SS研合同分科会

ディープフェイクの実態と対策
計算社会科学の見地から

笹原 和俊

Institute of
SCIENCE TOKYO

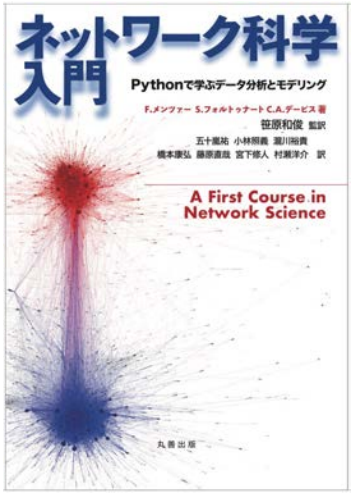
自己紹介

- 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

参考文献

令和7年国語教科書（三省堂）に一部掲載 👇

2024年10月発売 👇



発表内容

1. フェイクニュース現象の振り返り
2. 生成AIによるフェイクの高度化
3. 介入によるフェイクの拡散抑止
4. まとめ

発表内容

1. フェイクニュース現象の振り返り
2. 生成AIによるフェイクの高度化
3. 介入によるフェイクの拡散抑止
4. 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

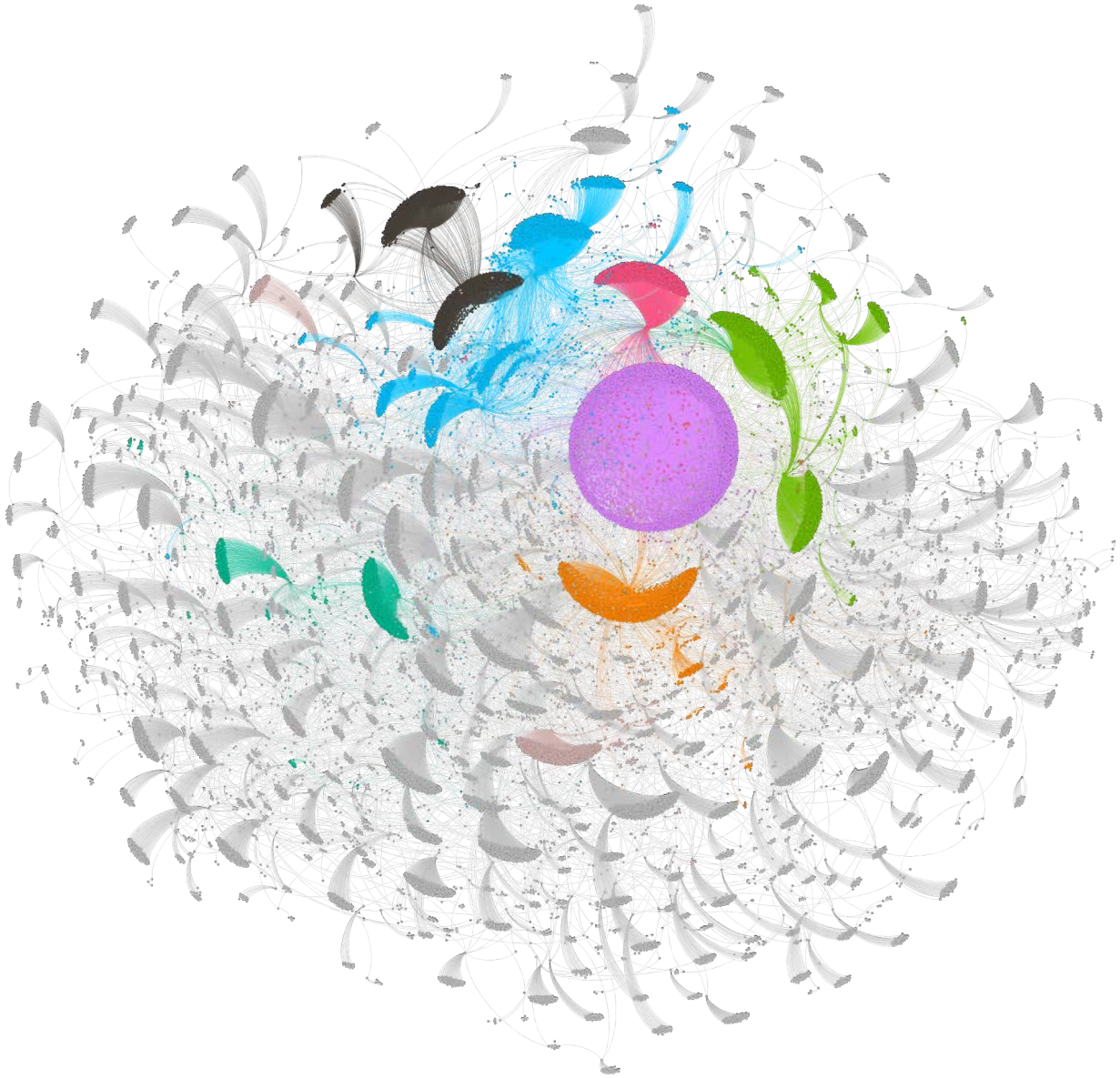
‘alternative facts’

‘post truth’

