実態解明と可視化
偽誤情報の特徴・拡散パターンを体系化
誤判断を招く形式・手法をデータで解明

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

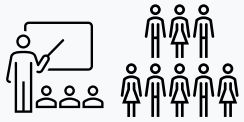


効果的な介入手法
警告ラベル等の介入の有効性を実証的に評価
ユーザー負担が少なく実効性の高い対策手法の設計

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

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

- 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）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
- 機械学習やネットワーク解析を用いて、頻出フレーズ、文体、拡散のために用いている手法などを分析し、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
- 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把握を目指す

【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、介入を組み込んだオンライン実験を設計する
- 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
- 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

2-2. 研究および有効性等に関する検証の個別詳細

研究項目1：現在及び将来の偽誤情報脅威の体系化

- 偽誤情報関連キーワードで収集したXデータ（2025年10-11月）で最も頻出したのは、政治・政党批判に関する言説、ついでメディア批判を主題とする言説などが続く
- 情報操作戦略として最も顕著であったのは信用失墜型（Discrediting）であり、発信元そのものの信頼性を攻撃するもの
  → 正誤の検証以前に信頼の基盤を侵食する構造を形成している可能性を示唆する

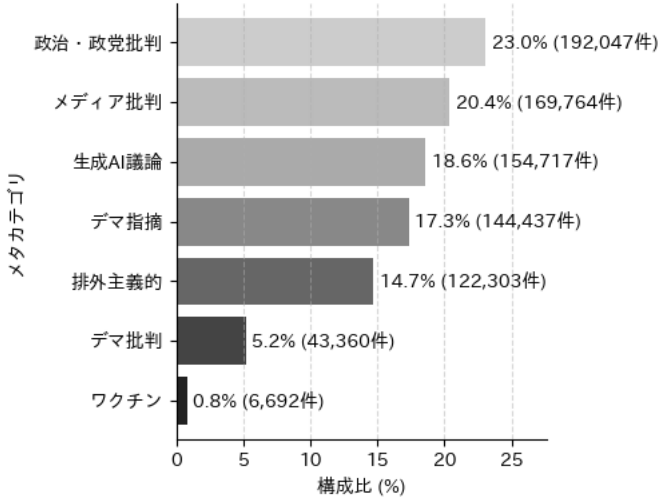


図. 分析期間中のデータから抽出したトピックのメタカテゴリごとの投稿割合

表. トピックごとに用いられている操作手法戦略等の例

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）」

2-2. 研究および有効性等に関する検証の個別詳細

研究項目1：現在及び将来の偽誤情報脅威の体系化

- 「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把握のため、対象・側面・評価の3軸分類体系を構築し、メタカテゴリごとにランダム抽出した100件を人手による3軸分類ラベリングを実施
- 言説の主要対象は 政府・政治およびメディア であり、制度的アクターへの言及が中心
- 評価は否定的なものが多数を占め、制度や情報源への不信が顕著
  - 政治・政党：約76%が否定評価（政策・政治行動への批判が中心）
  - メディア：約73%が否定評価（報道姿勢や情報操作への不信）
  - 生成AI：非評価が最多（40%）であり、評価が分化した過渡的段階

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件) / 個人 (13件) / 社会集団 (8件)	行為・対応 (59件) / 人となり・内的特性 (21件) / 主張の真偽 (19件)	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件) / 個人 (16件) / 対象不明 (16件)	行為・対応 (42件) / 人となり・内的特性 (26件) / 主張の真偽 (20件)	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件) / 社会集団 (16件) / 経済 (13件)	行為・対応 (49件) / 主張の真偽 (38件) / 人となり・内的特性 (12件)	科学技術に対する主張の真偽の非評価 (16件)

2-2. 研究および有効性等に関する検証の個別詳細

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

- 日米独仏印の5カ国でオンライン実験を実施（N=5,443）
- 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）に加え、ユーザーの迷い（不確実性）を測定
- 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

