「持続可能な社会を支える都市・インフラ学」第2回講演会

AIによる分断、AIによる包摂、そしてソーシャルAIへ

笹原 和俊

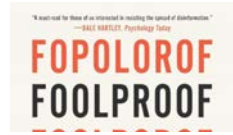
東京科学大学 環境・社会理工学院 イノベーション科学系

自己紹介

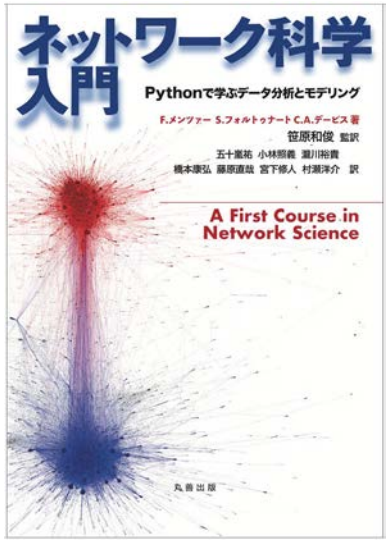
- 1976　福島県いわき市生まれ
- 2005　東京大学 大学院総合文化研究科修了（博士（学術））
- 2012～2020　名古屋大学 大学院情報学研究科 助教・講師
- 2016～2020　JSTさきがけ研究者（兼任）
- 2020～2024　東京工業大学 環境・社会理工学院 准教授・教授
- 海外経験　UCLA (FY2009), Indiana University (FY2016)
- 現在　東京科学大学 環境・社会理工学院 教授、系・課程主任
  国立情報学研究所 客員教授
- 研究　計算社会科学
- 主な受賞　第23回ドコモ・モバイル・サイエンス賞優秀賞（社会科学部門）

主な著書・訳書

令和7年国語教科書（三省堂）に一部掲載



監訳中



通底する問題意識

いかにして、信頼性・公平性・安全性を考慮しながら、AIと人の共生・共創を可能とする技術を実現するか

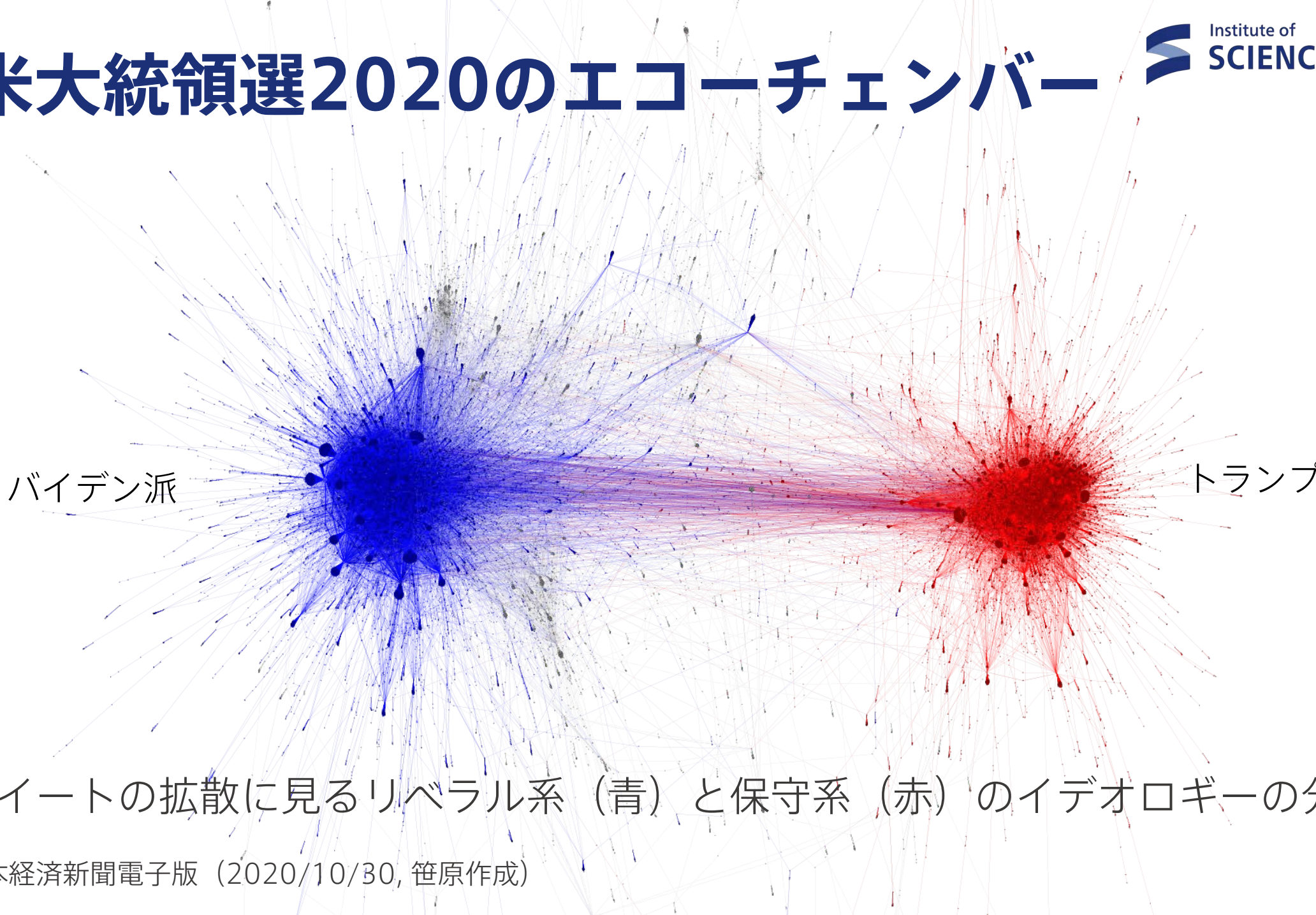
分断×AI

エコーチェンバー

同じ意見をもつ人々が集まり、自分たちの意見を強化し合うことで、多様な視点に触れることができなくなる現象

- インターネットの文脈ではCass Sunsteinが提起 (2001)
  - 意見の二極化と社会的分断（懸念）
- ホモフィリー（「類は友を呼ぶ」）による選択的接触
- それに加えて、アルゴリズムによる情報の取捨選択・調整