BBC NEWS

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

新型コロナ・インフォデミック

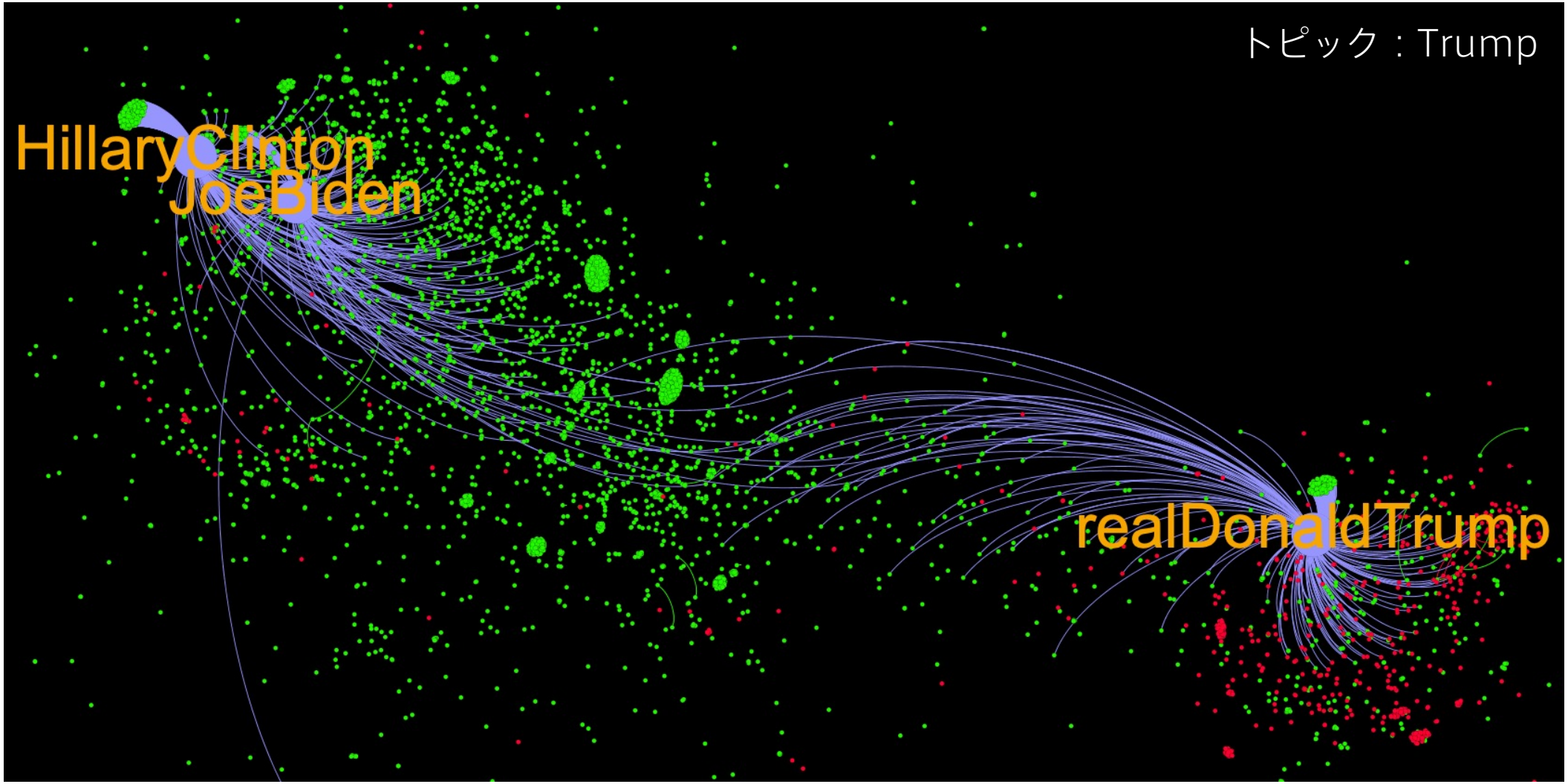


確かな情報と不確かな情報が混在する情報過多

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)

陰謀論を増幅するBot

赤：悪質なBot
緑：普通のBot



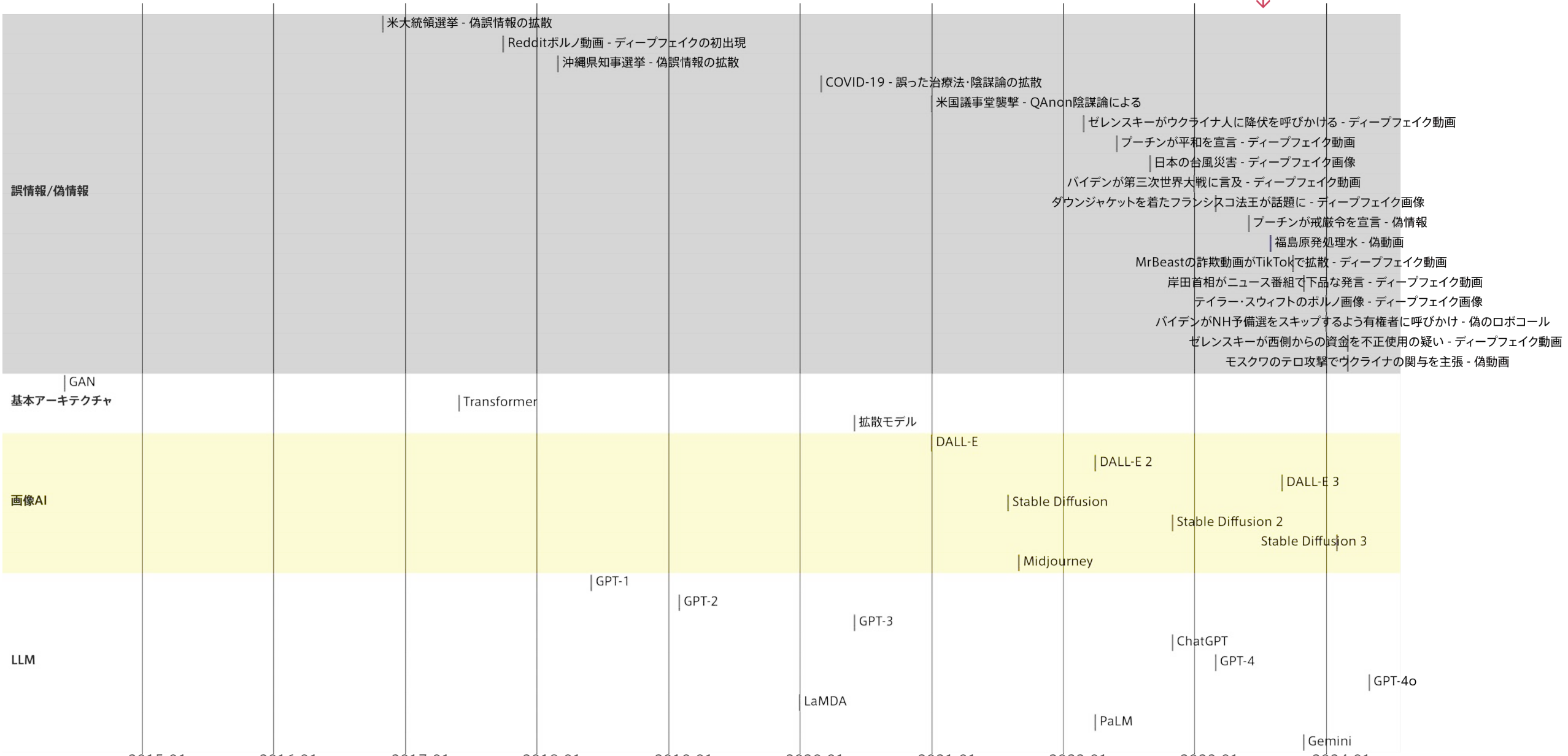
W. Xu and K. Sasahara, Journal of Computational Social Science, 2022　←米大統領経済報告2023で被引用