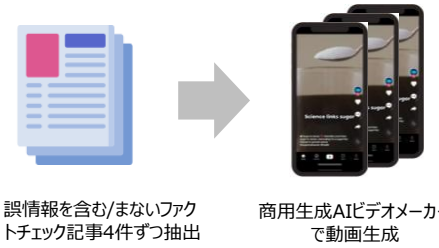


図. 実験で用いた動画の生成方法

図. 3条件の介入

表. 対象地域と被験者数

国	被験者数
米国	1,054
ドイツ	1,050
フランス	1,063
日本	1,285
インドネシア	991
合計	5,443

註：質問紙の前半・後半にアテンションチェック問題を3問配置し、うち2問以上に適切に回答できなかった被験者は対象被験者から除外。表中被験者数は除外後の数

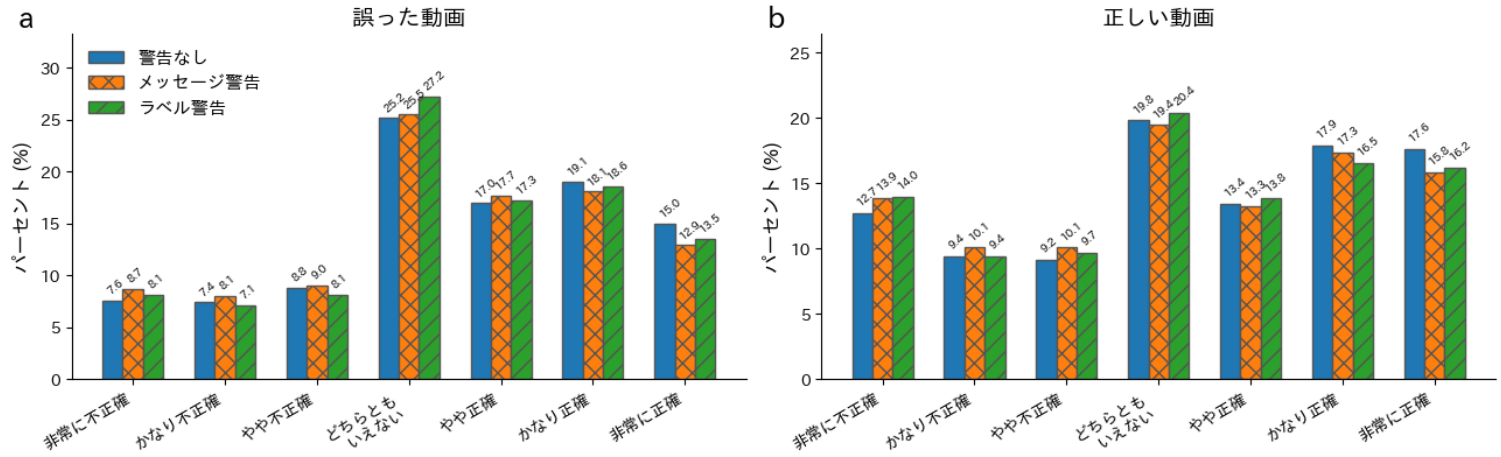


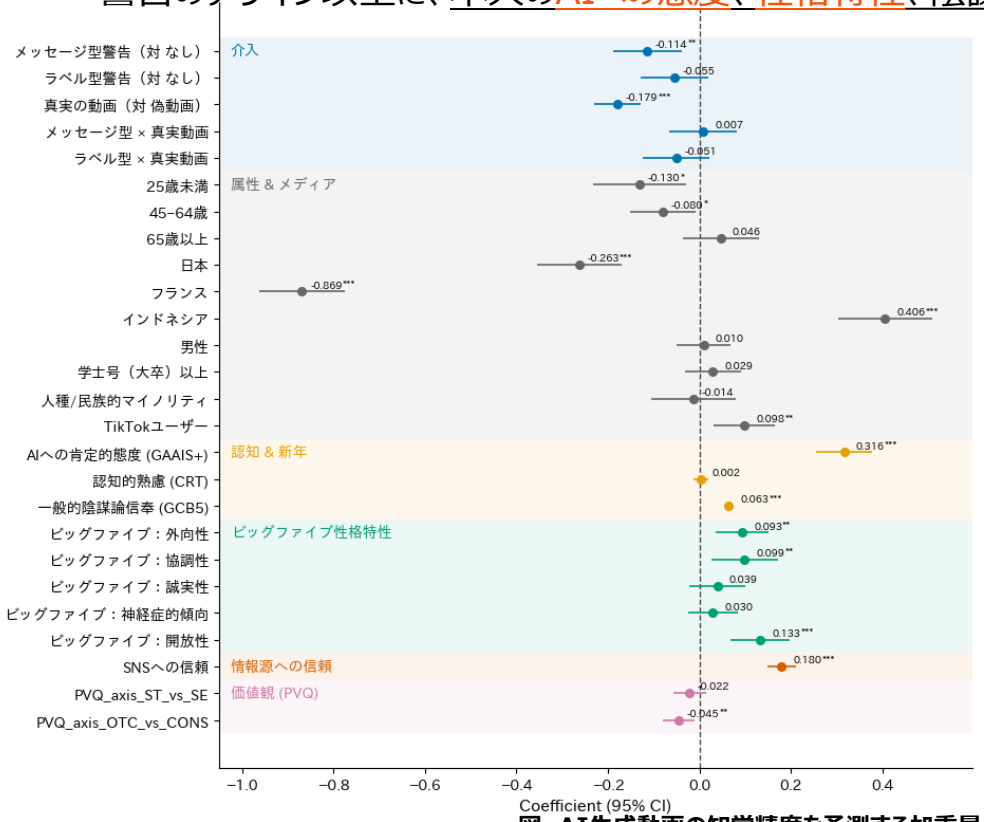
図. 生成AI動画に対する正確性評価の分布

2-2. 研究および有効性等に関する検証の個別詳細

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 動画の正確性評価に影響は与えない
- 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



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

