Institute of SCIENCE TOKYO

WebDB夏のワークショップ2025

フェイクニュース拡散のメカニズムと対策：計算社会科学の観点から

笹原 和俊

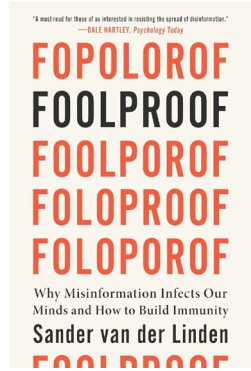
東京科学大学 環境・社会理工学院

自己紹介

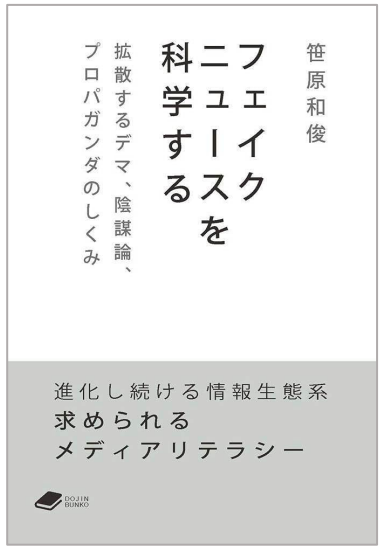
- 1976　福島県いわき市生まれ
- 2005　東京大学 大学院総合文化研究科修了（博士（学術））
- 2012～2020　名古屋大学 大学院情報学研究科 助教・講師
- 2016～2020　JSTさきがけ研究者（兼任）
- 2020～2024　東京工業大学 環境・社会理工学院 准教授・教授
- 海外経験　UCLA, Indiana University
- 現在　東京科学大学 環境・社会理工学院 教授、系・課程主任
  国立情報学研究所 客員教授
- 研究　計算社会科学
- 主な受賞　第23回ドコモ・モバイル・サイエンス賞優秀賞（社会科学部門）

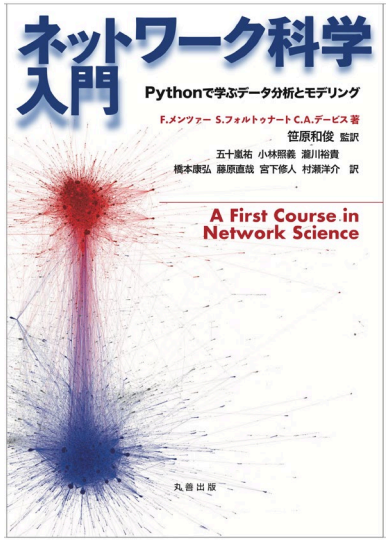
主な著書・訳書

監訳作業中…



令和7年国語教科書（三省堂）に一部掲載





発表内容

1. フェイクニュースの拡散
2. エコーチェンバーとフィルターバブルという環境要因
3. AIで高度するフェイクニュース
4. フェイクニュースの拡散抑止のための介入技術
5. まとめ

発表内容

1. フェイクニュースの拡散
2. エコーチェンバーとフィルターバブルという環境要因
3. AIで高度するフェイクニュース
4. フェイクニュースの拡散抑止のための介入技術
5. まとめ

フェイクニュースの氾濫

We are living in uncertain, confusing times – when it can be hard to know what to believe

Fake news

Internet hoaxes

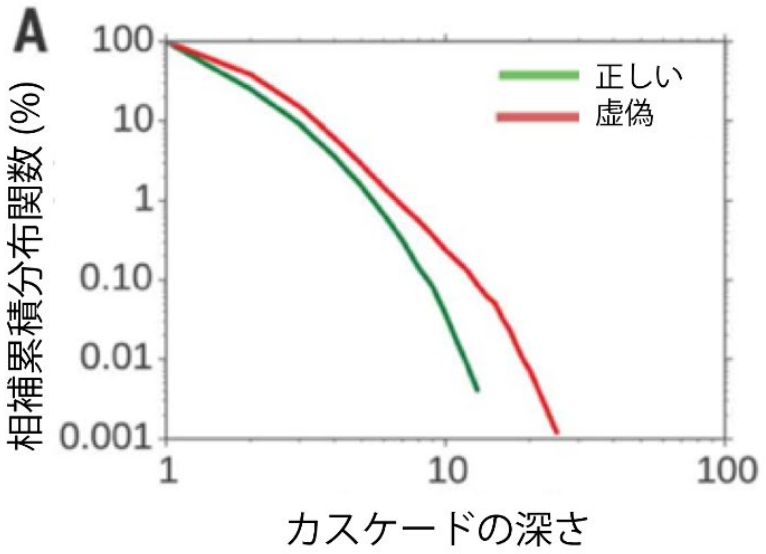
'alternative facts'

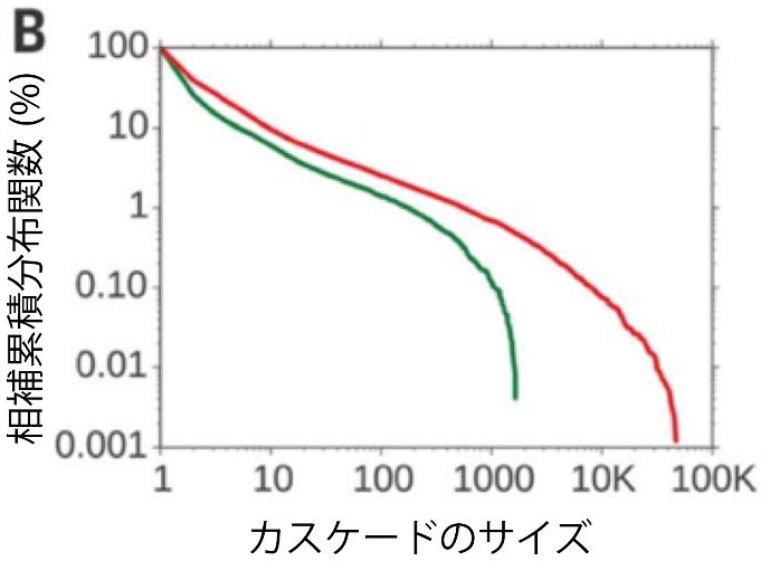
'post truth'