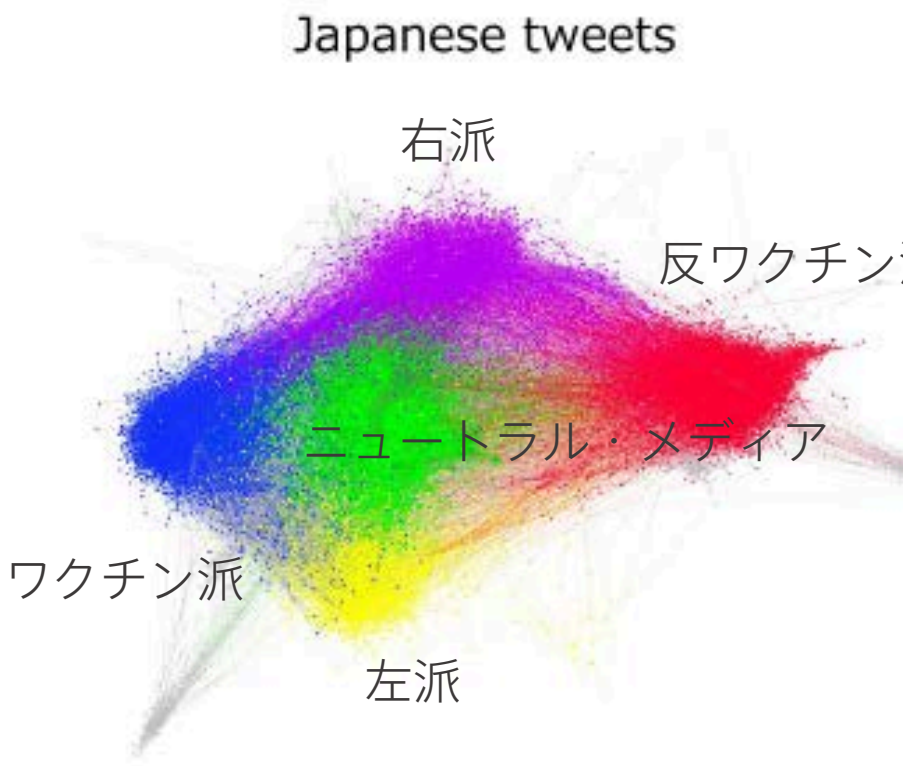
米大統領選2020のエコーチェンバー



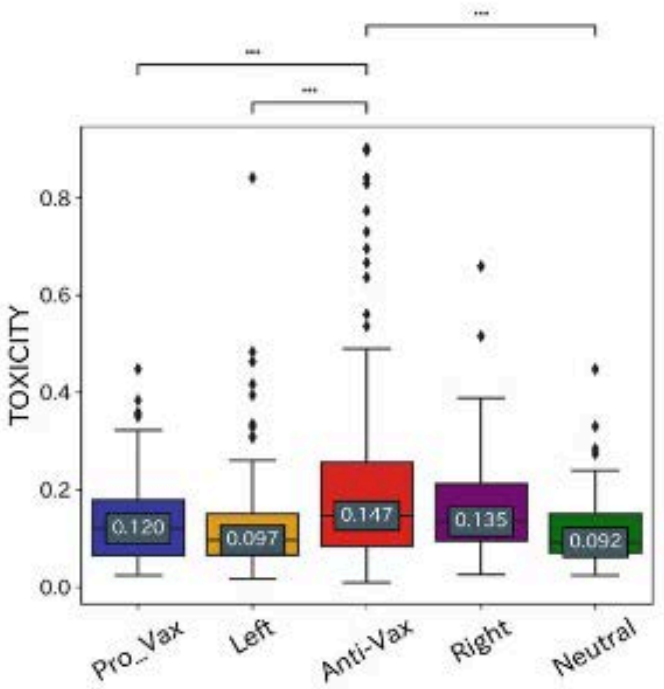
ツイートの拡散に見るリベラル系（青）と保守系（赤）のイデオロギーの分断

日本経済新聞電子版（2020/10/30, 笹原作成）

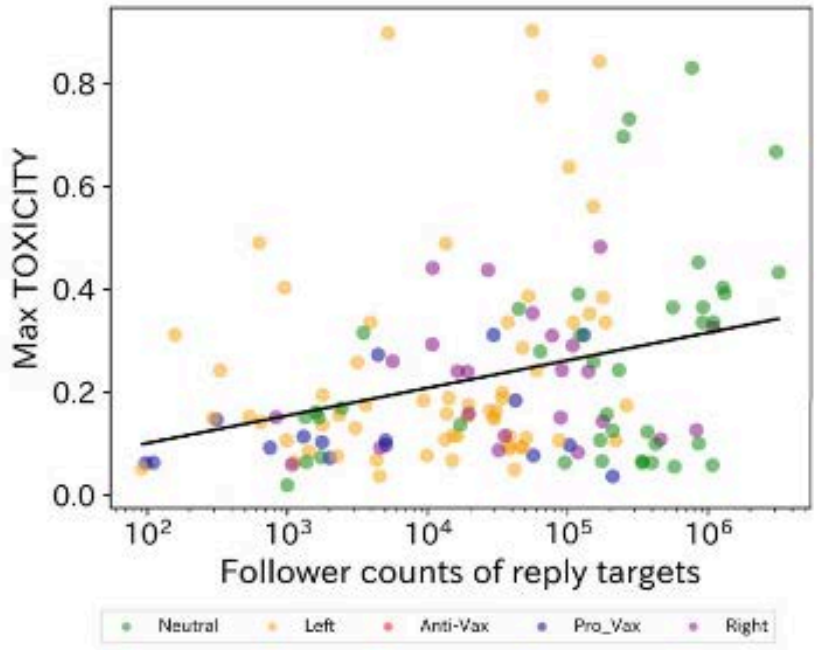
反ワクチン運動



反ワクチンの投稿の毒性は高い



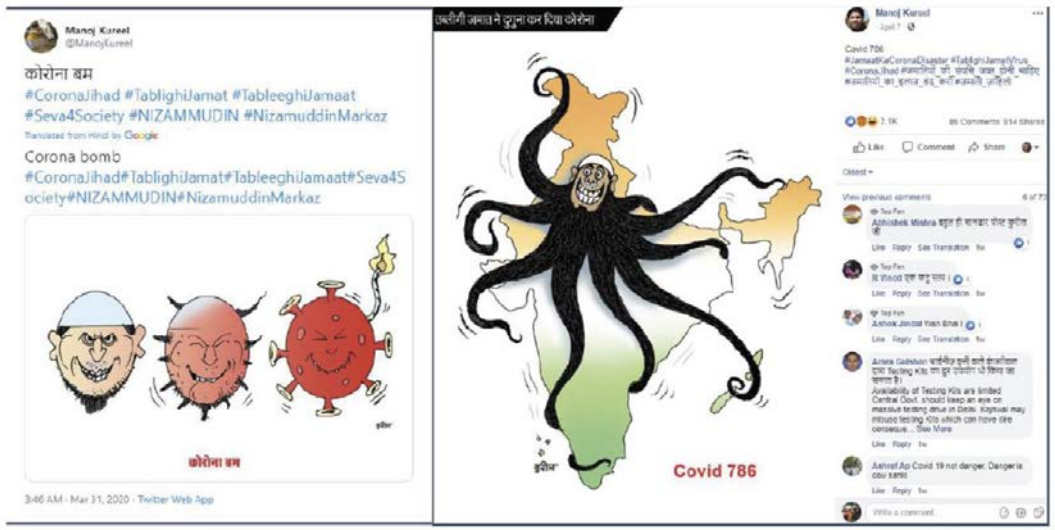
フォロワー数が多いほど毒性が高い
リプライを受け取る

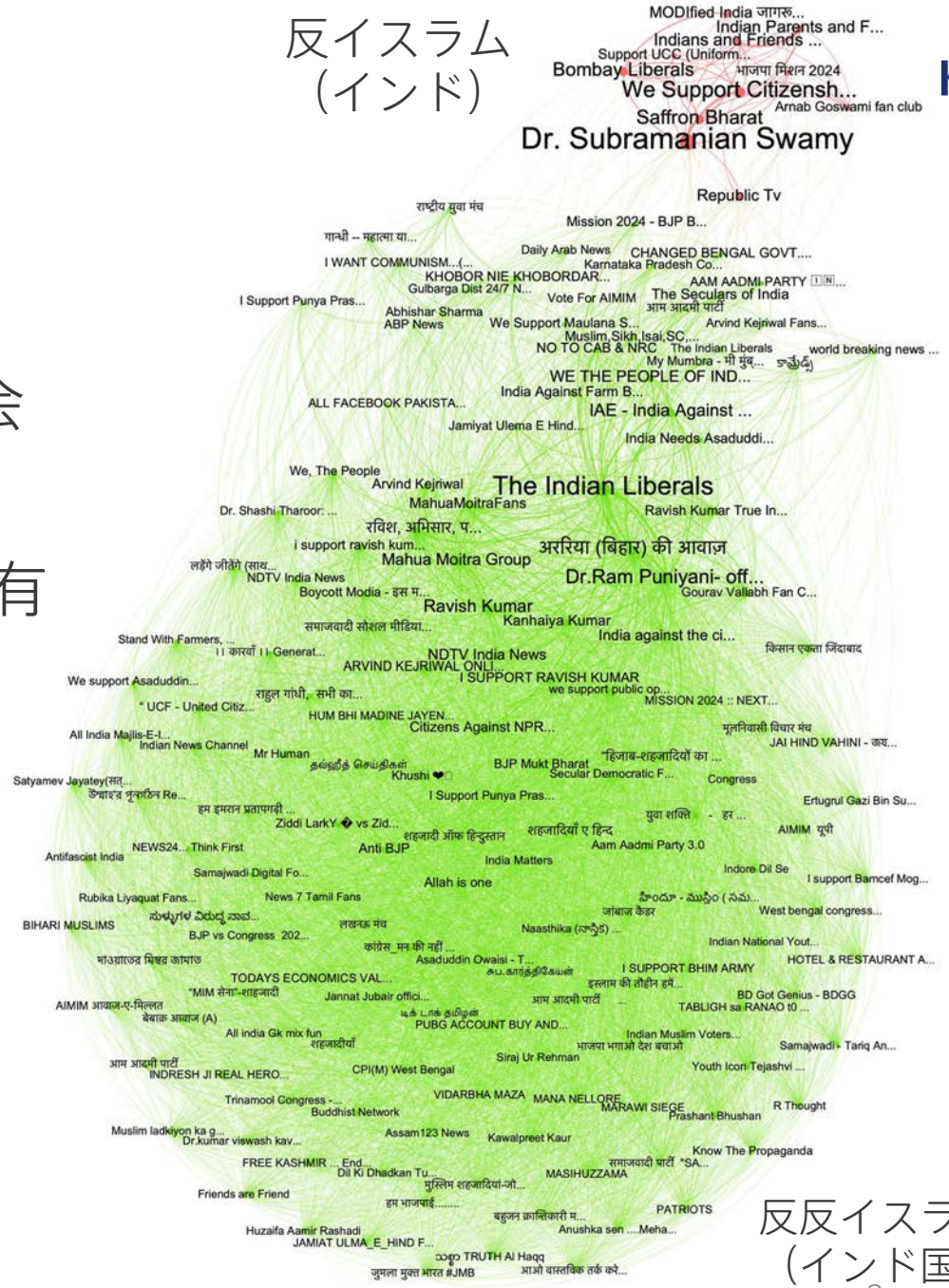