発表内容

1. フェイクニュース現象の振り返り
2. 生成AIによるフェイクの高度化
3. 介入によるフェイクの拡散抑止
4. 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

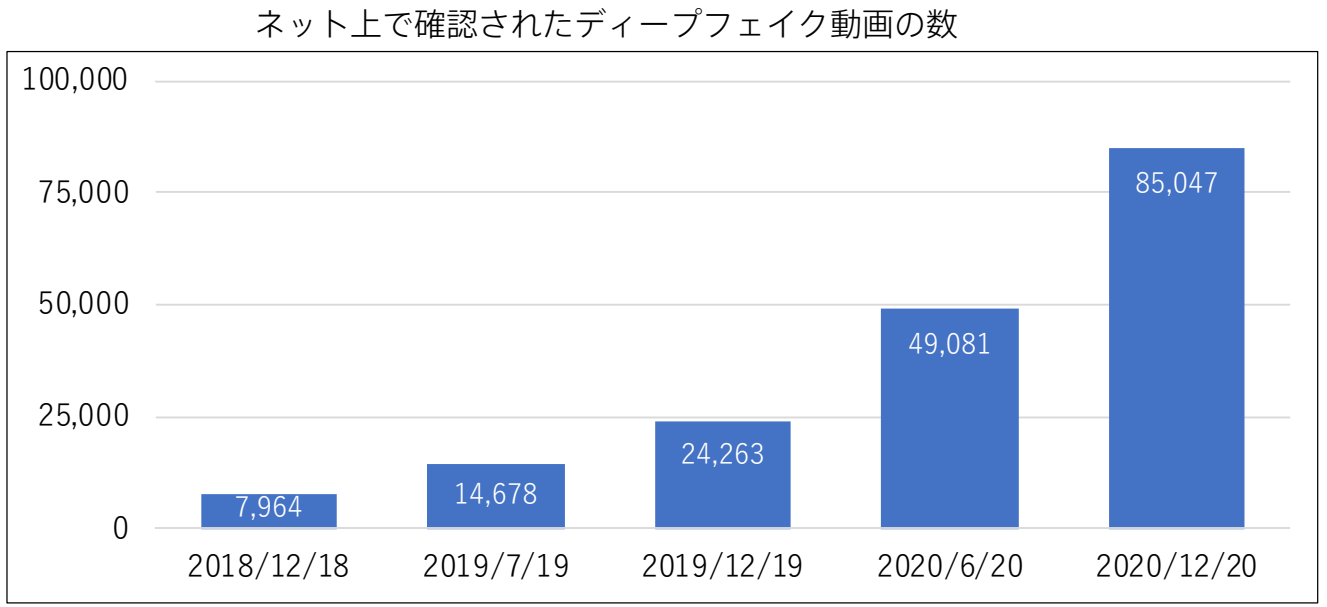
Governments have long employed human commentators—whether state officials, hired contractors, or party loyalists—to covertly manipulate information online. Generative AI technology is slowly beginning to enhance such distortion campaigns.

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)

ディープフェイクとは

- ディープラーニング（深層学習）とフェイク（偽物）を組み合わせた造語
- 広義：AIによって合成・生成されたメディアやその技術
  （cf. シンセティックメディア）
- 狭義：人をだます目的で、写真、音声、映像の一部を入れ替えて本物そっくりに合成された偽メディア



画像生成AIによるフェイクニュース

ドローンで撮影された静岡県の水害。
マジで悲惨すぎる...

