2026年2月16日

第1回宮城県選挙期間中の情報流通の諸課題への対処に関する検討会

生成AI時代の偽誤情報

東京大学大学院情報学環

澁谷遊野

私たちに必要な考え方

- AIでもっともらしい文章・画像・動画が誰でも作れる時代
- 大事なのは正解・誤り探しではなく、**騙されにくい構造を知る**こと

情報環境は大きく変化

- ソーシャルメディアは感情的な刺激や不安・怒りを誘う内容を拡散
- 私たちは毎日、検証しきれない量の情報に触れている

→ 問題は**偽誤情報の多さ**だけでなく、**人間の判断の限界**が突かれていること

偽誤情報はどう作られるか

よく使われる情報操作手法(DEPICT)

- 信用を失わせる(Discrediting)
- 感情に訴える(Emotion)
- 分断する/敵味方化(Polarization)
- なりすまし(Inpersonation)
- 陰謀化(Conspiracy)
- あらし(Trolling)

→ 日本語データ(2025年10-11月)では、「信用失墜」がよく見られる戦略の一つ

事実を否定するより、事実を伝える仕組みへの信頼を壊す

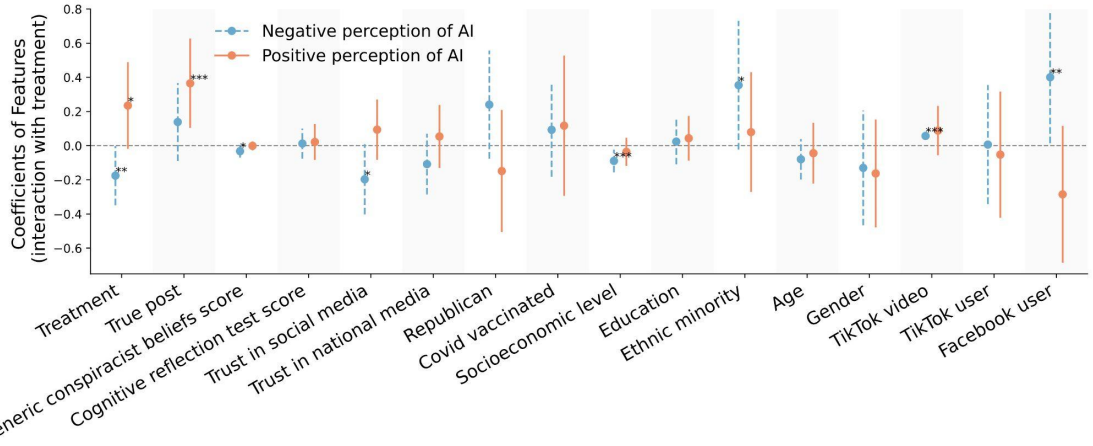
選挙と情報の空白

- ソーシャルメディアが情報の空白を補完
- 例） 兵庫県知事選
  - ショート動画（短尺）の閲覧数よりも長尺動画の閲覧数が伸びる

→ ユーザーは長尺動画を通じて政策内容を深く理解しようとする傾向か

介入効果は一様ではない；脆弱な層への重点的アプローチが必要

- 生成AIが動画を作ったことを示す警告表示は一定の効果を持つが、効果は個人の特性によって異なる　→ **一律の警告だけでは不十分**
- 人は警告よりも**すでに持っている信念**に基づいて判断する



Shibuya, Y., Nakazato, T., & Takagi, S. (2025). How do people evaluate the accuracy of video posts when a warning indicates they were generated by AI?. International Journal of Human-Computer Studies, 199, 103485.

個人だけの問題ではない

- 収益化構造
- アルゴリズム設計
- 制度的信頼への攻撃

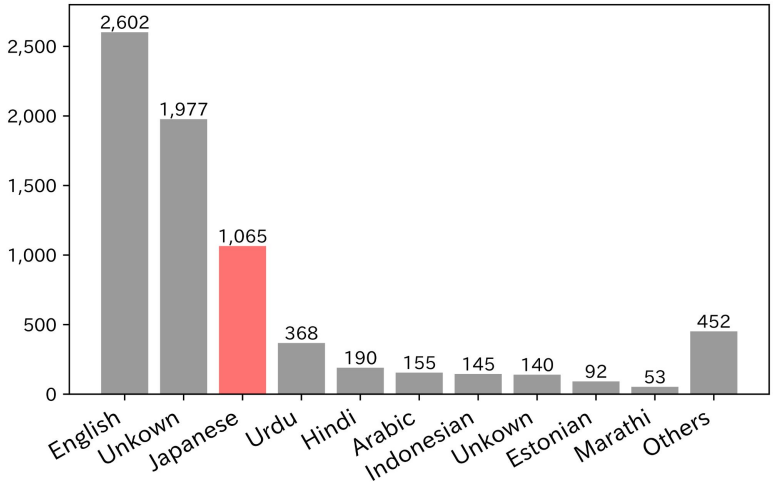
→ 偽誤情報は**構造**の問題でもある

→ 偽誤情報は内容よりも**仕組み**で広がる

収益化構造が拡散を歪める

2024年能登半島地震（X収益化後、初の大規模国内災害）では、

- 閲覧数稼ぎ目的とみられる虚偽・誇張・コピペ投稿が多数確認された
- 日本語話者以外と推定されるアカウントによるコピペ投稿が、全体の約**85%**



図：能登半島地震時複製投稿ユーザーの推定使用言語[2]
図中“Unknown”はプロフィール文の記載がないか記号や句読点，絵文字等のみで構成されるユーザの投稿数。

まとめ

偽誤情報対策とは

- 誤った情報との戦いではない
  - すべての偽誤情報を消すことではない
  - 単なる真偽判定ではなく**人がどう受け取り、どう判断し、どう行動するか**から設計する
- **認知と信頼**を守る設計
  - 事前の接種（prebunking／inoculation）によって耐性を育てること
  - 全員に同じ注意を促すのではなく、脆弱な層を取り残さない

→ 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

偽誤情報対応のため中長期的に取り組むべきこと

- 偽誤情報対策には特効薬はなく、多種多様なアクターが自分ごととして捉え・協力し合う**多面的・多層的なアプローチ**が求められる
- 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
  - 研究機関・外部団体など第三者が、偽誤情報の実態把握やその対応策の効果検証などを行えるかたちで、**データの提供やアルゴリズムの開示が行われることが望ましい**