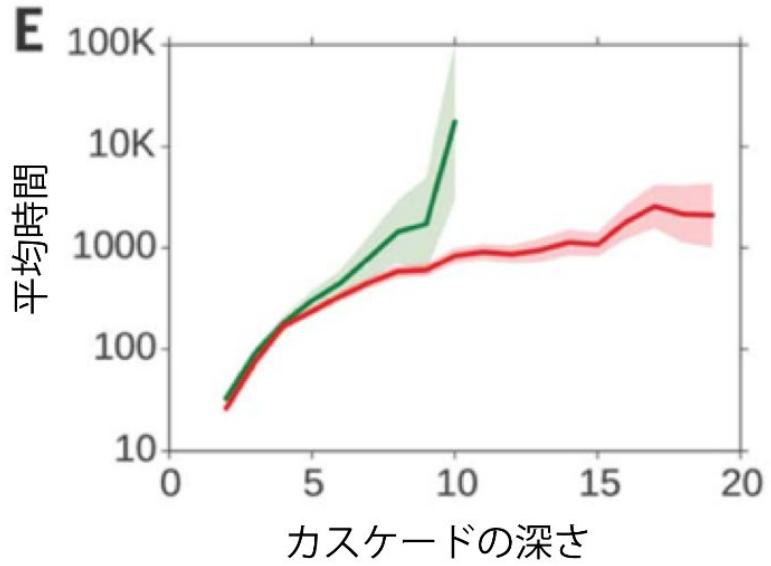
BBC NEWS

https://news.yahoo.co.jp/newshack/newshack/bbc_fakenews.html

偽ニュースは速く遠くまでたくさん伝わる





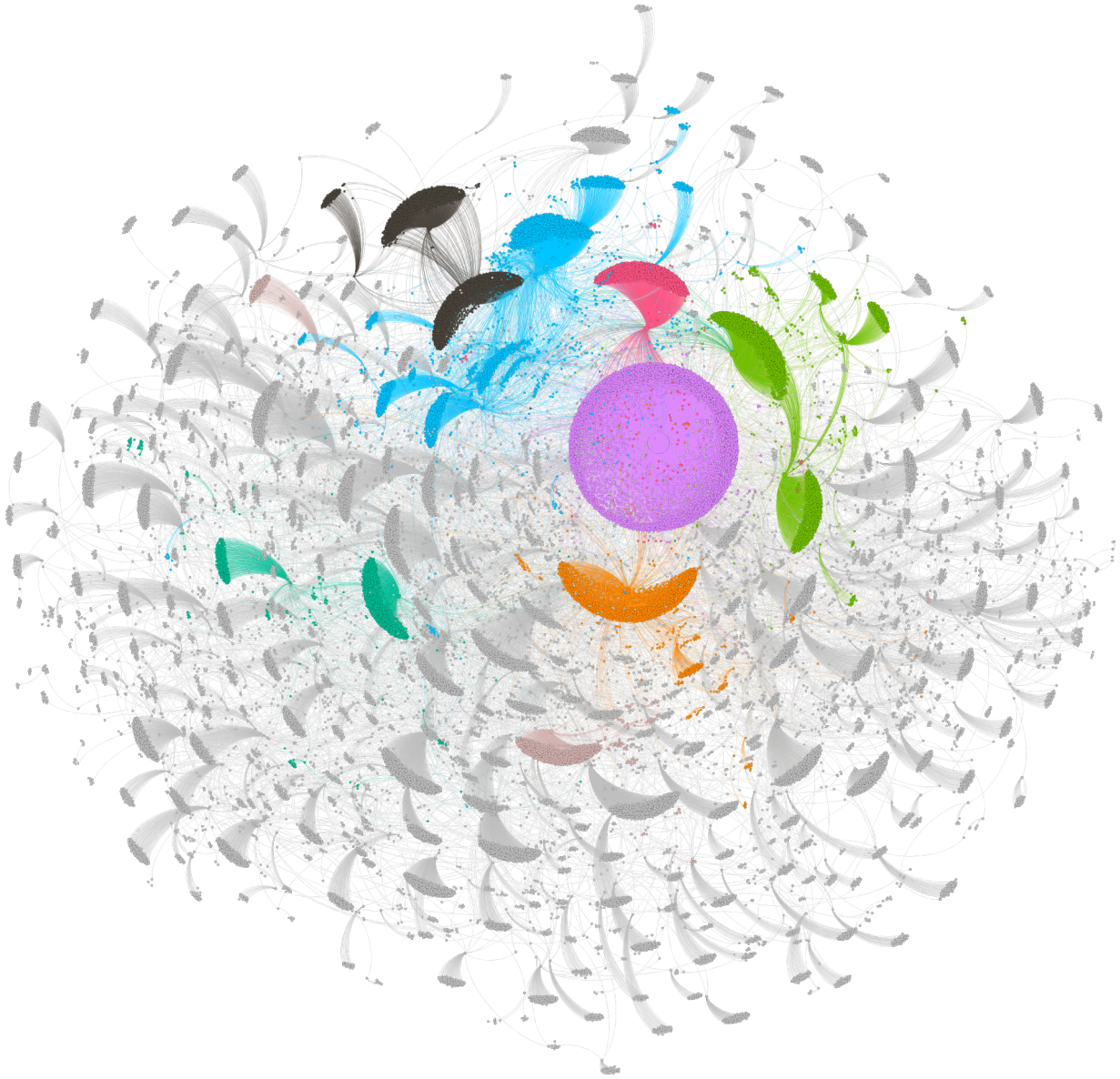


拡散されやすい話題

- 政治
- 都市伝説
- ビジネス
- テロ・戦争
- 科学
- エンタメ
- 自然災害

Vosoughi et al. Science, 2018

新型コロナ・インフォデミック



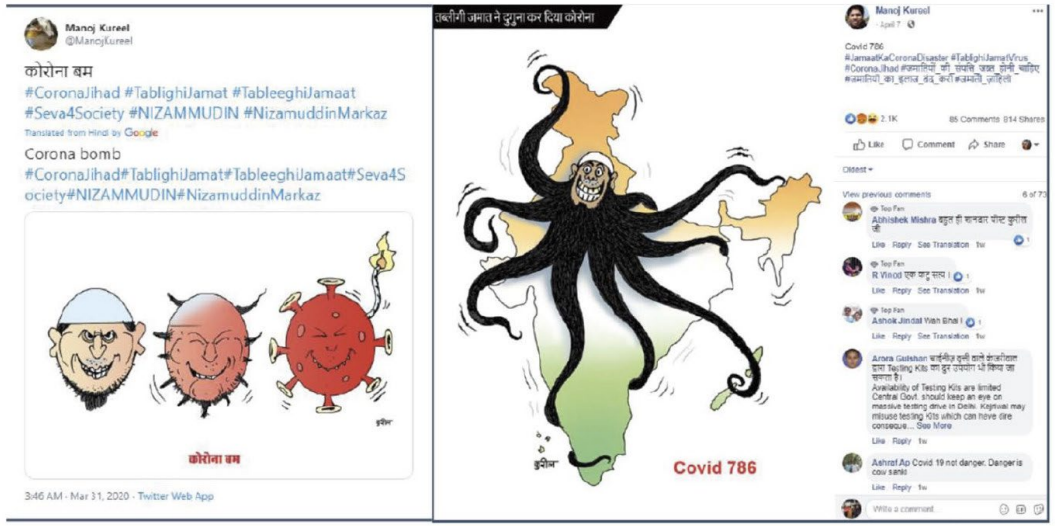
確かな情報と不確かな情報が混在する情報過多

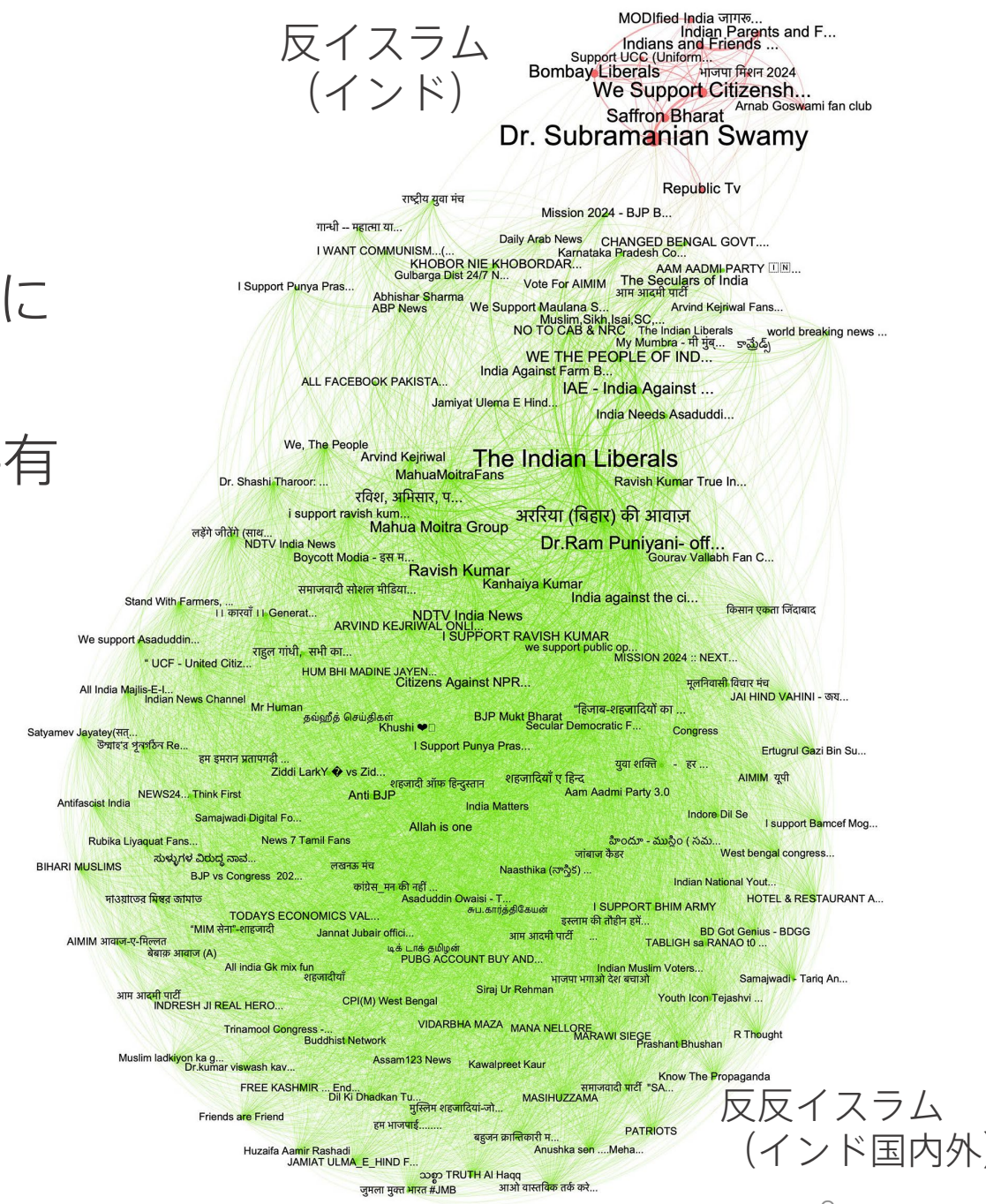
笹原和俊「デマや陰謀論はなぜネット上を拡散するのか」現代化学 (2020)

ヘイトの拡散

イスラム教のタブリーグ・ジャマート集会に関するFacebookの投稿の共有

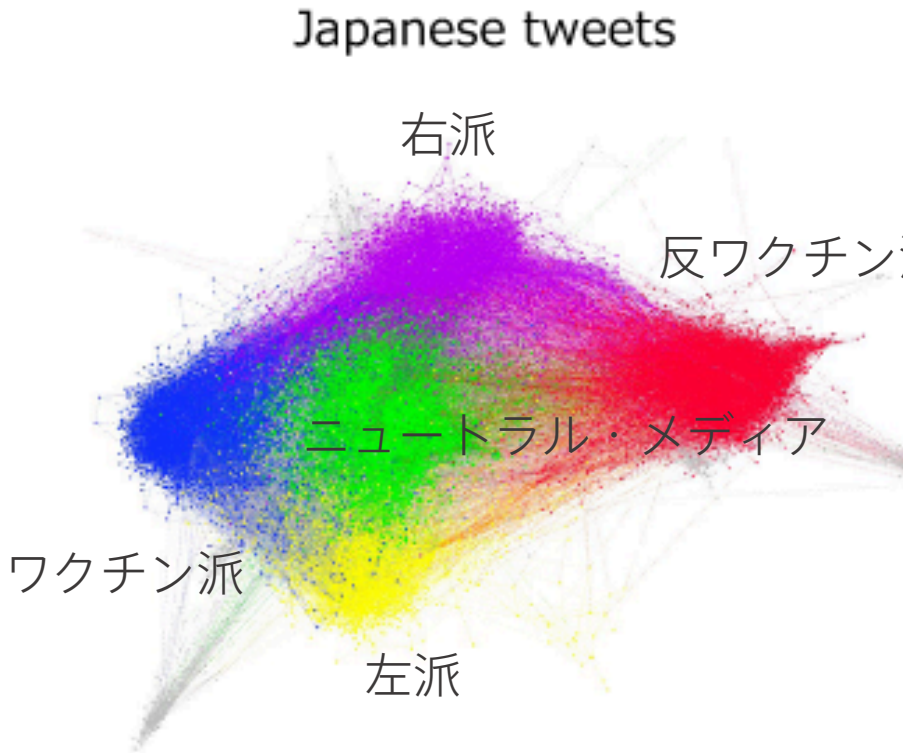
- 反イスラムは専らヘイト（偽情報）を共有
- 反イスラムの投稿は反反イスラムの3倍速く拡散