午前4:39 · 2022年9月26日 · Twitter for Android

https://twitter.com/kuron_nano/status/1574121450860007424

米国防総省の付近で爆発（偽画像）
→ドル円が急落

政治的人物のディープフェイク

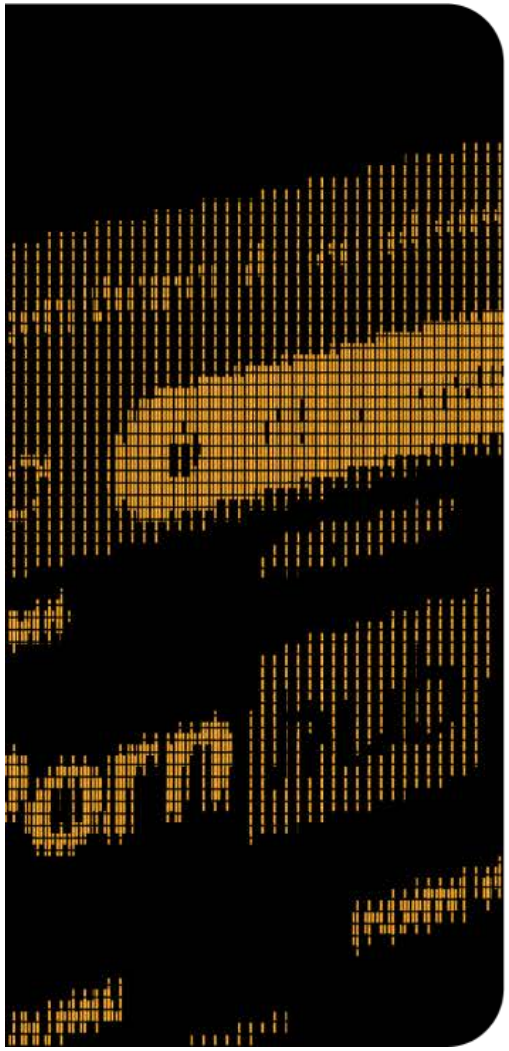
政敵への攻撃、印象操作、民主主義の混乱、
外国からの干渉、風刺

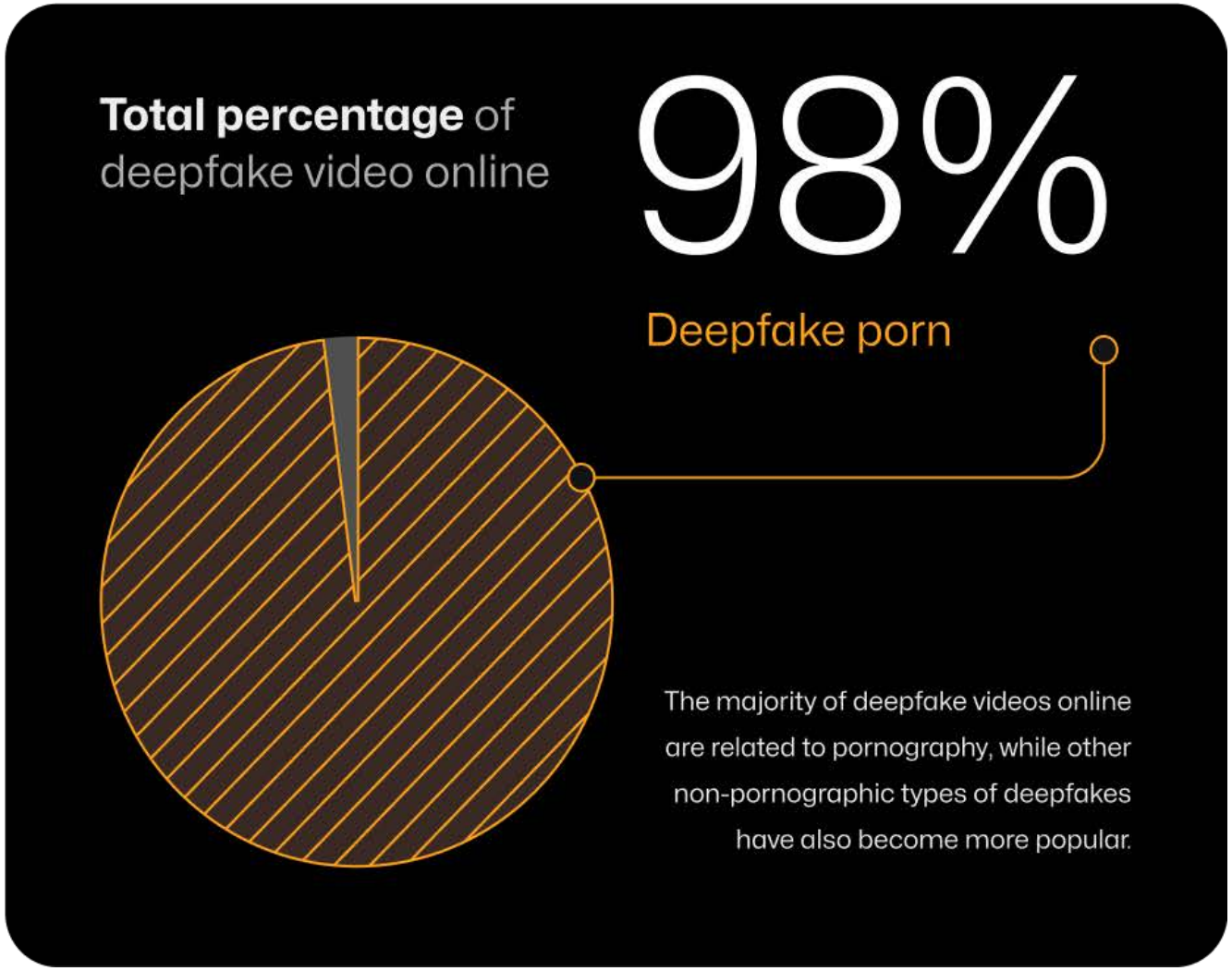
2022/3/16

2023/6/5

2023/11/2

ディープフェイクのほとんどはポルノ

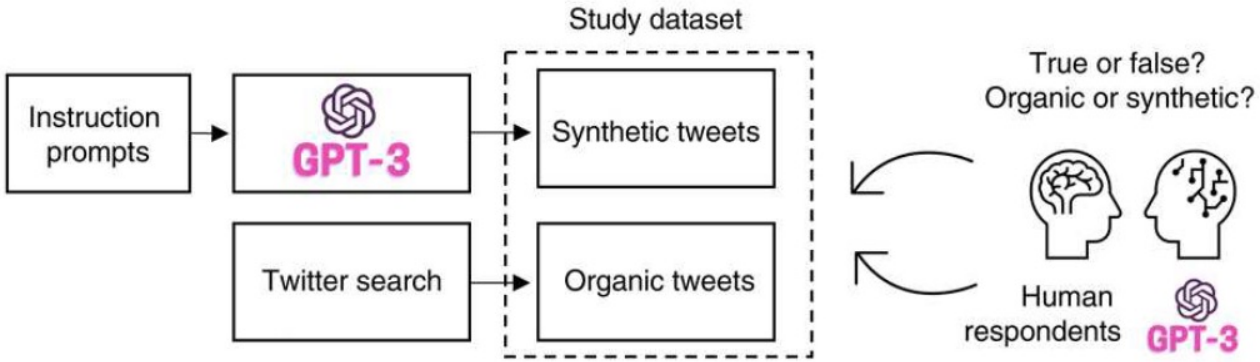


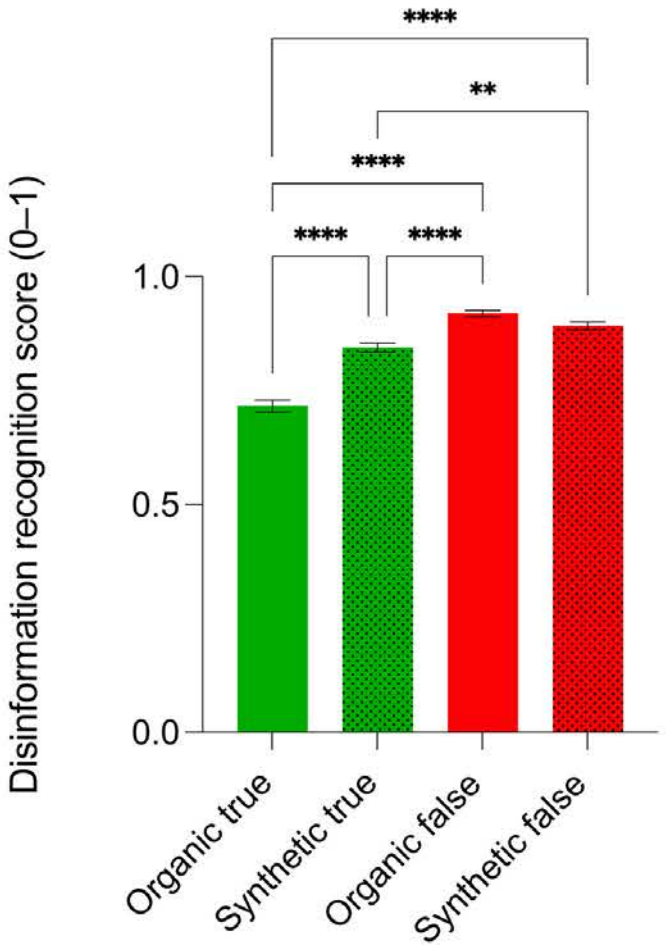


2023 State Of Deepfakes: Realities, Threats, And Impact (homesecurityheroes.com)

AIが生成した記事は説得力が高い

- GPT-3は人間よりも説得力のある偽情報を作り出す
- 人間は人間による情報とAIが作成した情報を区別できない





G. Spitale, et al. Science Advances (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

発表内容

1. フェイクニュース現象の振り返り
2. 生成AIによるフェイクの高度化
3. 介入によるフェイクの拡散抑止
4. 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

