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 簡易版

2026/3/19

研03_東京大学大学院情報学環

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

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。