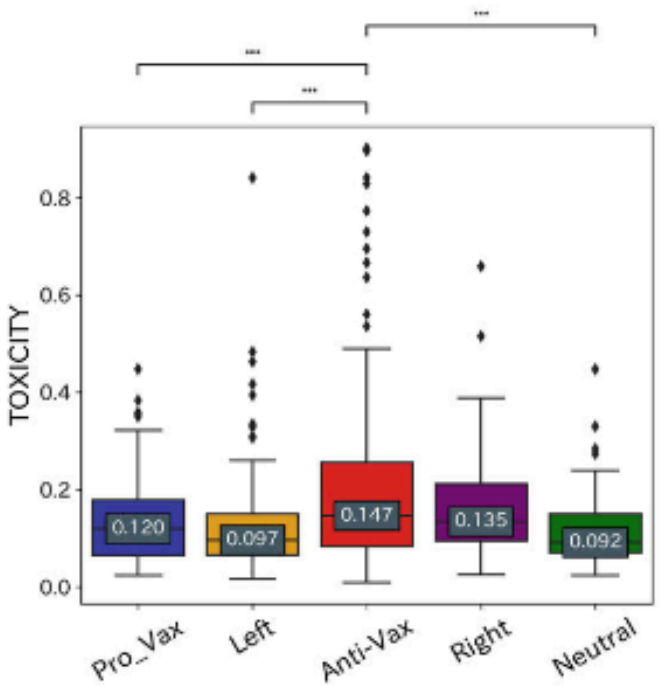


P. Ghasiya and K. Sasahara, Social Media and Society (2022)

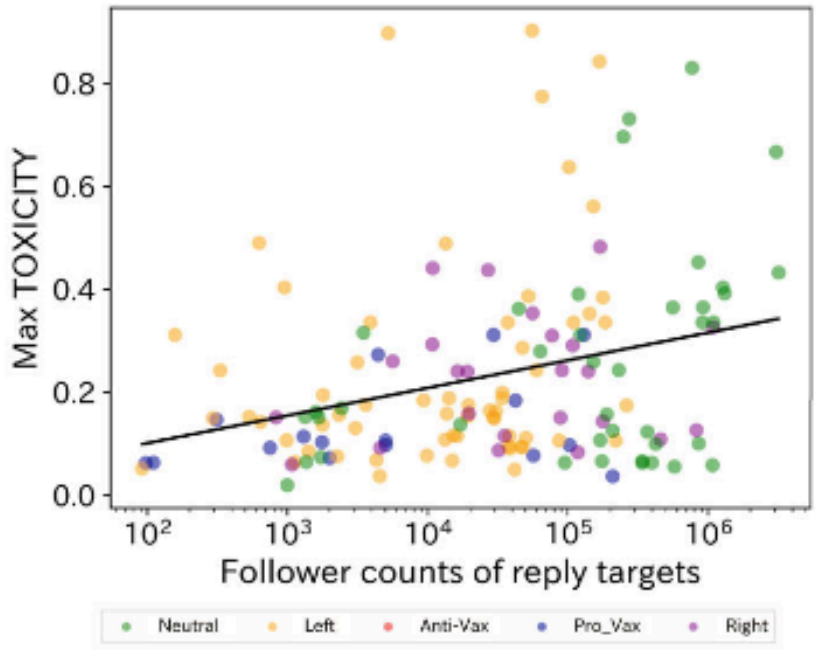
反ワクチン運動



反ワクチンの投稿の毒性は高い



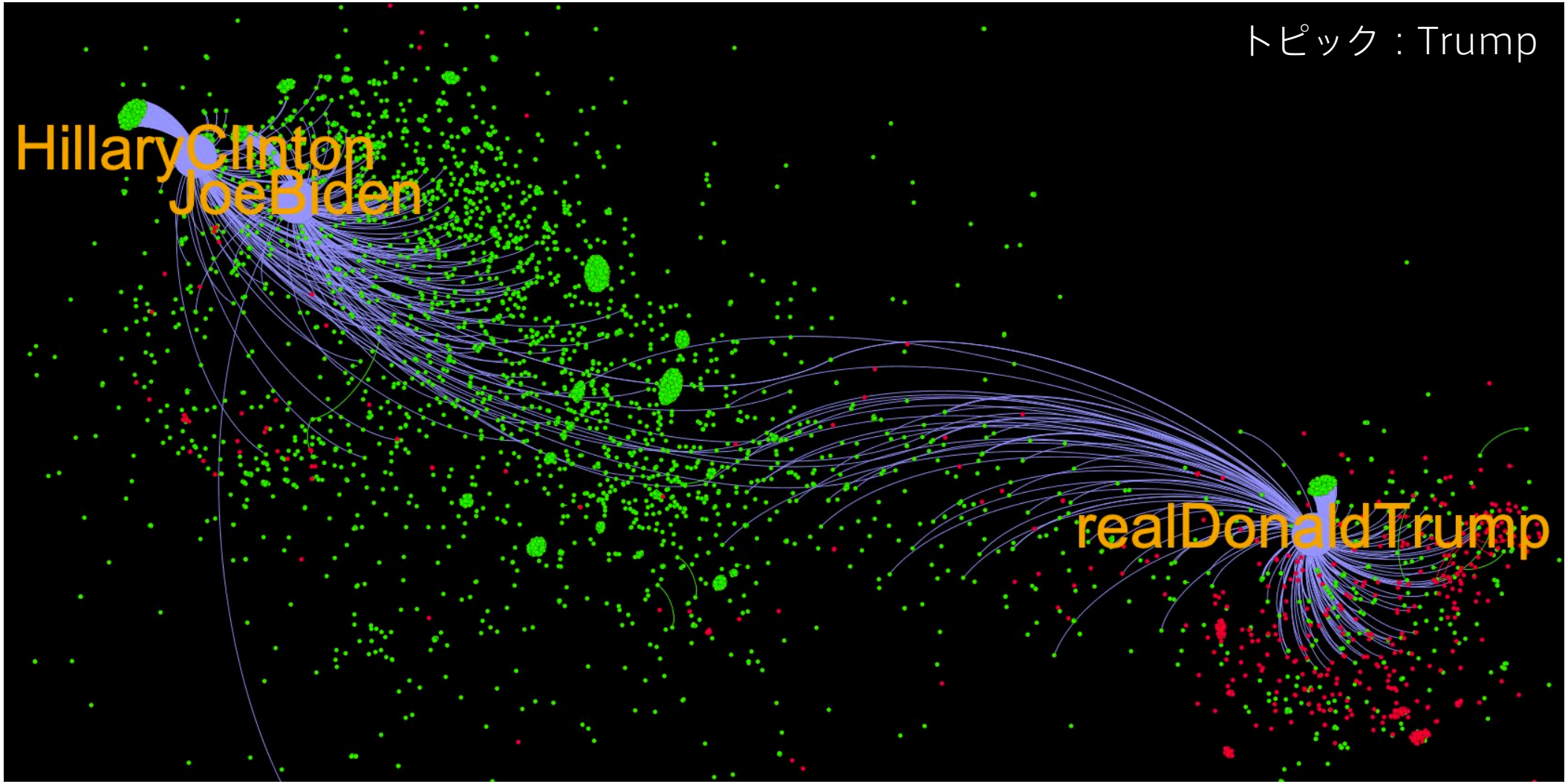
フォロワー数が多いほど毒性が高い
リプライを受け取る



K. Miyazaki, T. Uchiba., K. Tanaka and K. Sasahara, Humanities and Social Sciences Communications (2022)

陰謀論を増幅するBot

赤：悪質なBot
緑：普通のBot



W. Xu and K. Sasahara, Journal of Computational Social Science (2022) ← 米大統領経済報告2023で引用

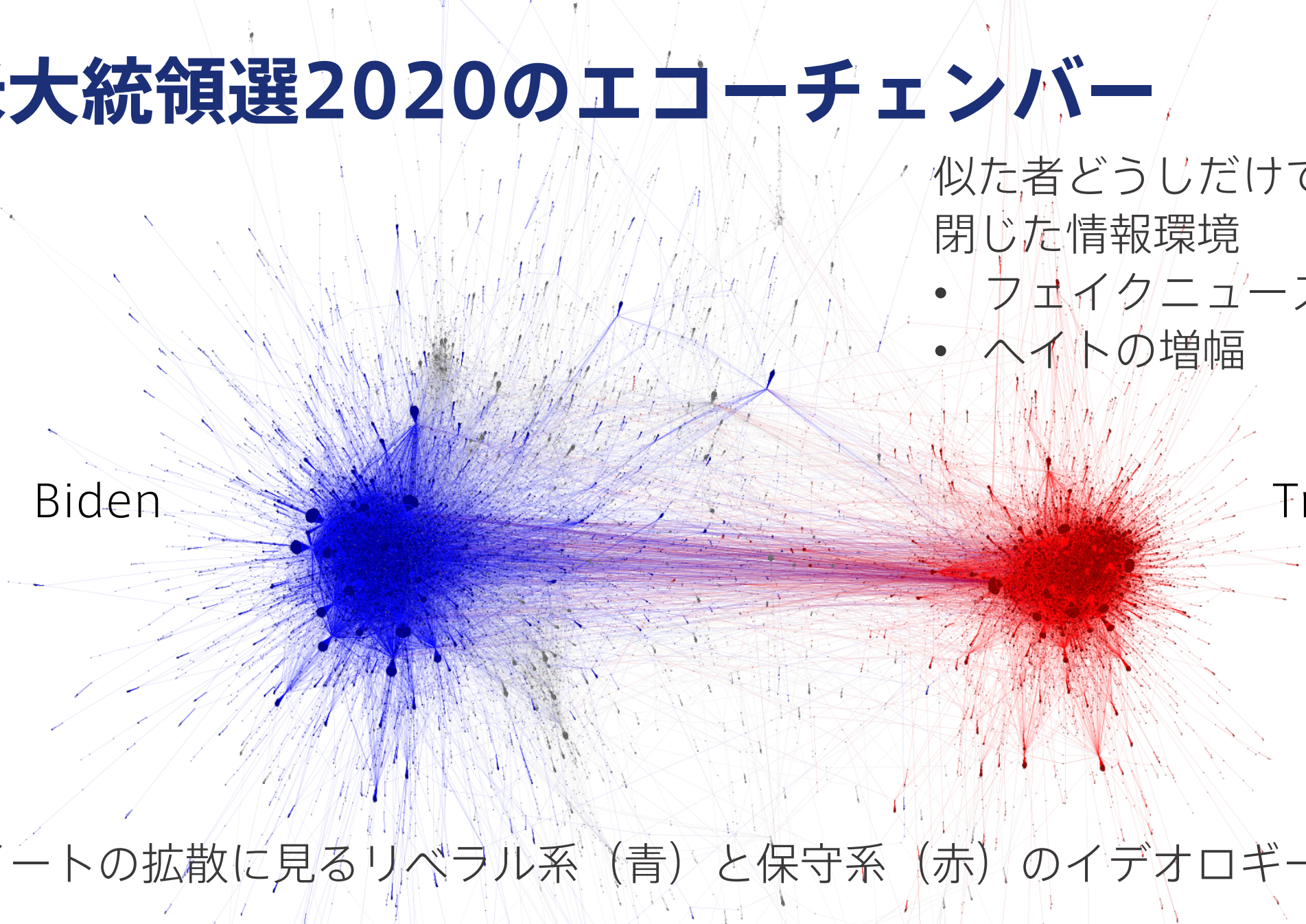
発表内容

1. フェイクニュースの拡散
2. エコーチェンバーとフィルターバブルという環境要因
3. AIで高度するフェイクニュース
4. フェイクニュースの拡散抑止のための介入技術
5. まとめ