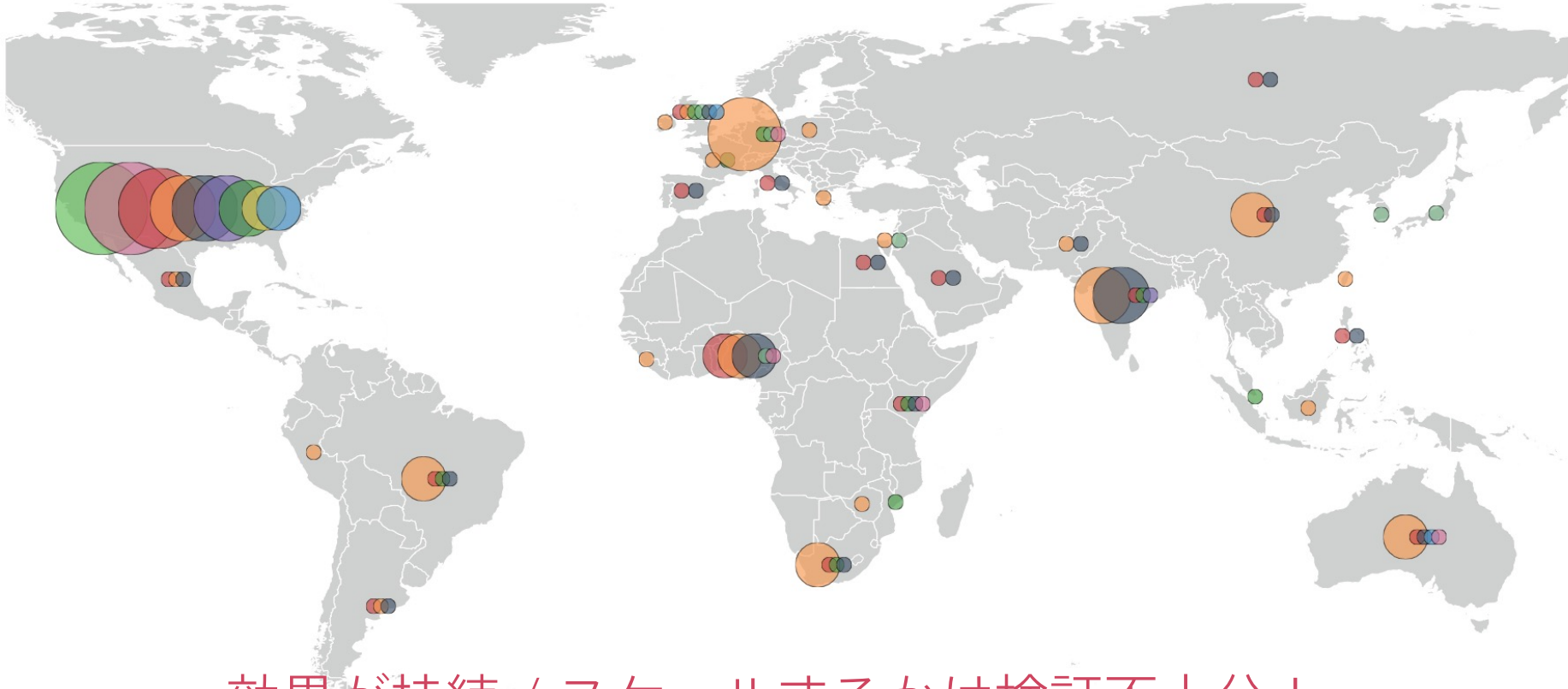
Check for updates

個人的介入の種類

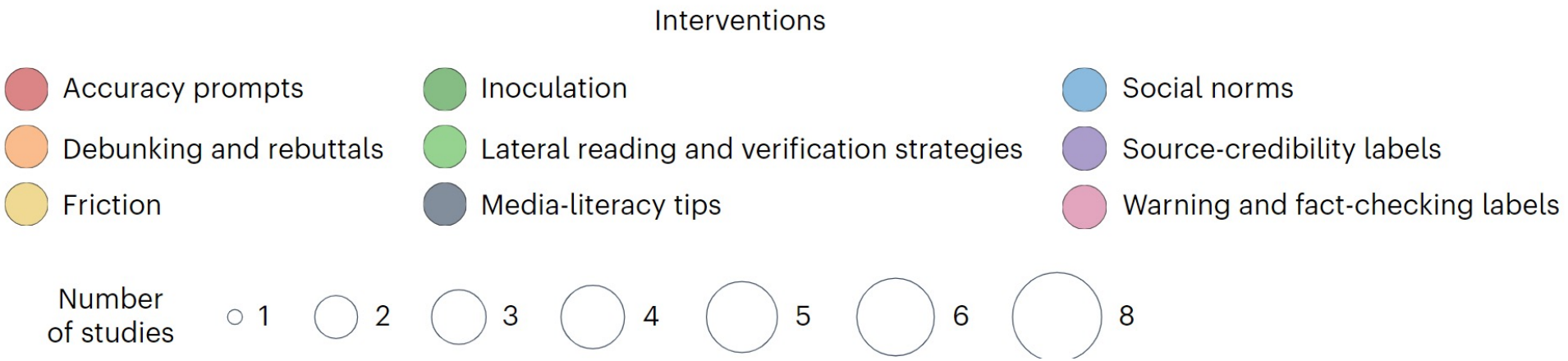
- ナッジ：
  正確性のプロンプト、摩擦、社会的規範
- ブースト・教育的介入：
  接種理論、横読みと検証、メディアリテラシーのヒント
- 反論戦略：
  プレバンキングとデバンキング、警告とファクトチェックのラベル、出典信頼性のラベル

個人的介入の効果が検証されている国々

Kozyreva, A. et al. Nature Human Behaviour (2024)