K. Miyazaki, T. Uchiba., K. Tanaka and K. Sasahara, Humanities and Social Sciences Communications (2022)

ヘイトの増幅

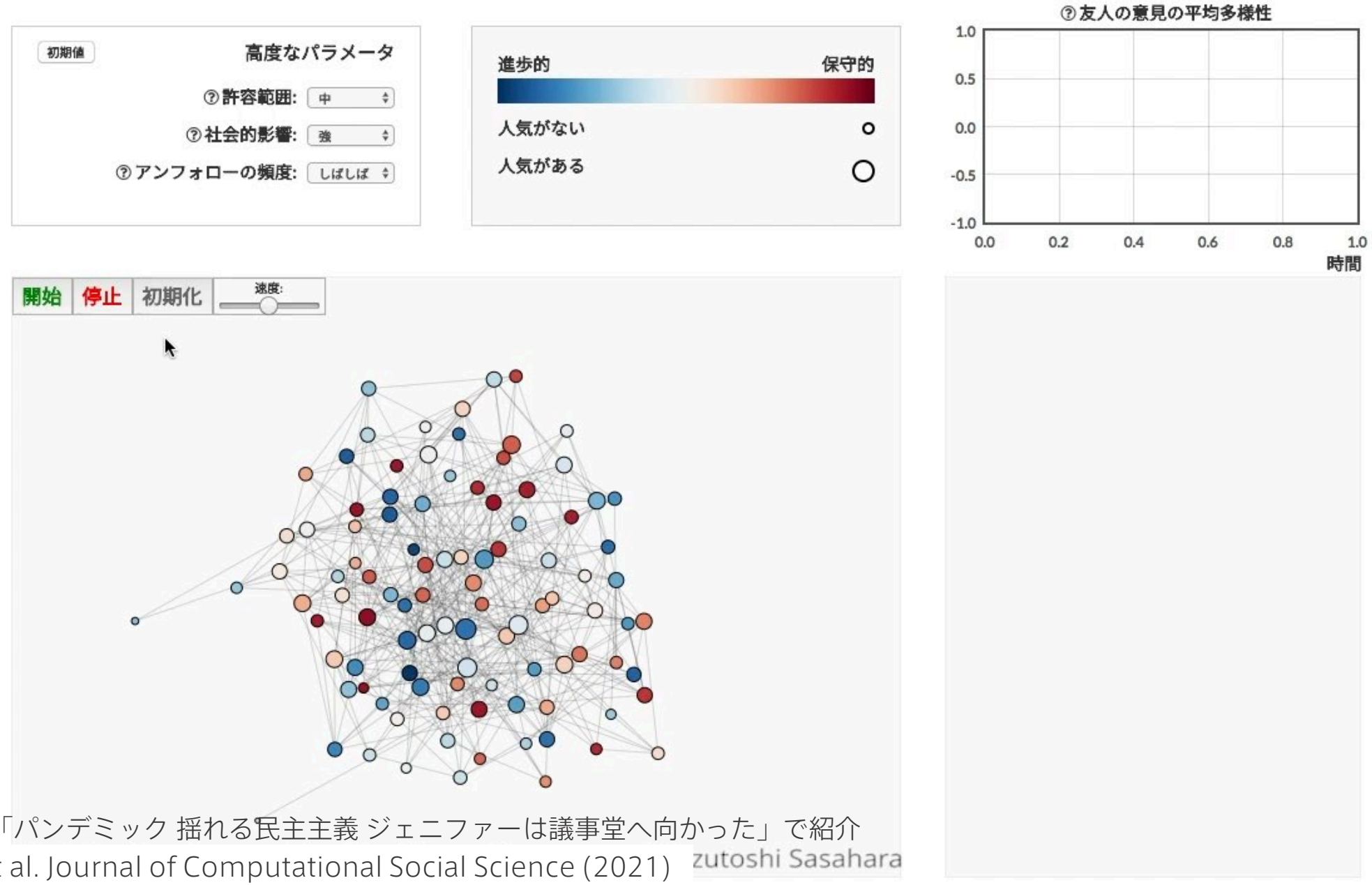
- イスラム教のタブリーグ・ジャマート集会に関するFacebookの投稿の共有
- 反イスラムは専らヘイト（偽情報）を共有
- 反イスラムの投稿は反反イスラムの3倍速く拡散





P. Ghasiya and K. Sasahara, SM+S, 2022

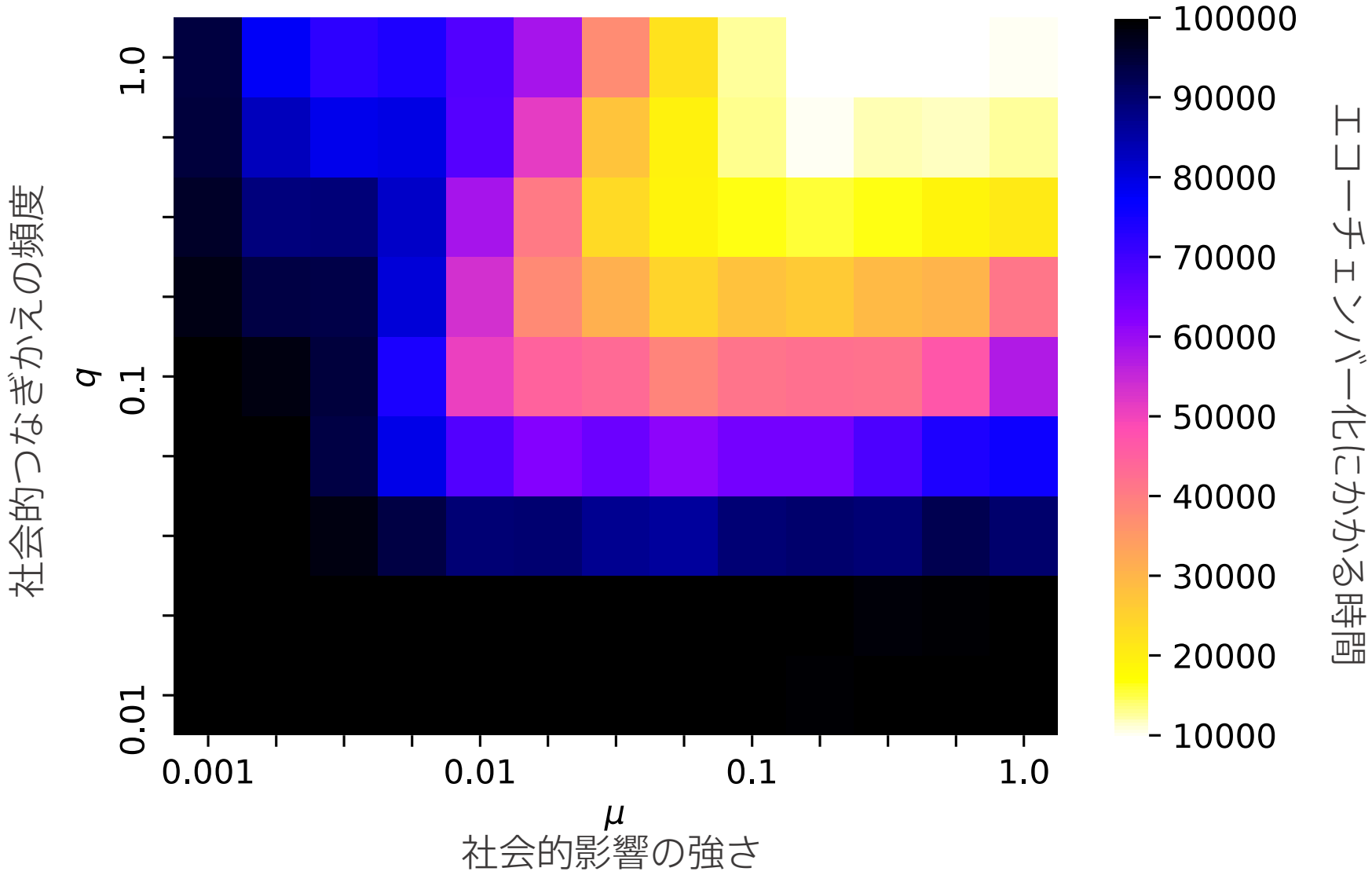
エコーチェンバーはどのようにしてソーシャルメディアから生じるのか



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)