米大統領選2020のエコーチェンバー

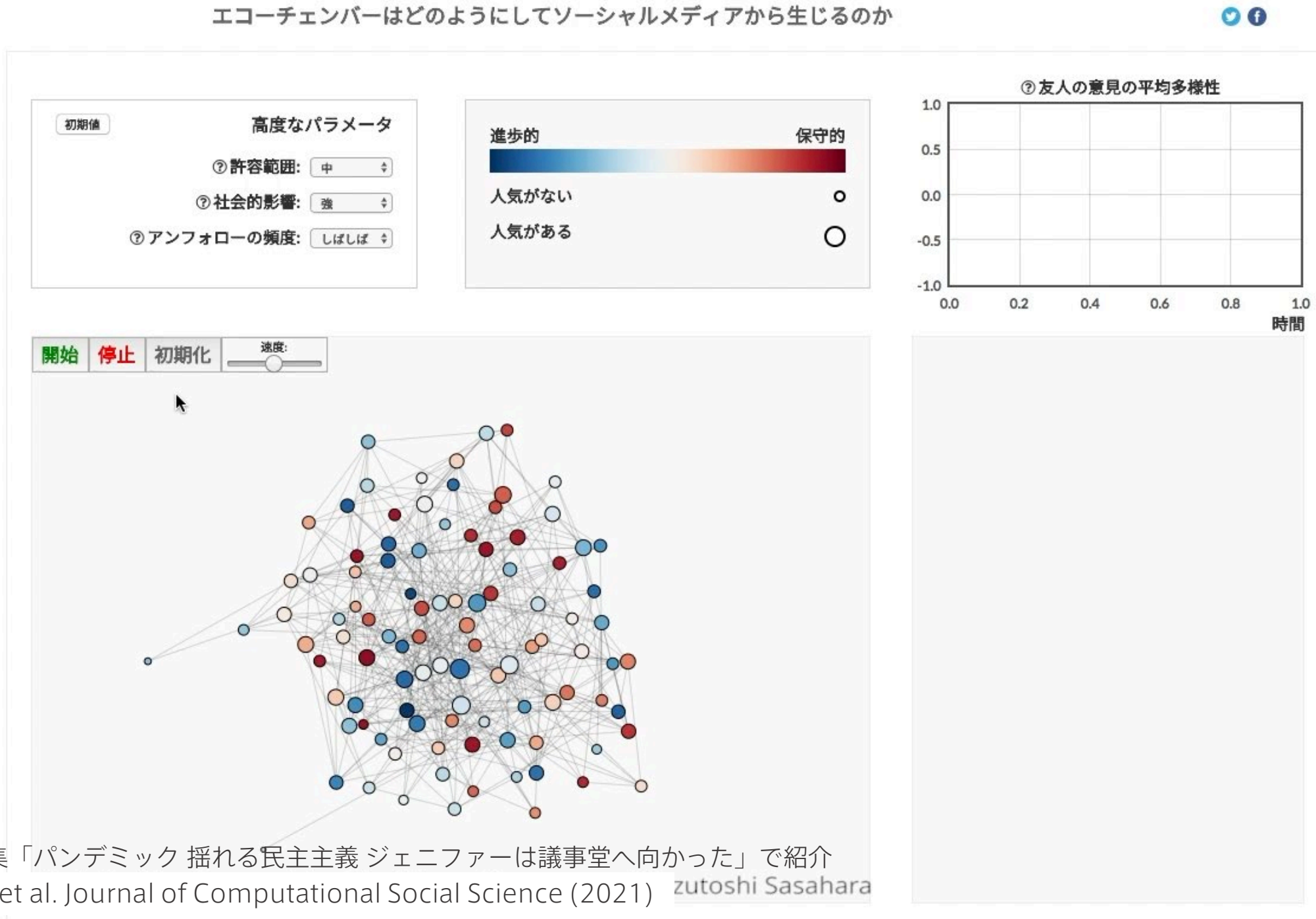
似た者どうしだけでつながった
閉じた情報環境

- フェイクニュースの温床
- ヘイトの増幅



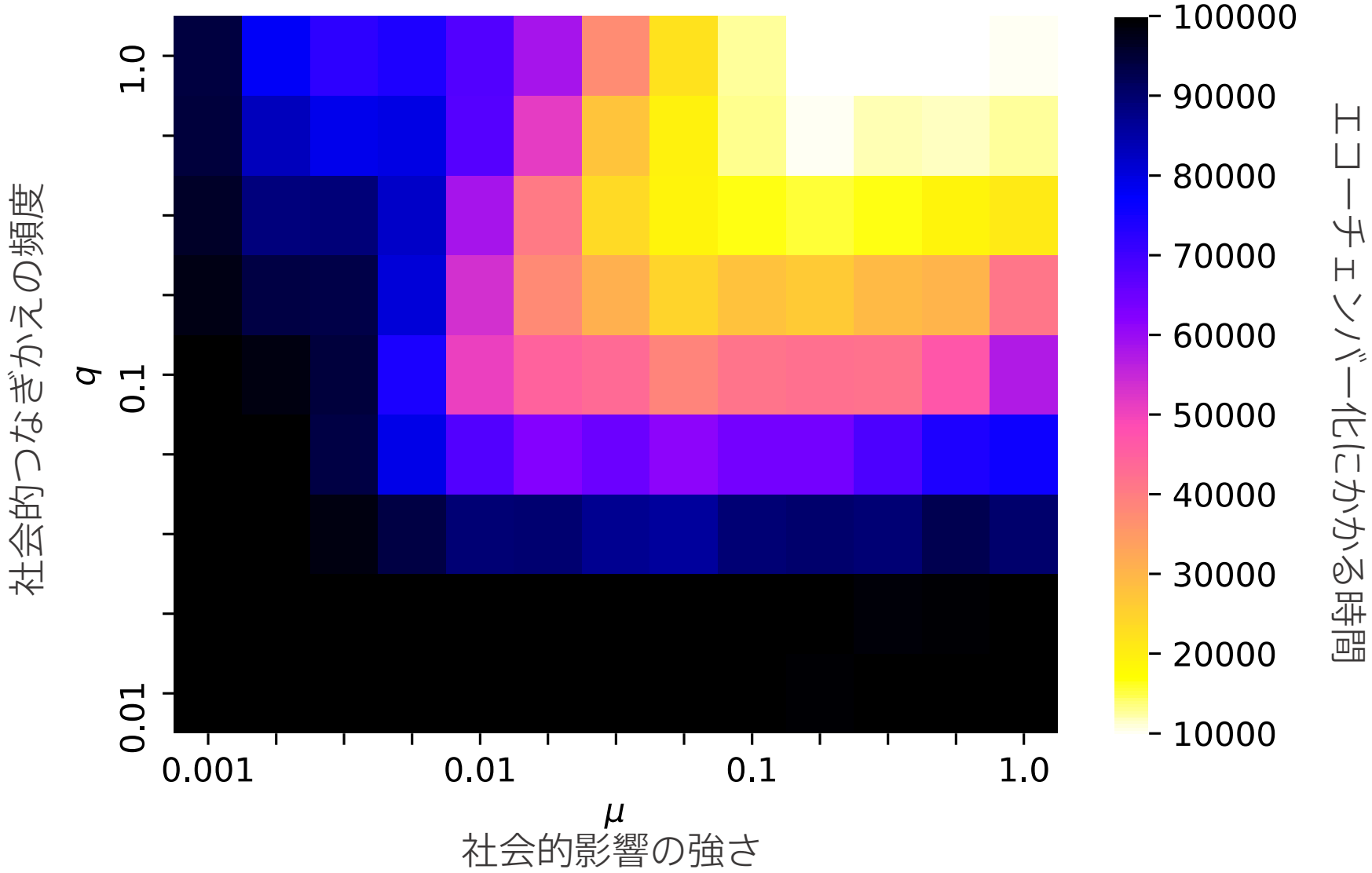
ツイートの拡散に見るリベラル系（青）と保守系（赤）のイデオロギーの分断

日本経済新聞電子版（2020.10.30, 笹原協力）



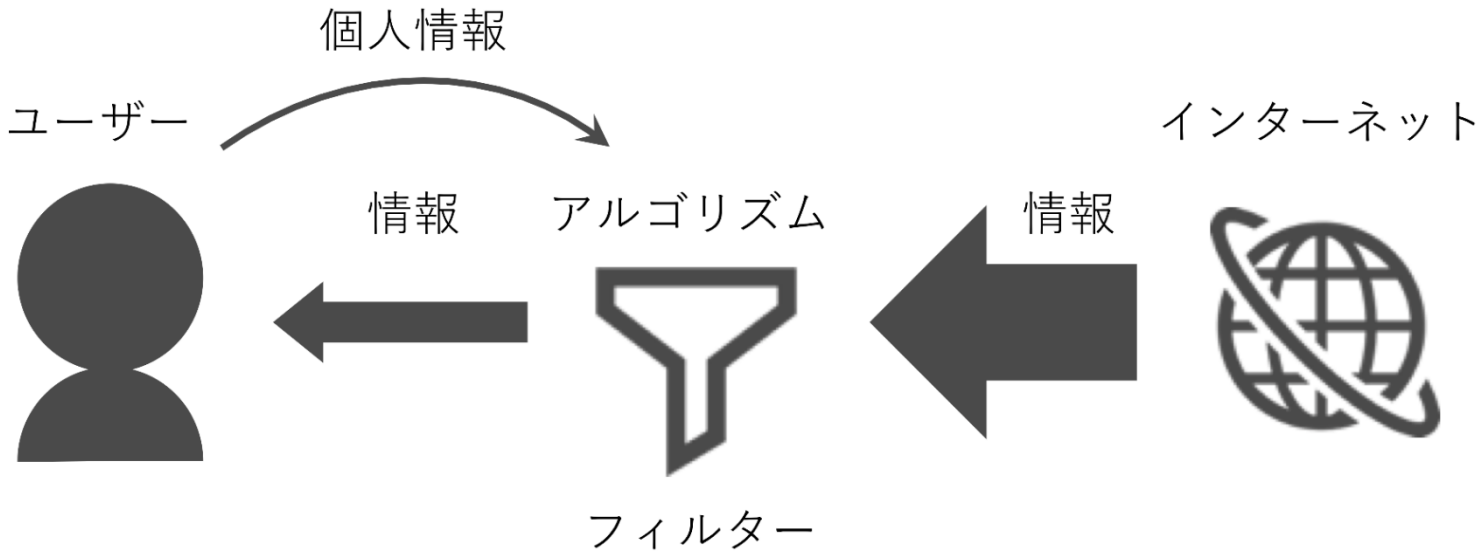
NHK ETV特集「パンデミック 揺れる民主主義 ジェニファーは議事堂へ向かった」で紹介
K. Sasahara et al. Journal of Computational Social Science (2021)