効果が持続 / スケールするかは検証不十分！



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%

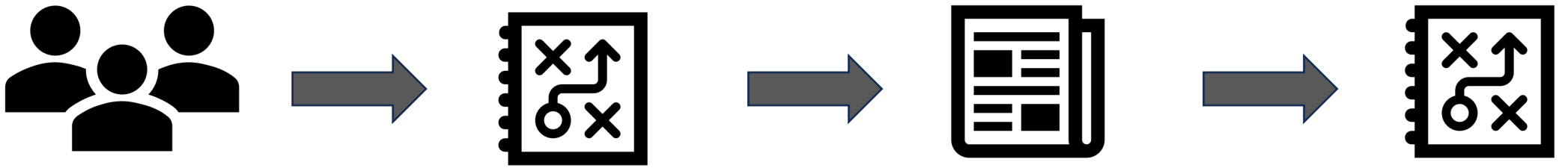
Pre-roll videos on YouTube drove up to an 11% increase in confidence, three weeks after exposure. *(Source: Jigsaw)*

画像差替によるディープフェイクの拡散抑止の検討

w/ 市川慧、陳佳玉

Presented at IC²S² 2024

サーベイ実験



Prolific
- 米国在住
- 18歳以上
- N=250（各条件）

Qualtrics
- 基本属性
- 政治的傾向
- 認知傾向（CTR）
- SNSの利用傾向

など

Qualtrics

AI生成コンテンツ（政治、エンタメ）の共有意欲
- 実験1：感情価の影響
- 実験2：画像差替

Qualtrics
- ディープフェイクの知識・経験
- 共有理由（感情等）

など

どのようなAI生成コンテンツが拡散されるのか？

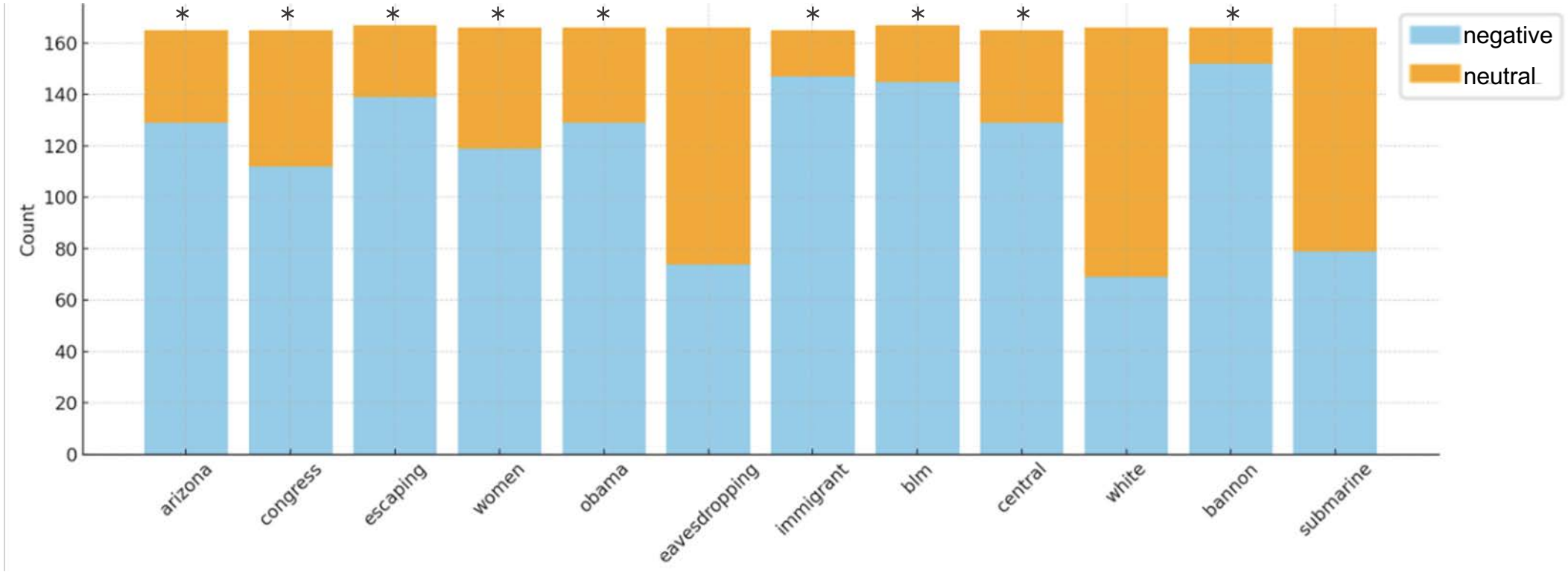
提示されたニュース（政治）に対して、感情価が異なる画像のうちどちらが一緒に表示された場合に、よりSNSで共有したくなるかを質問（1人12問，参加者250名）

GPT-3で作成→ A rally against the recent election fraud is scheduled to be held again at the Capitol starting early today.（最近の選挙不正に対する抗議集会が、本日早朝から再び議事堂で開催される）

Stable Diffusionで作成→

left　　right

ネガティブ画像を伴う政治記事は共有される



- 9/12個のニュースで、ネガティブ画像が統計的有意に選択されている
- 残りうち2個は、画像とニュースの内容が合致していない可能性あり

K. Ichikawa, J. Chen, and K. Sasahara (in prep.)

ネガティブ画像の共有と関係する変数

区分	説明変数	回帰係数	z値	p値
基本属性	年齢	0.0030	0.063	0.950
	性別（男性）	0.1168	1.067	0.286
	学歴	-0.0174	-0.377	0.706
	党派心(共和党)	0.1025	0.756	0.450
AI画像,CRT,SNS/internet使用率など	**画像共有率**	**-0.1499**	**-3.146**	**0.002**
	DF作成経験の有無	0.0729	1.444	0.149
	画像の信頼性	-0.0302	-0.635	0.526
	記事の信頼性	**-0.1012**	**-2.098**	**0.036**
	DFを知っているか否か	-0.0162	-0.093	0.926
	Internet使用率	-0.0134	-0.268	0.788
	SNS使用率	-0.0268	-0.525	0.600
	CRT正答率(平均値)	-0.2916	-1.472	0.141
感情価(Paul Ekmanの6感情)	怒り	-0.4370	-1.728	0.084
	嫌悪	0.0324	0.180	0.857
	恐怖	**-0.7357**	**-2.894**	**0.004**
	喜び	0.8624	2.893	0.004
	悲しみ	0.0247	0.056	0.955
	驚き	-0.1510	-1.148	0.251

K. Ichikawa, J. Chen, and K. Sasahara (in prep.)

「画像を差し替える」という介入

President Trump publicly cursed at a woman who expressed disapproval of him.

Your previous sharing rate 0.

I definitely prefer to share the previous image. I definitely prefer to share this image.