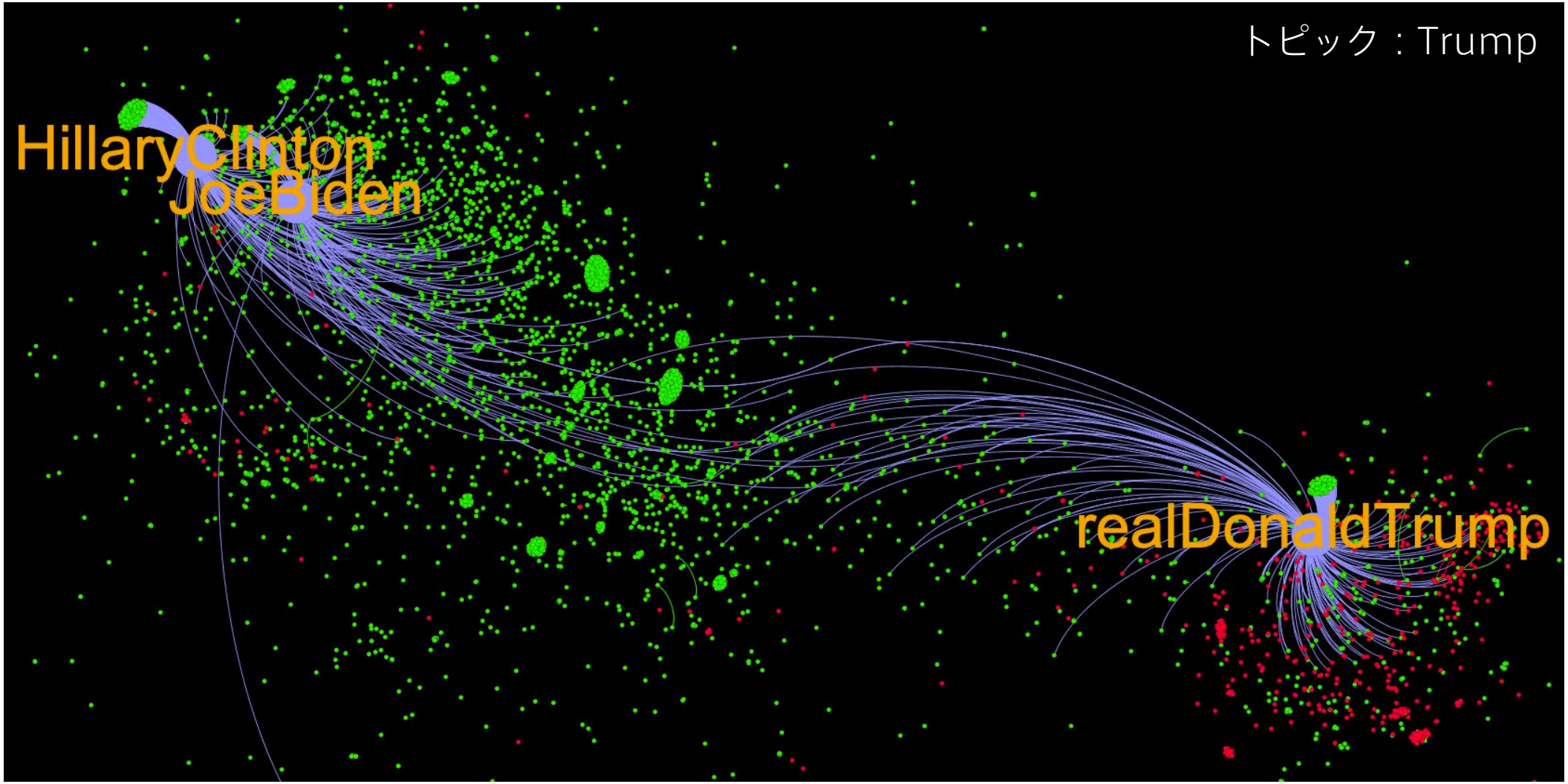
SNSはエコーチェンバーを加速する



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

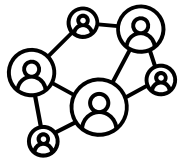
陰謀論を増幅するBot

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

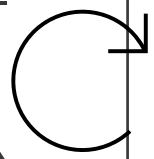
プラットフォームに埋め込まれた認知



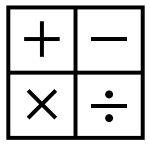
エコーチェンバー

ホモフィリー

アルゴリズム



情報環境の
局所最適化

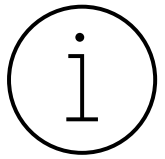


フィルターバブル

What is
engaged
with

What is
thought

What is
shown

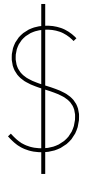


承認欲求

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

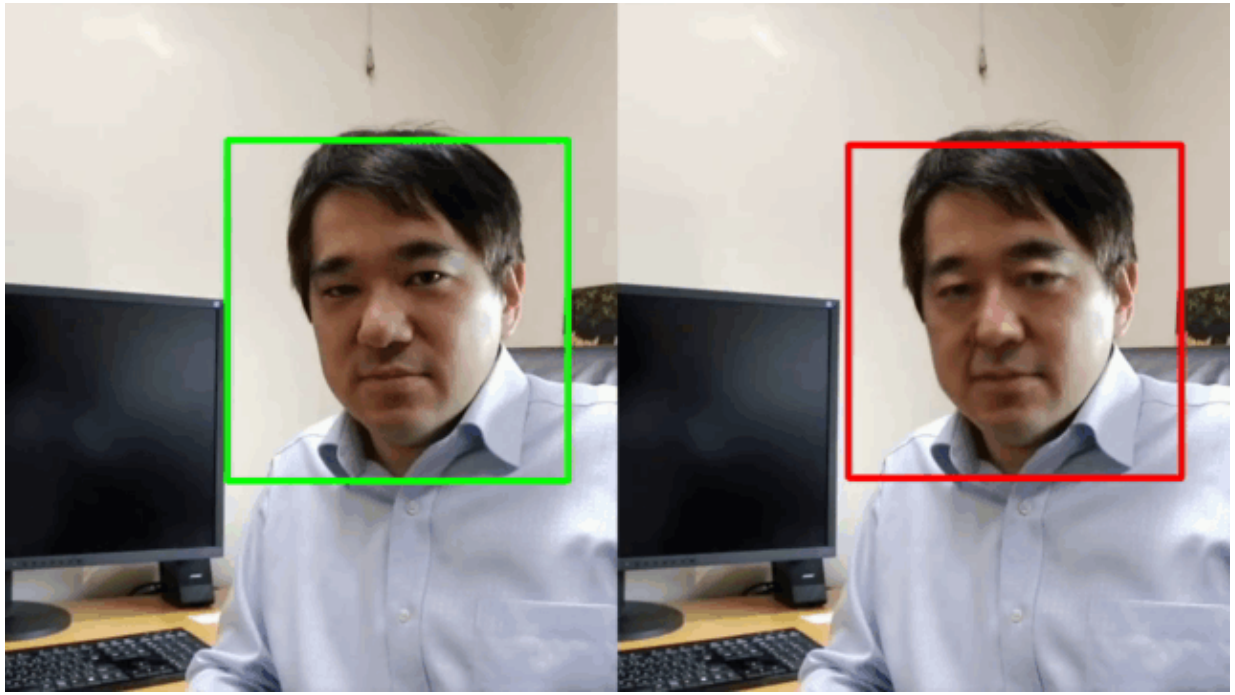


↑AIによるハッキング（認知戦）

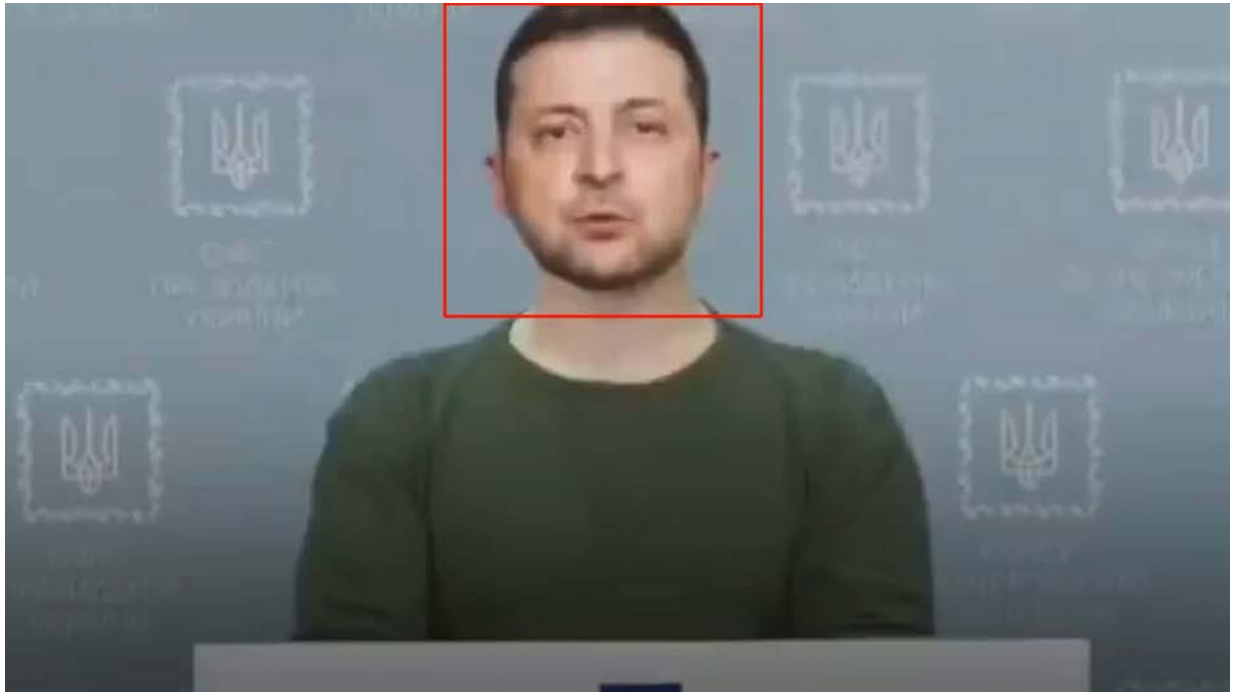