SNSはエコーチェンバーを加速する



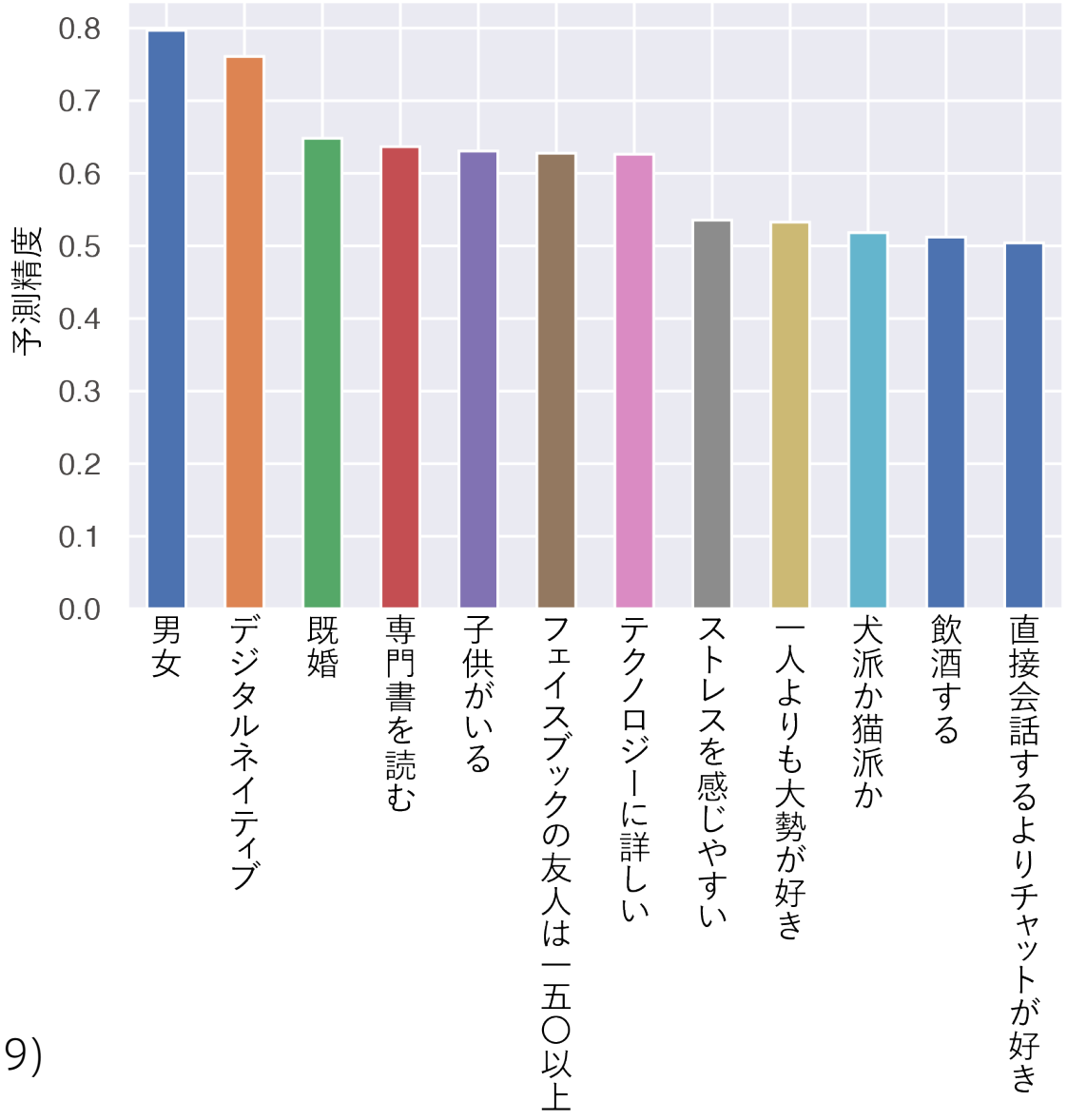
K. Sasahara et al. Journal of Computational Social Science (2020)

フィルターバブル



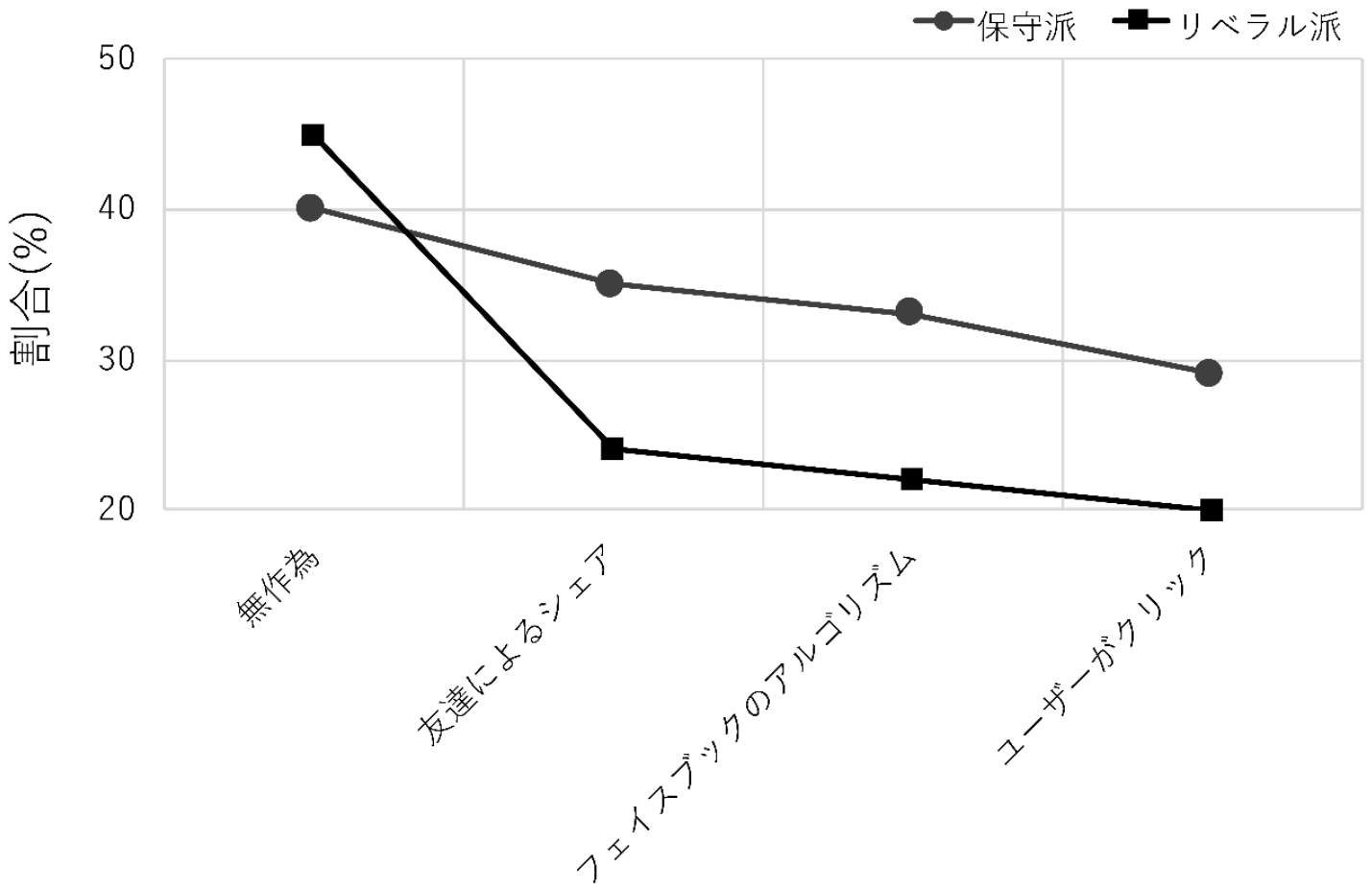
ユーザの個人情報を学習したアルゴリズムによって、その人にとって興味関心がありそうな情報ばかりがやってくる情報環境