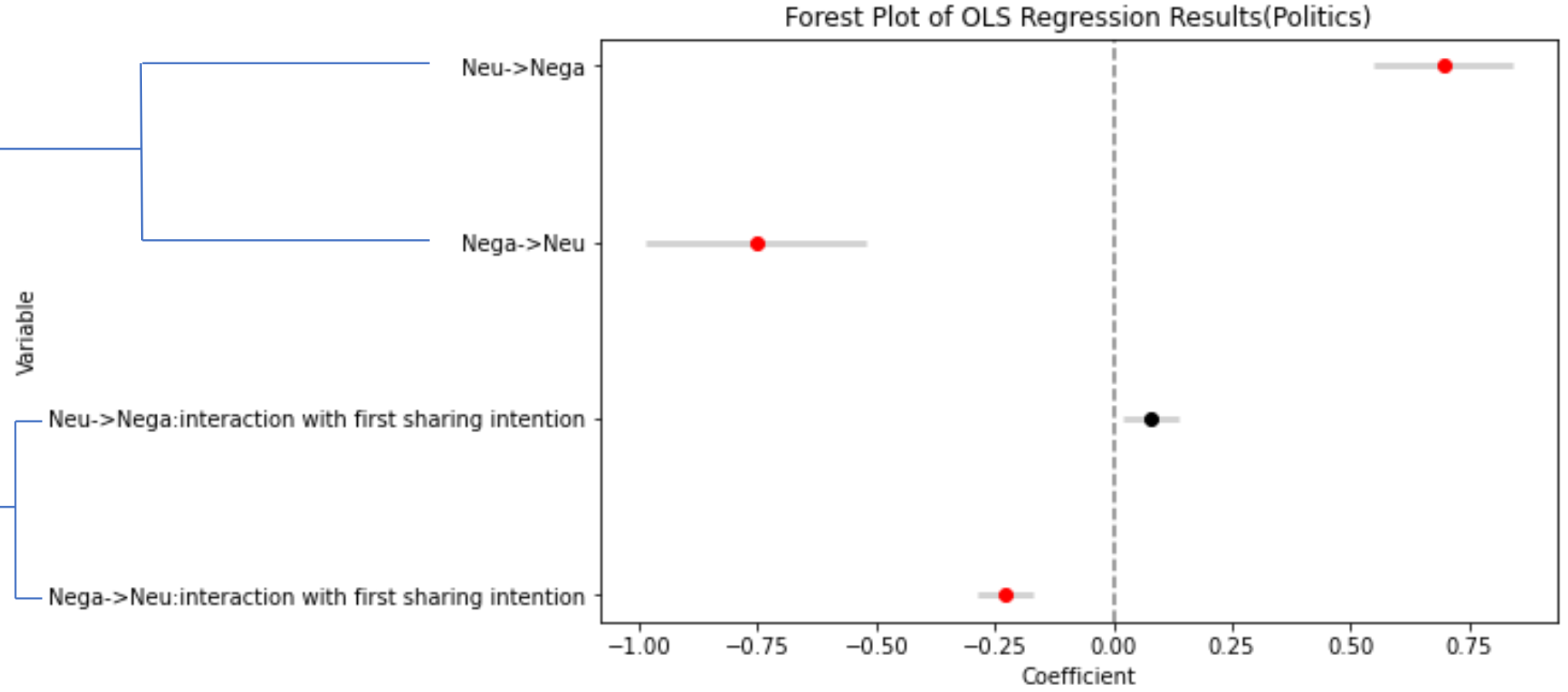
-5 -4 -3 -2 -1 0 1 2 3 4 5

ネガティブ画像を差し替えると共有意欲が減少

政治記事における画像の感情価を操作：ネガティブ→ニュートラル

主効果

交互作用を考慮

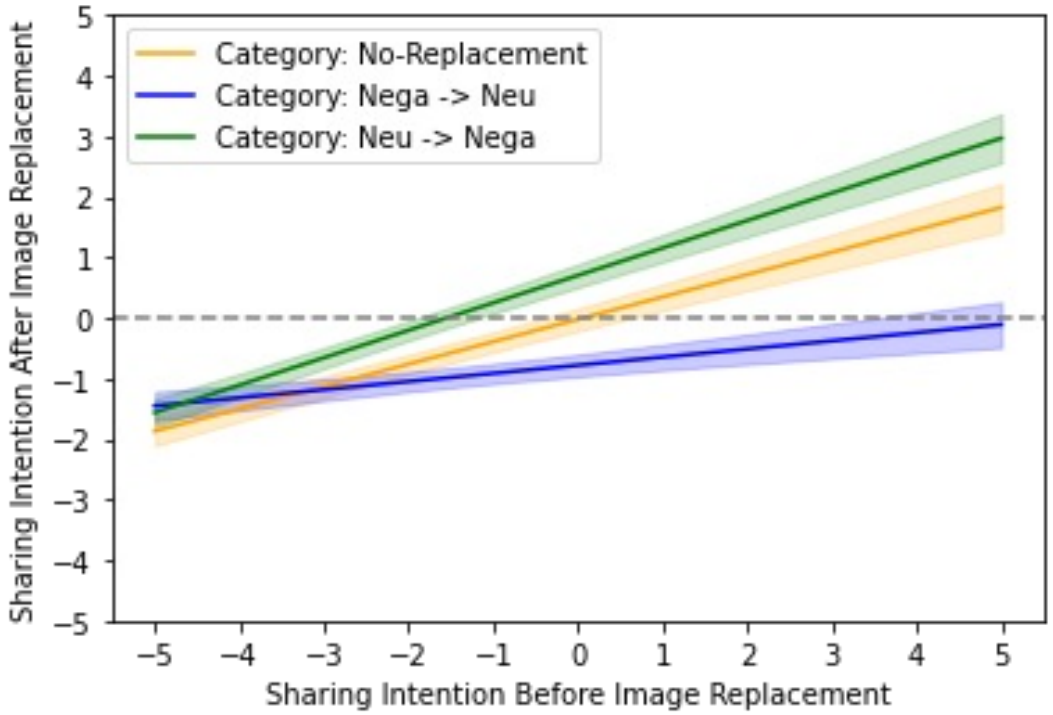


K. Ichikawa, J. Chen, and K. Sasahara (in prep.)

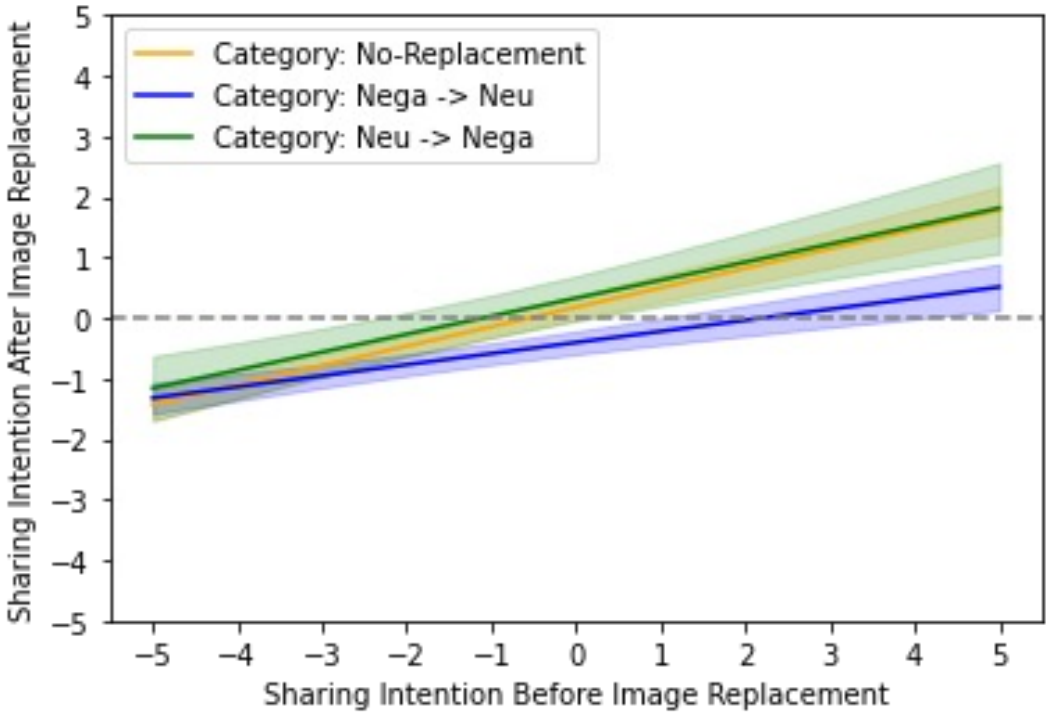
共有意欲の強さと差替効果の交互作用

- 政治記事の場合、元の共有意欲が強くても、画像差替の効果がある（共有が減る）
- エンタメ記事の場合、画像差替で逆効果になる場合もある（共有が増える）

政治



エンタメ



差し替え後の共有意欲

差し替え前の共有意欲

K. Ichikawa, J. Chen, and K. Sasahara (in prep.)

ここまでのまとめ

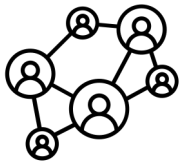
- 政治に関するAIコンテンツは感情価が共有に大きく影響
  - 感情価を抑えた画像に差し替えることで、元々の共有意図の強さを考慮しても、共有意欲が減少
- エンタメに関するAIコンテンツは多様な要因が共有に影響
  - 感情価を抑えた画像に差し替えると、元々、共有意図が強い人は返って共有意欲を促進する危険性がある
- AIコンテンツが当たり前化した社会で望まれる介入技術の設計原理

発表内容

1. フェイクニュース現象の振り返り