包摂×AI

ディープフェイクの生成と検出

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/

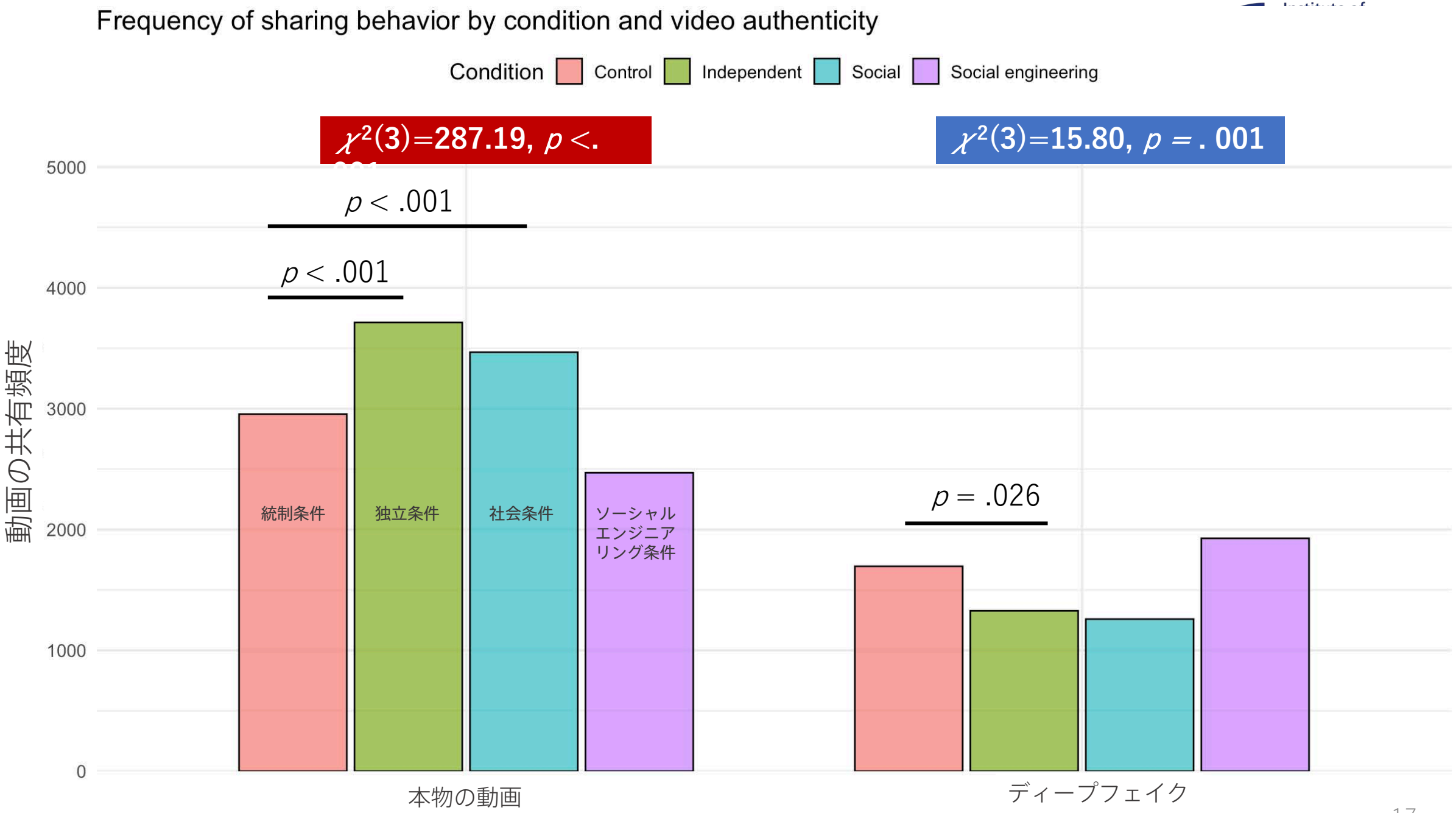
XFinch実験

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

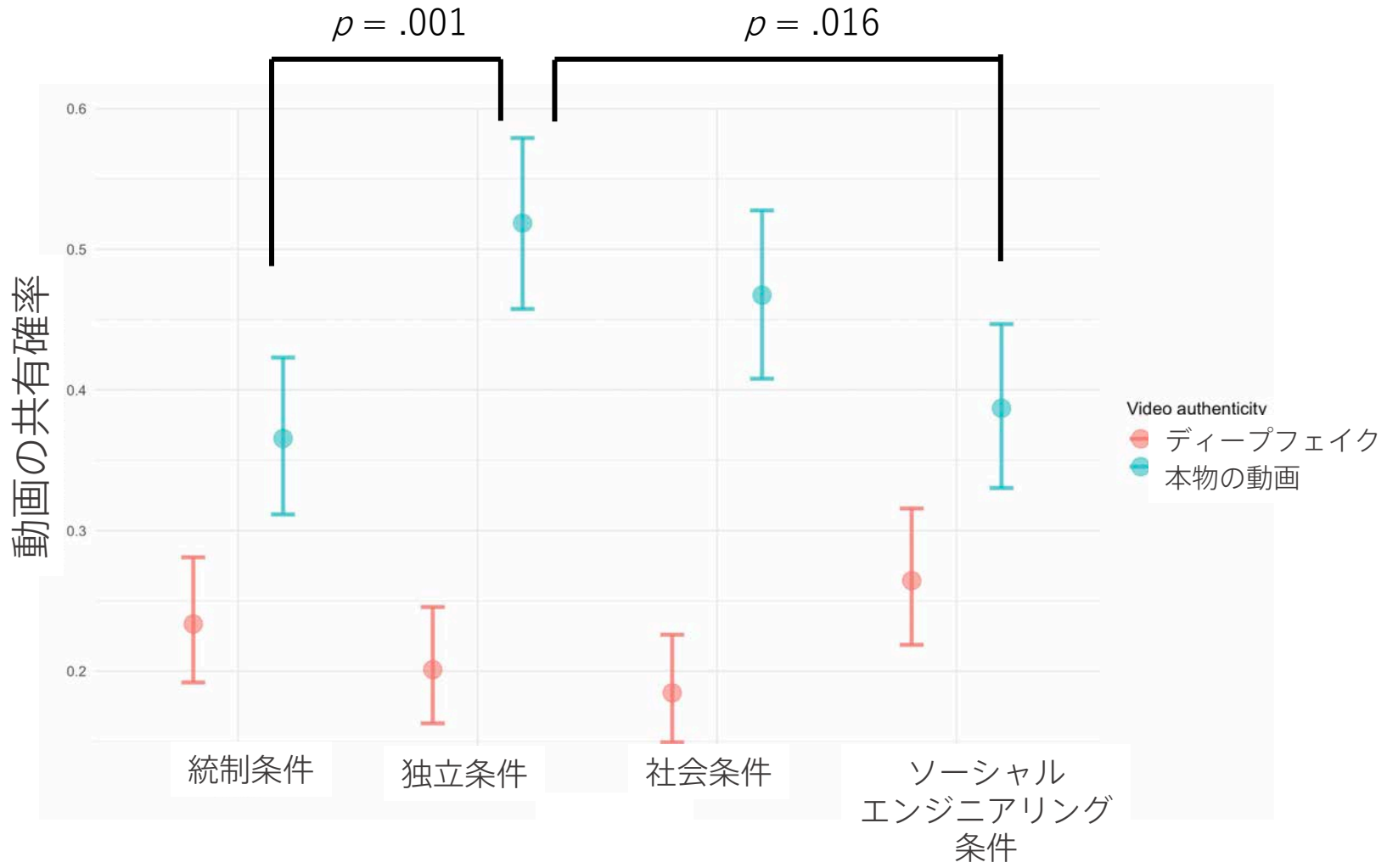
- XFinchは**動画を視聴・共有するSNS**
- 動画は、**普通の動画**と**ディープフェイク**を含む
- **検出ツールは越前GのAIモデル**を使用

  H. H. Nguyen, J. Yamagishi, and I. Echizen, IJCB 2024
  84% (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**）

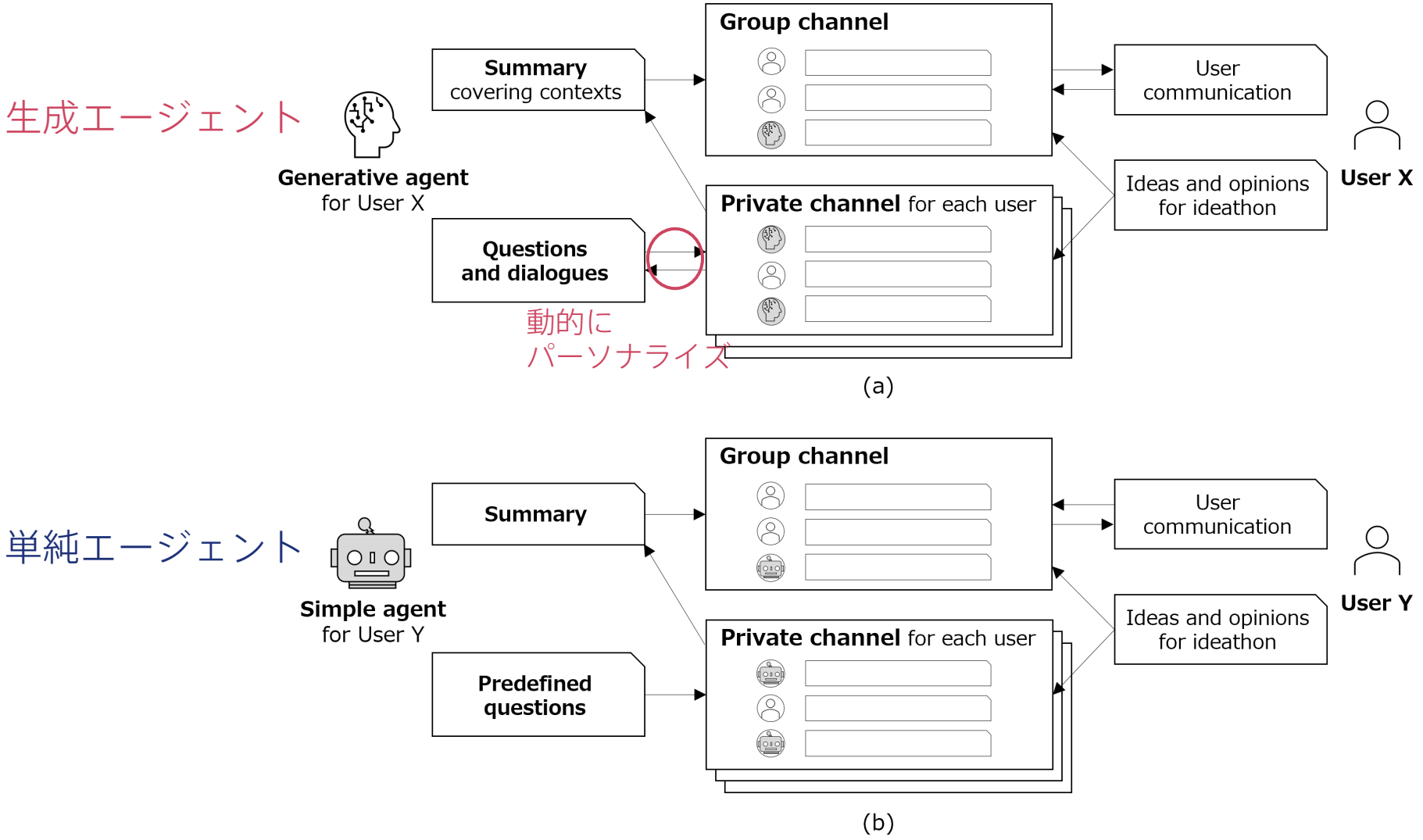
統制条件 n=329	• 検出ツールなし
独立条件 n=315	• 検出ツールあり • 疑わしい動画を報告可能 ただし、その集計結果は表示されない
社会条件 n=326	• 検出ツールあり • 疑わしい動画を報告可能 かつ、その集計結果も表示される
ソーシャルエンジニアリング条件 n=312	• 社会条件とほぼ同じ ただし、検出ツールの結果がウソ： 例：改竄の可能性が高い → 低い