ツイートが運ぶ個人属性



T. Yo and K. Sasahara (2019)

アルゴリズムは悪さをしない？



政治的に多様なニュースをフィルターしているのは、FBのアルゴリズムではなく、あなたの友人たち（社会的ネットワーク）

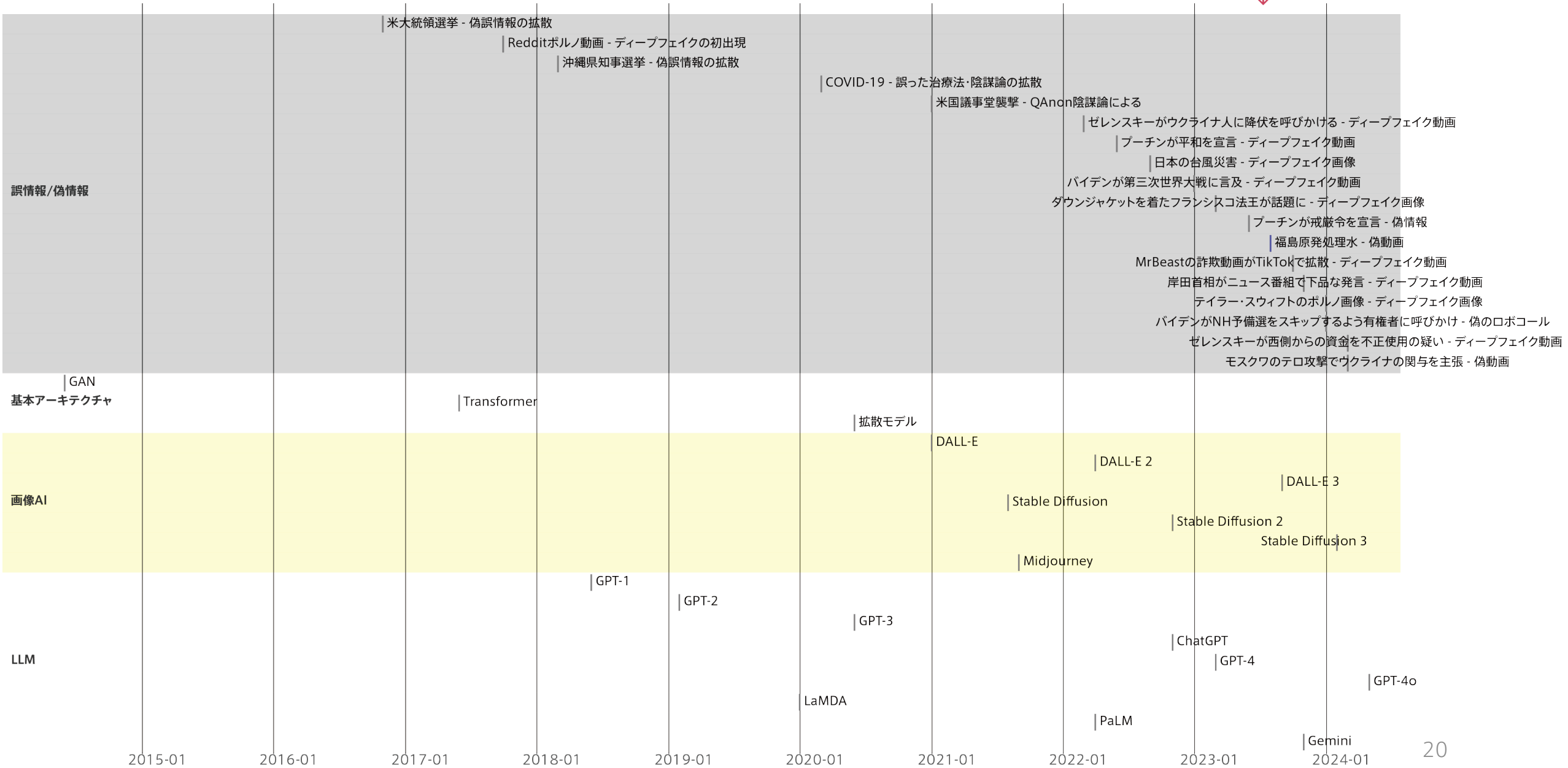
Bakshy et al. Science (2015)を元に作成

発表内容

1. フェイクニュースの拡散
2. エコーチェンバーとフィルターバブルという環境要因
3. AIで高度するフェイクニュース
4. フェイクニュースの拡散抑止のための介入技術
5. まとめ

AIと”フェイク”の共進化

近年、ディープフェイクが増加



HARNESSING AI TO AUGMENT DISINFORMATION CAMPAIGNS

Governments have long employed human commentators—whether state officials, hired contractors, or party loyalists—to covertly manipulate information online. Generative AI technology is slowly beginning to enhance such distortion campaigns.

政治・社会問題に関する情報を歪曲するために、画像、テキスト、音声を生成するAIツールを16カ国が使用

Freedom House. The Repressive Power of Artificial Intelligence (2023)

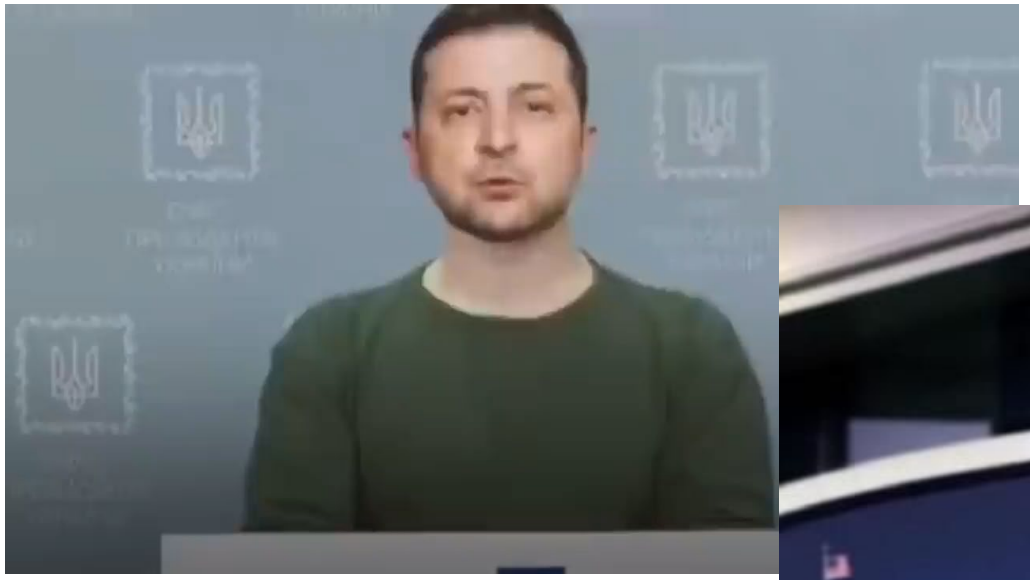
生成AIを利用した多言語による世論操作

国	工作・企業の名称	活動内容
ロシア	Bad Grammar	ウクライナ、モルドバ、バルト諸国、米国を標的に投稿を生成し、Telegramで拡散
ロシア	Doppelganger	ウクライナ紛争を巡り、ウクライナや西側諸国の批判・ロシア支持の投稿を生成し、Xや9GAGで拡散
中国	Spamouflage	福島の処理水放出を非難する投稿を生成し、X、Medium、Blogspot、アメブロで拡散
イラン	International Union of Virtual Media (IUVM)	米国やイスラエルを批判、パレスチナ人を支持する記事を生成し、Webサイトに投稿
イスラエル	STOIC	ガザ情勢を巡り、イスラエル寄りの記事や対立を煽る投稿生成し、X、Facebook、Instagramで拡散

https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations/ （2024/5/30）を元に作成

政治利用されるディープフェイク

政敵への攻撃、印象操作、民主主義の混乱、
外国からの干渉、風刺



2022/3/16

2023/6/5

2023/11/2

発表内容

1. フェイクニュースの拡散
2. エコーチェンバーとフィルターバブルという環境要因
3. AIで高度するフェイクニュース
4. フェイクニュースの拡散抑止のための介入技術
5. まとめ

nature human behaviour

Review article

https://doi.org/10.1038/s41562-024-01881-0

Toolbox of individual-level interventions against online misinformation

Received: 1 February 2023

Accepted: 5 April 2024

Published online: 13 May 2024

Check for updates

個人的介入の種類

- ナッジ：
  正確性のプロンプト、摩擦、社会的規範
- ブースト・教育的介入：
  接種理論、横読みと検証、メディアリテラシーのヒント
- 反論戦略：
  プレバンキングとデバンキング、警告とファクトチェックのラベル、出典信頼性のラベル

Info Interventions

Info Interventions: A set of approa
informed by behavioral science res
validated by digital experiments, to
resilience to online harms.

https://interventions.withgoogle

HOW IT WORKS

1

The individual scrolls through their social feed and comes across content with potential misinformation.

2

An accuracy prompt is triggered and pops up over the content.

3

A bite-sized explanation on why they are seeing the reminder is served to the individual and their attention is shifted to the accuracy of the content with information literacy tips.

4

The individual is now prompted to be more aware and may think twice when coming across similar content in their feed.

FINDINGS

50%

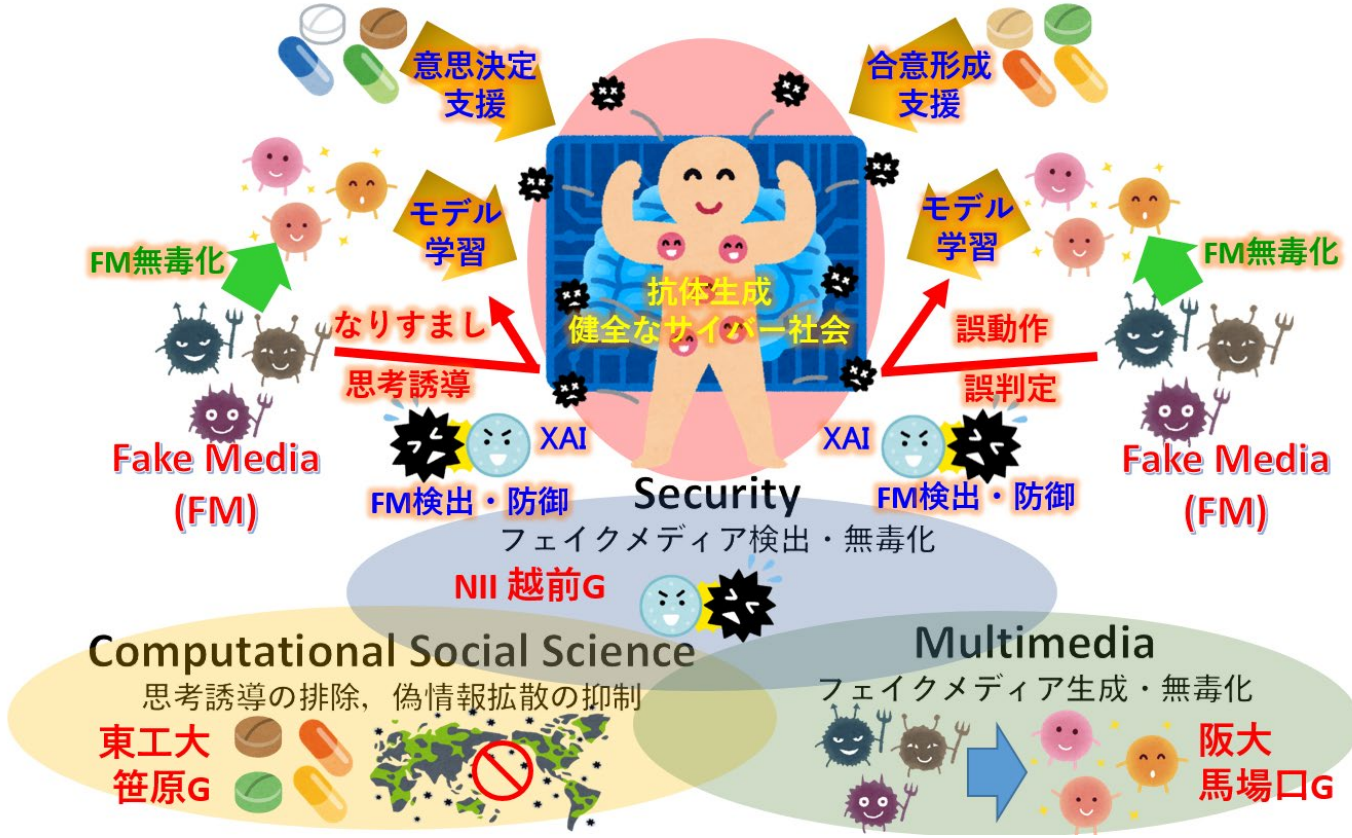
Those who received accuracy tips were 50% more discerning in sharing habits versus users who did not. *(Source: Jigsaw)*

