生成AI時代の偽誤情報対策

NICT サイバーセキュリティ研究所 サイバーセキュリティ研究室 研究員
松田美慧

2025/06/27@情報通信セミナー2025 in 静岡

自己紹介：松田 美慧 (まつだ みさと)

- **略歴：**
  - ✓ **2023.3　中央大学大学院理工学研究科情報工学専攻（情報通信工学研究室）　修了**
  - ✓ **2023.4～現在**
    - **NICTサイバーセキュリティ研究所サイバーセキュリティ研究室 研究員**
    - **横浜国立大学大学院環境情報学府 博士課程後期3年**
- **研究テーマ：ソーシャルメディア上の偽誤情報や有害情報の特徴分析ならびに検出技術の開発**
- **従事している業務：セキュリティ分野のキュレーションを自動化するプロジェクトなど**
- 主な論文
  - ✓ 松田美慧, 藤田彬, 吉岡克成: デマの検知に向けたモダリティの活用の検討, 情報処理学会論文誌66(3), pp.1-16, 2025.
  - ✓ 松田 美慧, 川口 大翔, 藤田 彬, 吉岡 克成: オンライン詐欺と犯罪へ誘導するSNS投稿文の類型化と特徴の分析, コンピュータセキュリティシンポジウム 2023, 2023.（CSS2023奨励賞受賞）など

NICTサイバーセキュリティ研究所の紹介

NICTが実施する重点5分野の1つ「サイバーセキュリティ」を担う研究所

CYBERSECURITY Research Institute

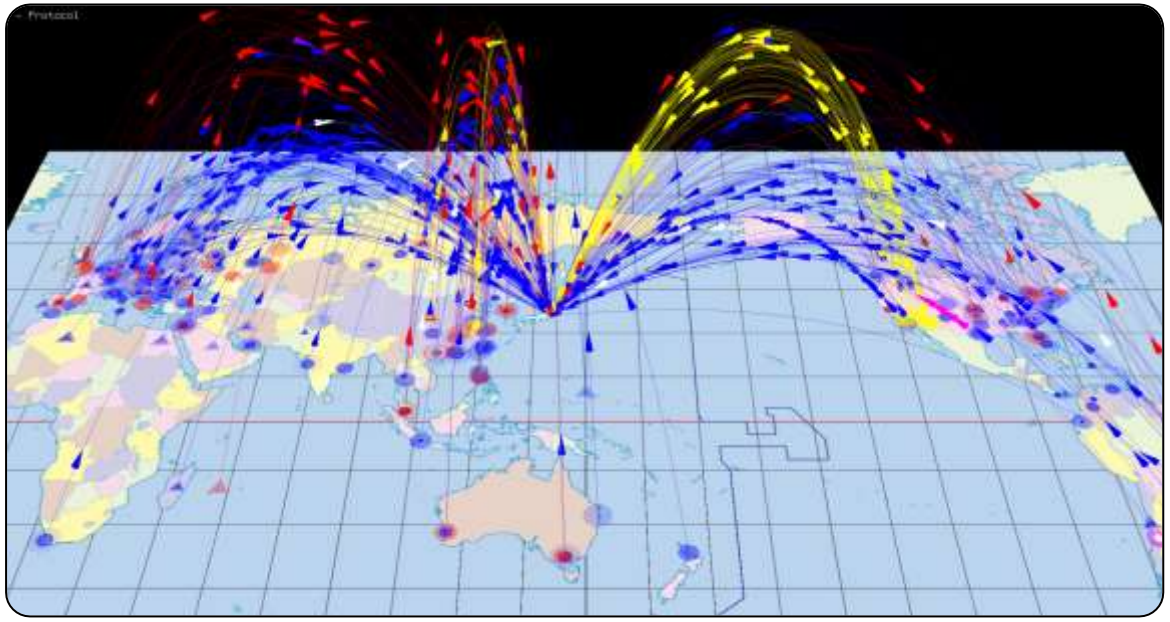
SECURITY FUNDAMENTALS Laboratory	CYBERSECURITY Laboratory	CYNEX CYBERSECURITY NEXUS	National Cyber Training Center	NATIONAL CYBER OBSERVATION CENTER	NEW CREATE（2025年2月設立）
暗号認証技術の研究開発	CS技術の研究開発	産学官の『結節点』	セキュリティ人材育成	IoT機器の脆弱性調査	AIセキュリティ