Frequency of sharing behavior by condition and video authenticity
Condition
Control
Independent
Social
Social engineering
$\chi^2(3)=287.19, p <.$
$\chi^2(3)=15.80, p = .001$
$p < .001$
$p < .001$
$p = .026$
5000
4000
3000
2000
1000
0
動画の共有頻度
統制条件
独立条件
社会条件
ソーシャル
エンジニア
リング条件
本物の動画
ディープフェイク

AIツールの使用は「本物の動画」の共有を促進



生成エージェントが促進する包摂（w/富士通）



アイディアソンによる実証実験

東京科学大生（東工大生）を含む25名が参加

東京工業大学倫理審査No.2023319

【アイデアソン参加者募集】「分散型SNS × AIチャットボットの活用アイデアを考えよう！」（応募〆切1/28まで延長しました）

《次世代SNSを創造してみませんか？》

皆さんは生成AIやSNSを活用していますか？生成AIは、膨大なデータから学習して人間が作成するようなテキストや画像、音声を生成する能力を持っています。これにより、迅速かつ高精度な情報処理を実現し、様々な作業をサポートすることが可能になり、現在その活用が急速に進んでいます。一方SNSは、従来の大規模運営事業者が一方的にルールやサービスを決めるような集中型SNSが問題視される中、利用者が自由に独自のSNSを立ち上げたり、所属するSNSを自由に選択可能な分散型SNSが登場し、柔軟でユーザ中心のなサービス形態が期待されています。

現在、生成AIはSNSの投稿を直接扱うことはできませんが、SNSの進化に伴いその可能性は広がっています。SNSに生成AIの機能が加わったら何ができるでしょうか？富士通の研究部門では、分散型SNS上で生成AIチャットボットがユーザとコミュニケーションする機能（質問や雑談への回答）を有する次世代SNSを開発しています。分散型SNSがAIチャットボットとのコミュニケーション機能を持つことで、これまでになかったような新しいコミュニケーションやビジネスなどへの活用が考えられます。

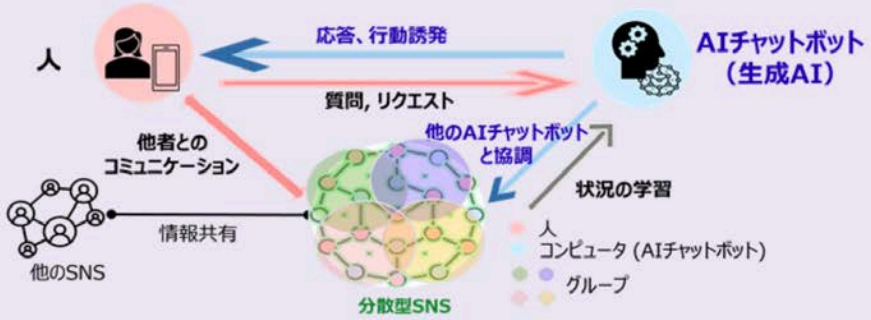
そこで今回、「分散型SNS×AIチャットボットの活用アイデアを考えよう！」を企画しました。富士通が開発中の次世代SNS（プロトタイプ）を実際に体験していただきながら、AIチャットボットを活用した次世代SNSの可能性や新しい活用アイデアを、グループごとに考えていただき、発表していただきます。

《日程》

- 参加募集期限：**2024年1月28日(日)**
- 専用SNS (富士通が開発中の次世代SNS) での情報交換期間：2024年1月29日(月)～2月2日(金)（*この期間での参加時間の制約はありません。自由な時間にSNSでの情報交換（投稿・閲覧）が可能です。*）
- アイデアソン開催日時：2024年2月5日(月)AM10-12時