2. 生成AIによるフェイクの高度化
3. 介入によるフェイクの拡散抑止
4. まとめ

プラットフォームに埋め込まれた認知

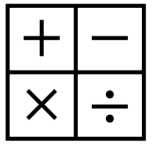
ホモフィリー



エコーチェンバー

情報環境の
局所最適化

アルゴリズム



フィルターバブル

情報発信・収集
コミュニケーション
つながり

What is
engaged
with

What is
shown

What is
thought

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

CREST
FakeMedia

意思決定
支援

合意形成
支援

モデル
学習

FM無毒化

モデル
学習

FM無毒化

なりすまし

思考誘導

誤動作

誤判定

Fake Media
(FM)

XAI

FM検出・防御

XAI

FM検出・防御

Fake Media
(FM)

Security
フェイクメディア検出・無毒化

NII 越前G

Computational Social Science
思考誘導の排除，偽情報拡散の抑制

東工大
笹原G

Multimedia
フェイクメディア生成・無毒化

阪大
馬場口G

K Program

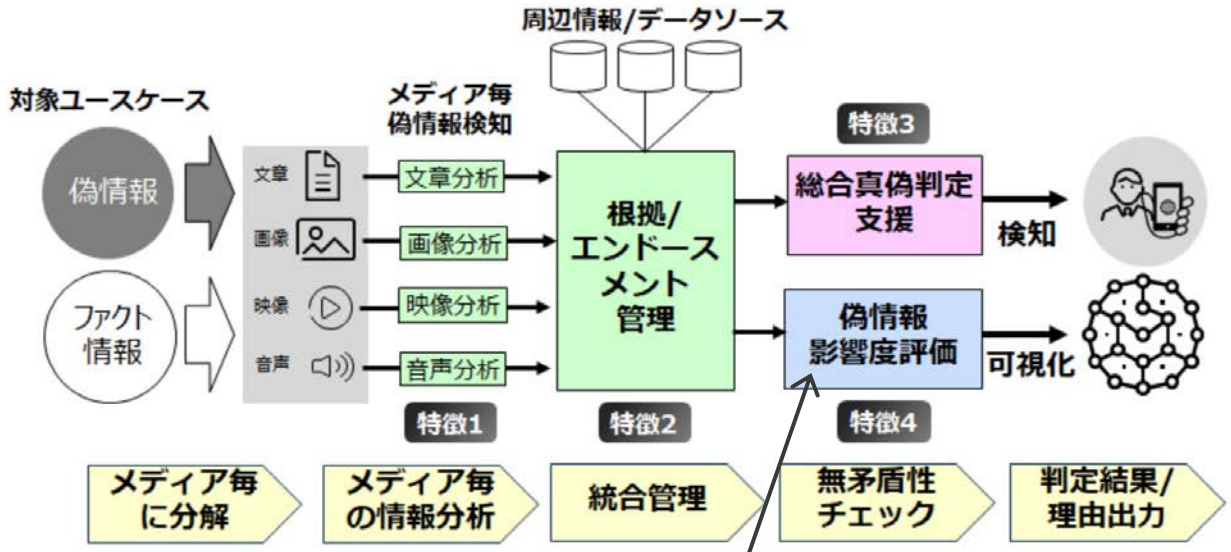
PRESS RELEASE

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。



笹原（東工大）、豊田（東大）、橋本（会津大）