11%

Pre-roll videos on YouTube drove up to an 11% increase in confidence, three weeks after exposure. *(Source: Jigsaw)*

CREST
FakeMedia

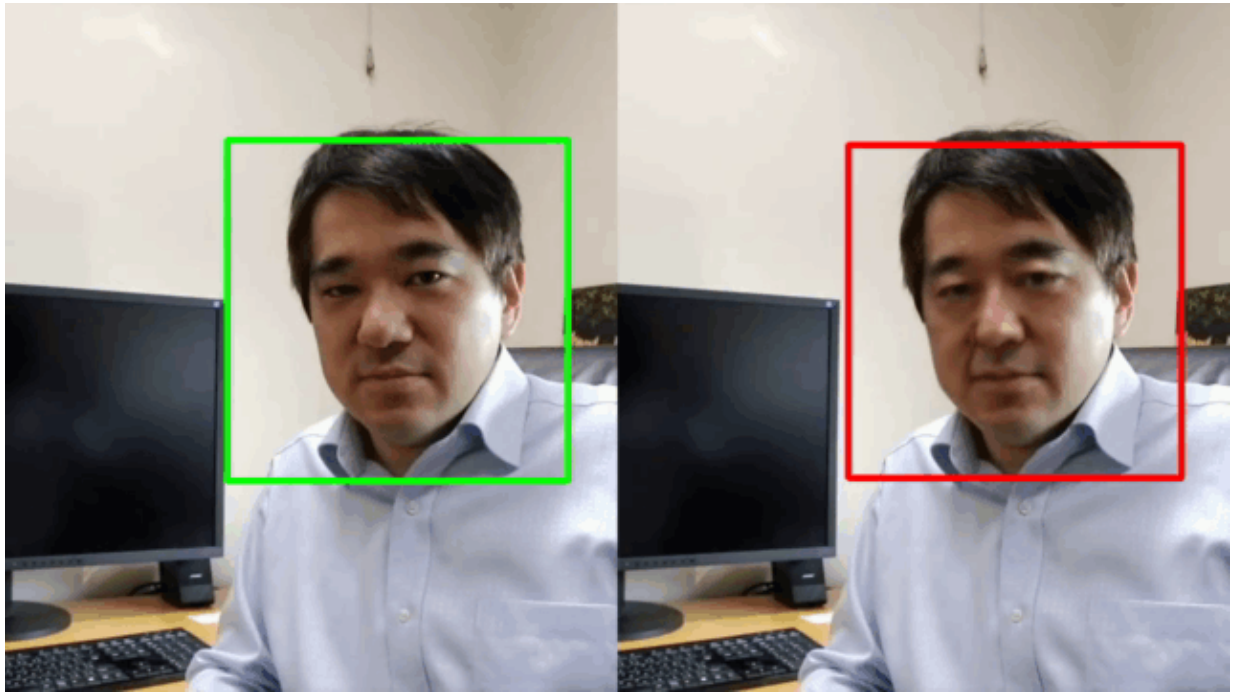
笹原 和俊（東京科学大・教授）
五十嵐 祐（名大・教授）
橋本 康弘（会津大・上級准教授）
Piyush Ghasiya（博士研究員）
陳 佳玉（博士研究員）
Bruno Toshio Sugano (研究員)



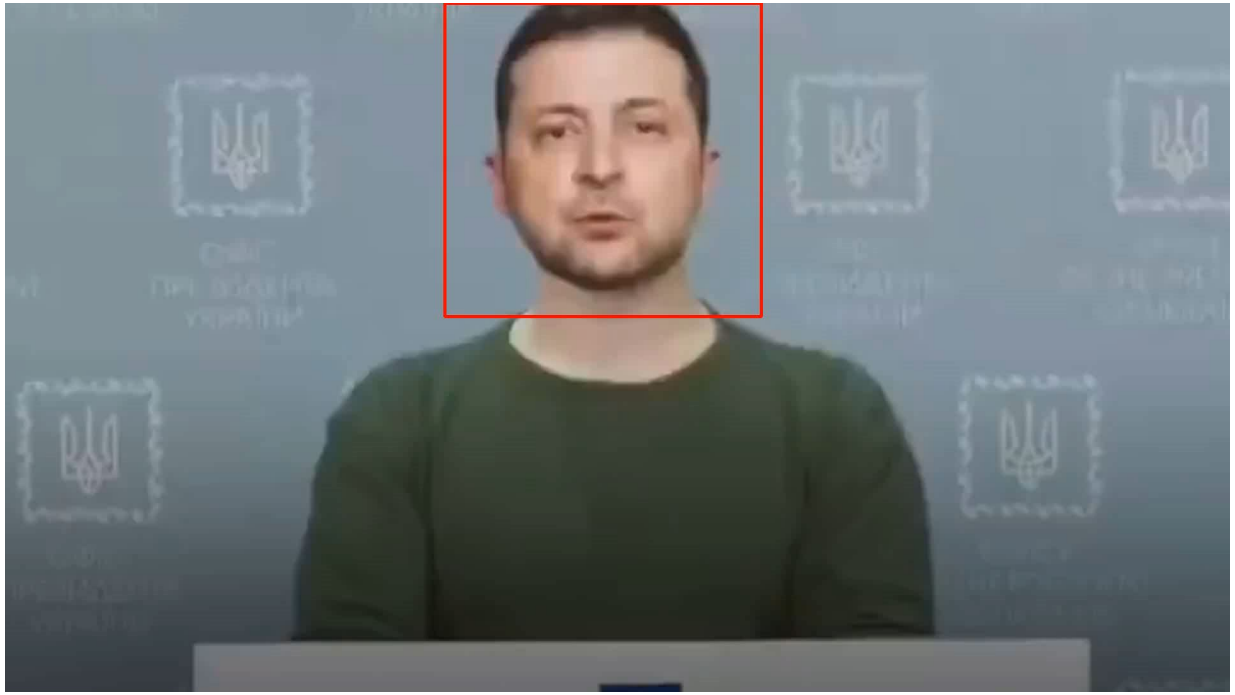
http://research.nii.ac.jp/~iechizen/crest/research.html