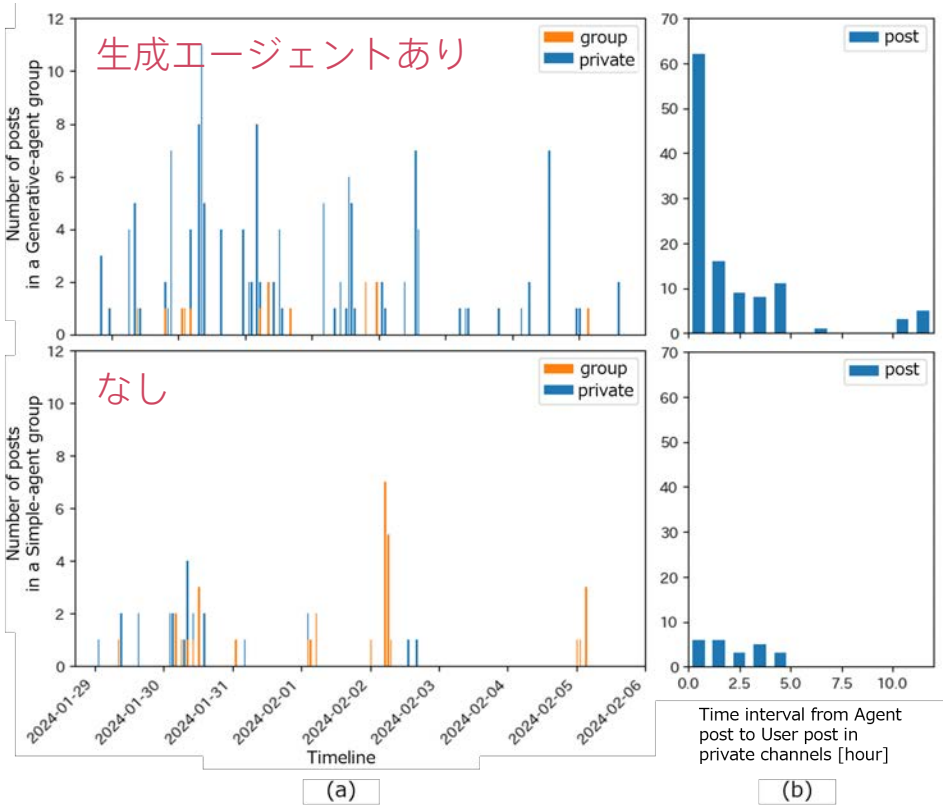
《入賞者は富士通の公開ページで発表！》

アイデアソンの結果は、入賞者の名前とともに富士通の公開ページで発表させていただきます。参加者の皆様の実績情報としてご活用いただけます。

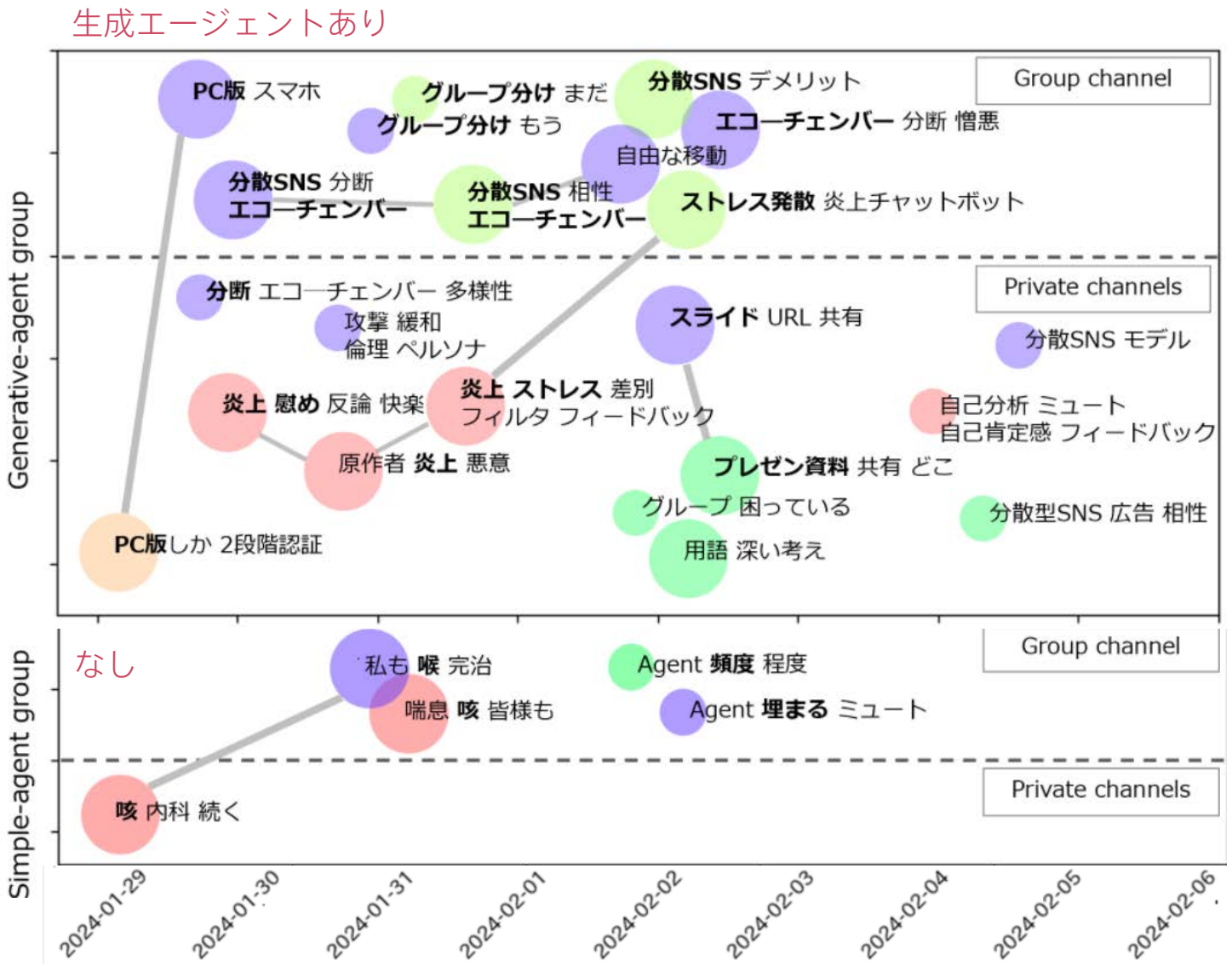


分散型SNSと生成AIの組み合わせ

生成エージェントは集合知を促進



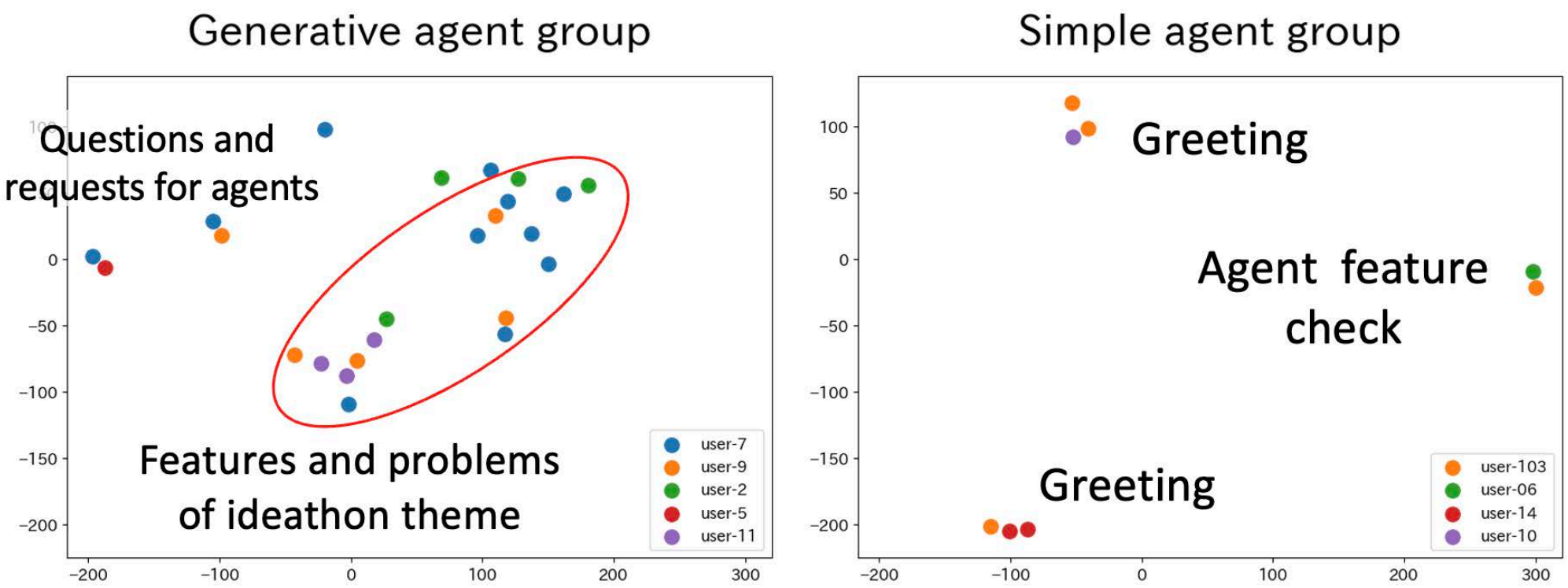
投稿数の推移と分布



トピックの変遷

M. Tanaka, A. Naito, K. Onoue, and K. Sasahara, IC2S2 2024

生成エージェントは集合知を促進（続）



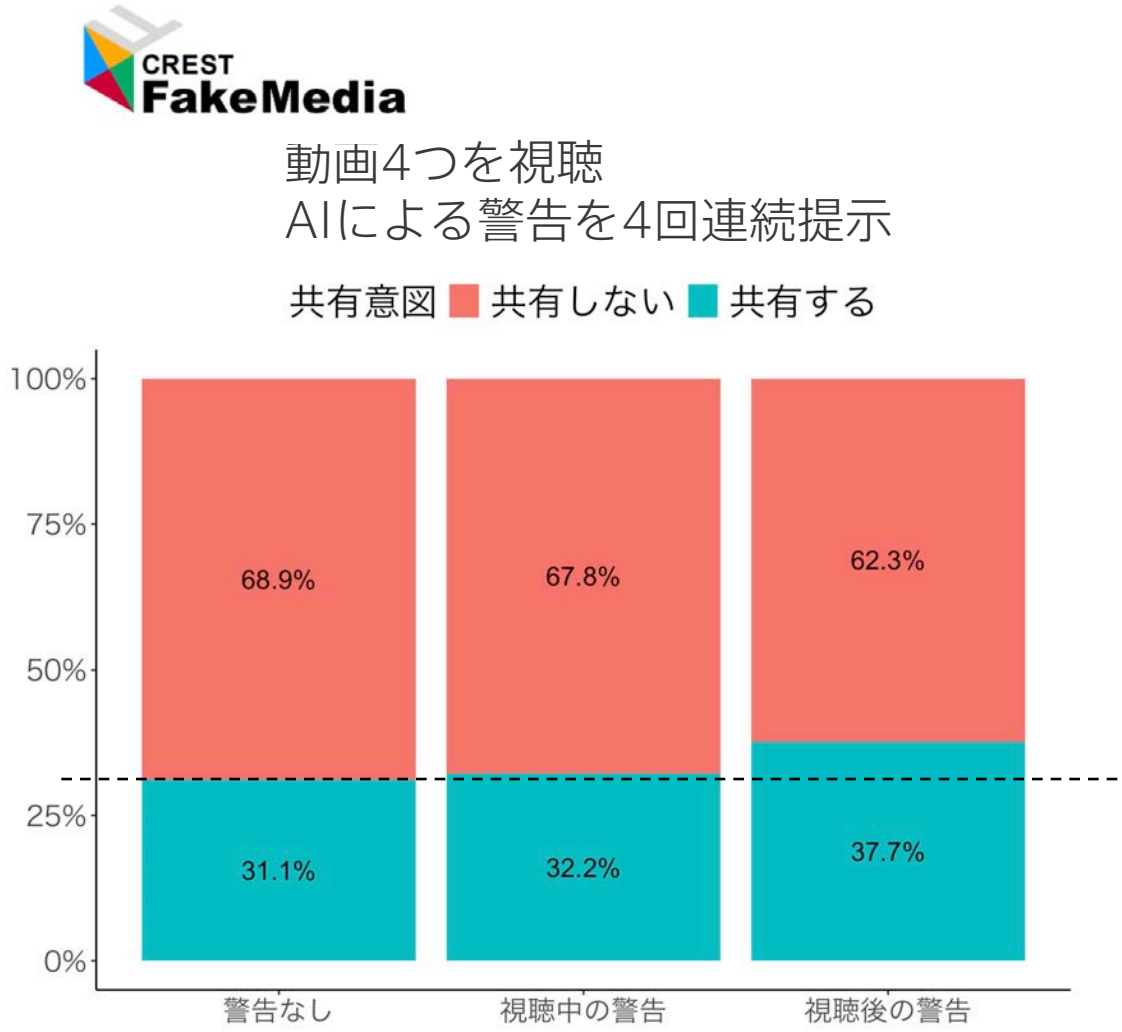
Vector representation visualizing posts that contain topics exchanged among participants by sentence-transformer and t-SNE.