2-2. 研究および有効性等に関する検証の個別詳細

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

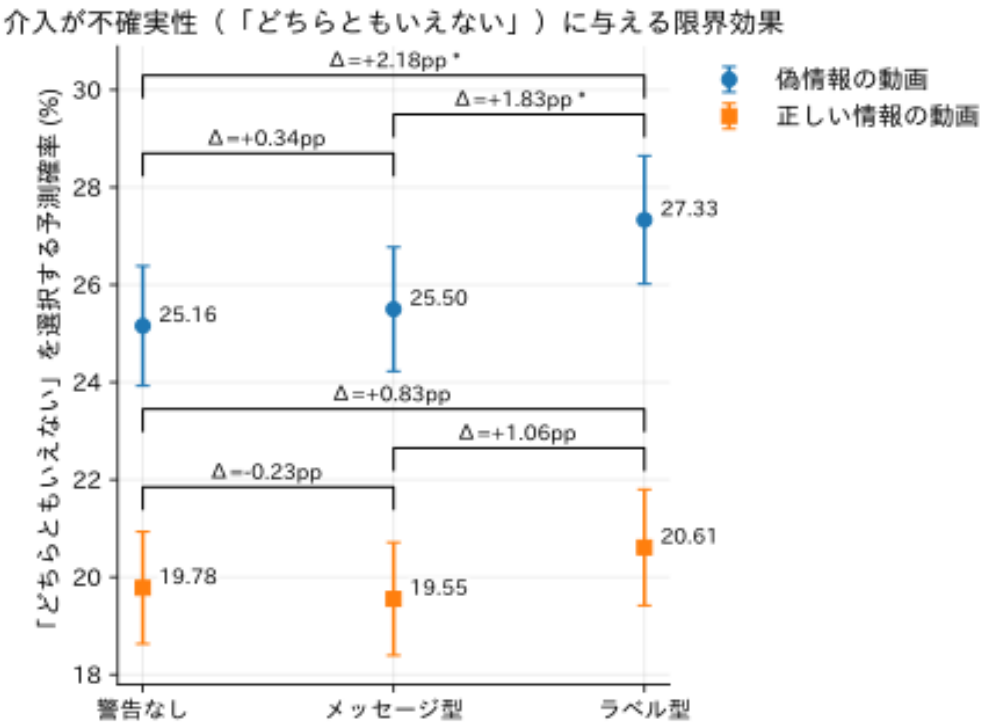


図. 介入が不確実性（「どちらとも言えない」）に与える限界効果

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

3-1. 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可能性
   - 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

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

- 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

上記課題を踏まえた今後の展望

- 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