Synthetiq Vision

越前教授と安倍元首相の顔を入れ替えたディープフェイク



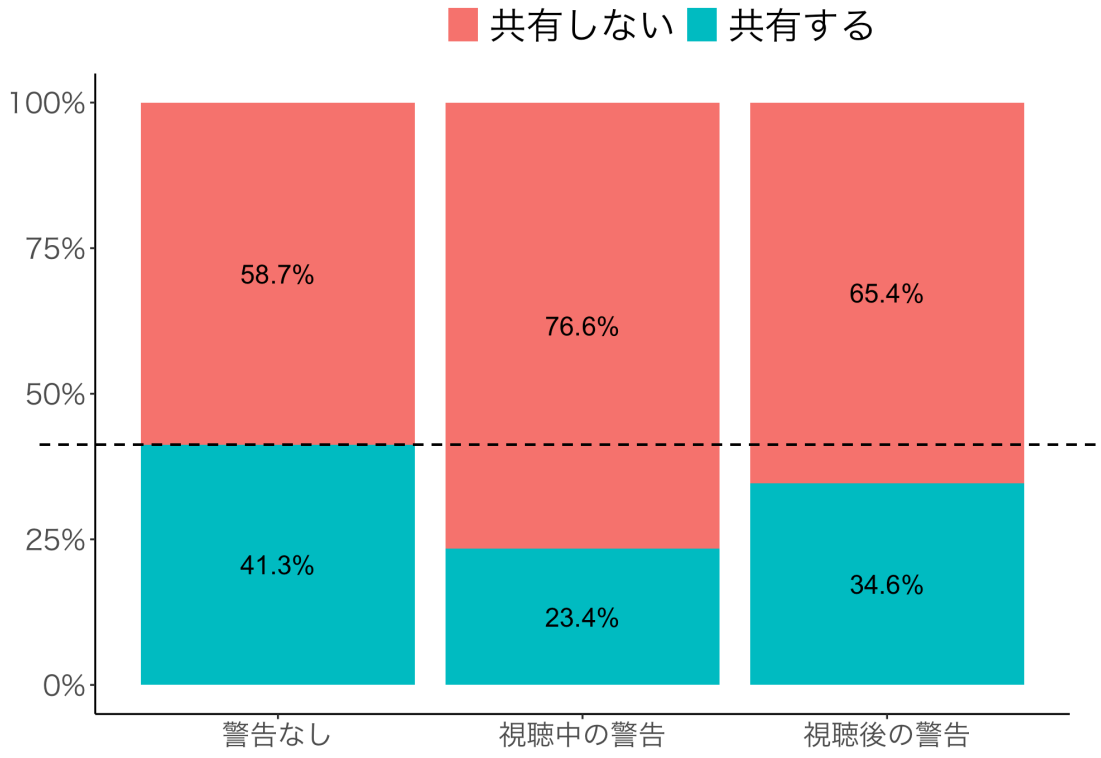
ゼレンスキー大統領のディープフェイク



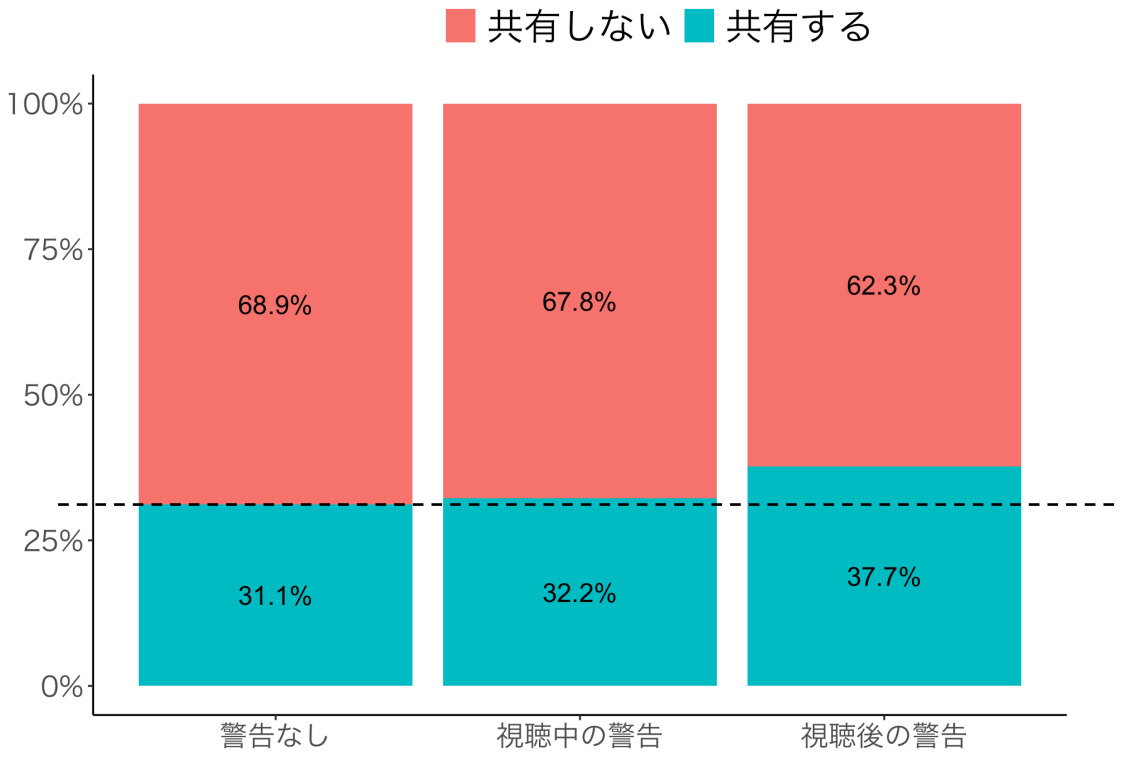
https://www.synthetiq.org/

警告介入はタイミングと頻度によっては逆効果

動画1つを視聴
警告の表示は1回のみ



動画4つを視聴
警告の表示が4回連続



J. Chen, B. T. Sugano, A. Frik, H. H. Nguyen, J. Yamagishi, I. Echizen, T. Igarashi, and K. Sasahara (under review)
https://osf.io/preprints/osf/xzjvg_v1

本物に翼を。

XFinch実験

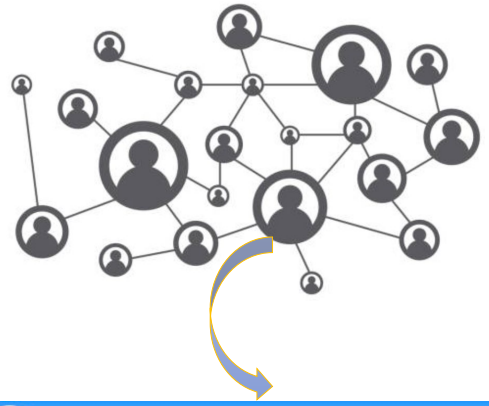
1282名の日本人参加者を対象とする大規模実験
2025年6月6～8日の3日間, 計5回
東京科学大学倫理審査許可番号2025050

陳佳玉, B.T.Sugano, 上山健太, 五十嵐祐, 笹原和俊 w/越前G

RQ：ソーシャルメディアにおけるディープフェイクの拡散を抑制するために、**検出ツールをどのように活用するのが有効か**?また、**潜在的な問題**は？

実験の流れ

- XFinchは、**動画を視聴・共有するSNS**
- 動画は、**普通の動画**と**ディープフェイク**を含む
- **検出ツールには越前Gのモデル**を使用
  （H. H. Nguyen, J. Yamagishi, and I. Echizen, IJCB 2024）
  True~94%, Deepfake~74%
- 参加者は他の参加者と繋がっている
  （**社会的ネットワーク**）
- 参加者は、システムから定期的に供給される動画の他、フォローしている参加者が共有した動画も視聴・共有できる
- 自分が共有した動画がフォロワーに共有されると、通知を受け取る
  （**通知の数 ～ 評判**）



実験の操作

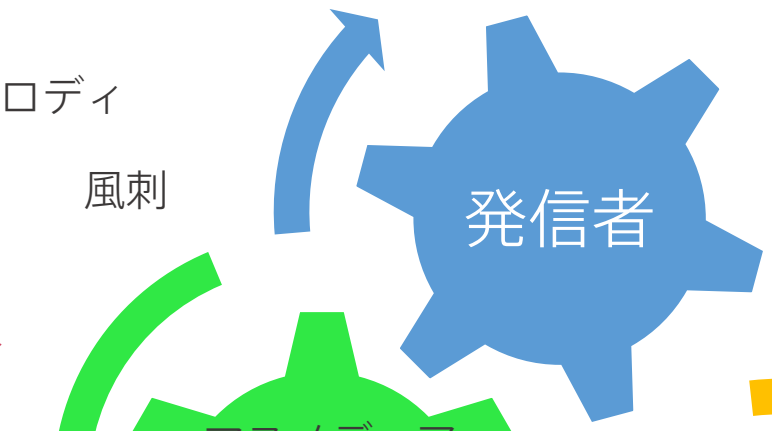
統制条件 n=329	• 検出ツールなし
独立条件 n=315	• 検出ツールあり • 疑わしい動画を報告可能 ただし、その集計結果は表示されない
社会条件 n=326	• 検出ツールあり • 疑わしい動画を報告可能 かつ、その集計結果も表示される
ソーシャルエンジニアリング条件 n=312	• 社会条件とほぼ同じ ただし、検出ツールの結果がウソ： 例：改竄の可能性が高い → 低い

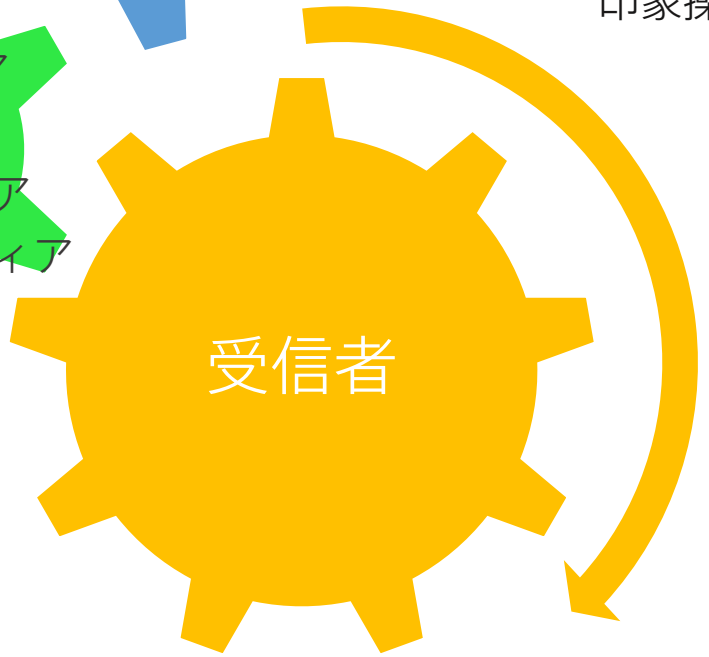
発表内容

1. フェイクニュースの拡散
2. エコーチェンバーとフィルターバブルという環境要因
3. AIで高度するフェイクニュース
4. フェイクニュースの拡散抑止のための介入技術
5. まとめ

フェイクニュースの情報生態系

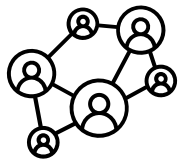
本当のような嘘
>> 嘘のような本当





経済的動機
釣り記事
陰謀論
プロパガンダ
選挙ビジネス
クリックベイト
政治的意図
ガイドライン
パロディ
ネット広告
風刺
党派的情報
法的枠組み
情報戦
ターゲティング
偏向報道
世論誘導
偽情報
アルゴリズム
印象操作
善意
誤情報
感情（怒り）
（生成）AI
不信感
Bot
ディープフェイク
アテンション
認知バイアス
フィルターバブル
社会的影響
エコーチェンバー
情報過多
情報リテラシー
インフォデミック

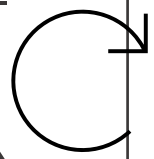
プラットフォームに埋め込まれた認知



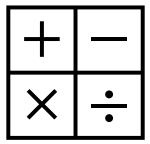
エコーチェンバー

ホモフィリー

アルゴリズム



情報環境の
局所最適化

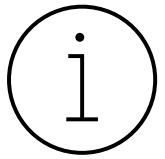


フィルターバブル

What is
engaged
with

What is
thought

What is
shown

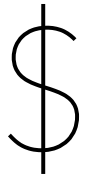


承認欲求

プラットフォーム

ネットワーク効果

アテンション
エコノミー



↑AIによるハッキング（認知戦）

K Program

経済安全保障重要技術育成プログラム

PRESS RELEASE

2024年7月19日
富士通株式会社

フェイクニュースの社会課題を解決する偽情報対策システムの研究開発を加速

「経済安全保障重要技術育成プログラム」にて採択

当社は、このほど、内閣府や経済産業省、その他の関係府省が、経済安全保障を強化・推進するため連携し創設した「経済安全保障重要技術育成プログラム（通称"K Program"）」(注1)のもと、国立研究開発法人新エネルギー・産業技術総合開発機構（以下、NEDO）が公募した、「偽情報分析に係る技術の開発」(注2)（以下、本事業）に実施予定先として採択され、偽情報の検知・評価・システム化に関する研究開発に着手します。事業の規模は60億円で、期間は2024年から2027年までの予定です。

https://pr.fujitsu.com/jp/news/2024/07/19.html

https://pr.fujitsu.com/jp/news/2024/10/16.html

ネット偽情報 検知から真偽判定まで行う総合的システム開発へ

2024年10月16日 15時33分

https://www3.nhk.or.jp/news/html/20241016/k10014610951000.html