M. Tanaka, A. Naito, K. Onoue, and K. Sasahara, IC2S2 2024

Institute of SCIENCE TOKYO

問題意識

- **AIの精度と受容性の乖離**
  - AIの正確な判定も、警告の繰り返しは効果がないか、逆に共有意欲を高めるパラドックス
- **人-AIの相互信頼の重要性**
  - 透明性・公平性と並ぶAI設計の国際的課題（EU・OECD・NIST）
- **ソーシャルAI設計学の必要性**
  - AIを「真偽判定装置」から「信頼を媒介・形成する社会的存在」へ再定義



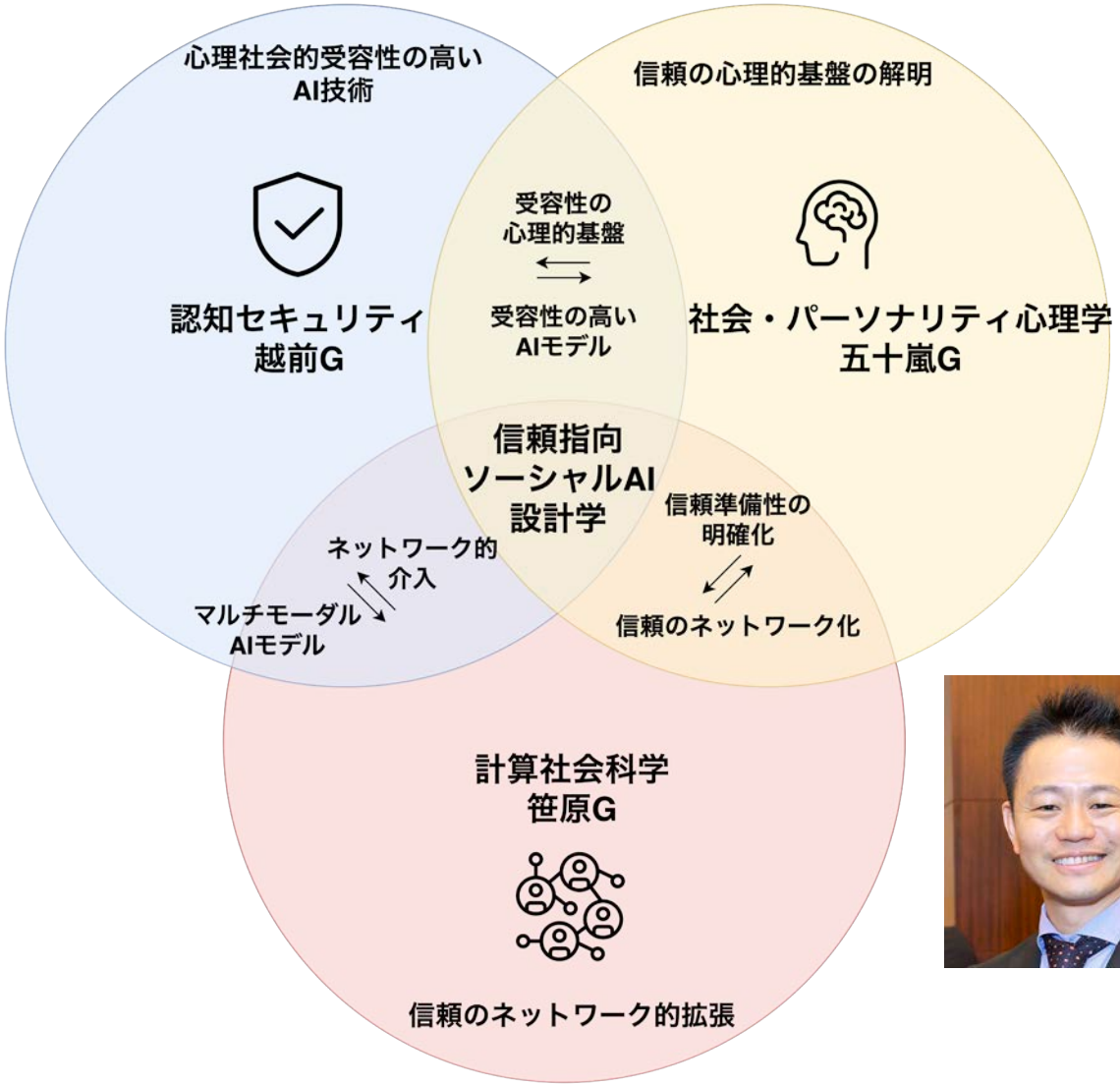
J. Chen et al. (under review)
https://osf.io/preprints/osf/xzjvg_v1

JST CREST（「人とAIの共生・協働社会を実現する学際的システム基盤の創出」）に採択！

Institute of SCIENCE TOKYO

不確実性社会を克服する信頼指向ソーシャルAI設計学

分担
越前功（NII）

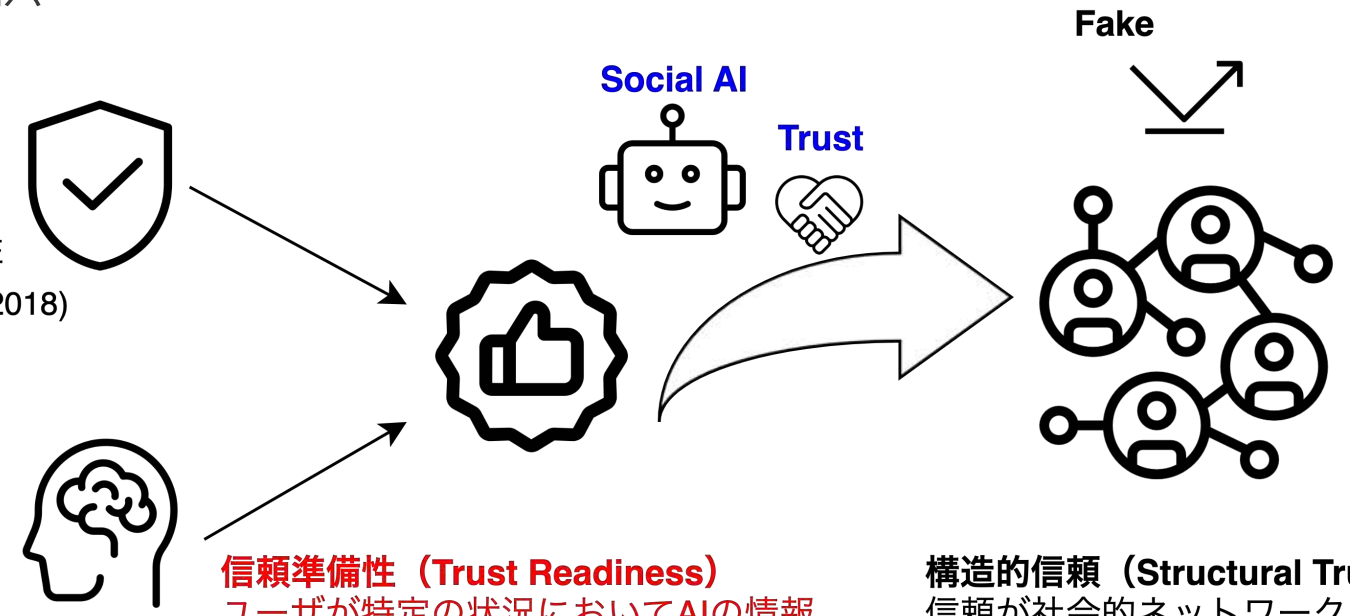


分担
五十嵐祐（名大）

代表
笹原和俊（東京科学大）

信頼準備性という基盤概念

信頼準備性 (Trust Readiness) は、技術的な信頼相当性（Trustworthiness）や個人の信頼傾向（Trust Propensity）を補完しつつ、動的かつ文脈適応的な信頼の側面を捉える視座を提供



信頼相当性（Trustworthiness）
AIやその提示する情報の正確さ・透明性・公平性・一貫性などに基づく、客観的でシステム側における設計指標となる信頼性

Thielsch, M. T., Meeßen, S. M., & Hertel, G. (2018)

信頼傾向（Trust Propensity）
ユーザの文化的背景やパーソナリティに依存する、AIに対する信頼のしやすさという個人差的傾向

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995)

信頼準備性（Trust Readiness）
ユーザが特定の状況においてAIの情報や介入を受け入れる心理的・社会的準備状態。感情、文脈、社会関係性に応じて動的に変化する（本研究で導入）

構造的信頼（Structural Trust）
信頼が社会的ネットワークを通じて集団的に拡張される創発的プロセス。個人の内面を超えて社会スケールでの信頼の形成・維持・強化に関わる

Luhmann, N. (1979)

図1　本研究における信頼の4つの側面
（JST CRDSのトラストに関する資料を踏まえた上での整理）