NICTサイバーセキュリティ研究室

データ駆動型サイバーセキュリティ技術

- サイバー攻撃の観測・データの蓄積・可視化・分析・周知・対策検討

例：NICTER（無差別型攻撃の可視化）



エマージングセキュリティ技術

- 安全と使いやすさの両立
- 5G/Beyond 5G
- クラウドセキュリティ　　など

例：ユーザ調査研究の地理的偏りを証明

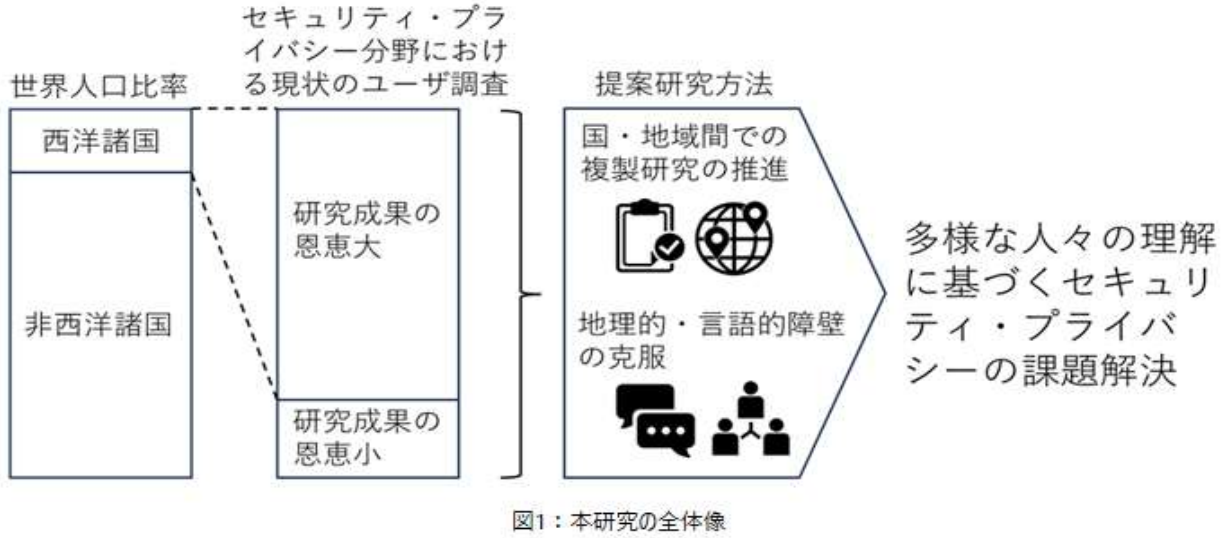


図1：本研究の全体像

引用：NICT，https://www.nict.go.jp/press/2024/09/03-1.html

クイズ：本物でしょうか？偽物でしょうか？

問題１

問題２

10秒間でお答えください

クイズ：本物でしょうか？偽物でしょうか？

答え：両方とも“偽物”（発表者が生成AI*を用いて作成）

*imageFX (https://imagefx.org/ja)を使用

生成AIへの指示

“日本の「地方銀行」が倒産するという噂を聞いた100名を超える人が銀行の入り口に押し寄せている様子”を出力して下さい。

生成AIへの指示

“大企業のオフィスで、30代の日本人男性社員が60代の日本人女性役員に大勢の社員の前で叱られている様子”を出力してください。

生成AIで作られた画像の見分け方の例

文字やサインの異常

・文字が読めるか？
・意味が通じるか？

”似せ字”に注意

人の顔の不整合

・パーツの位置は？
・間隔が似てない？
・同じ人はいない？

状況の不自然さ

・背景のピントは？
・表情や視線は？

文脈を踏まえた
リアリティに注意

手・指・関節の不整合

・手や指の本数は？
・手と腕の接続が自然か？

テクスチャや照明の矛盾

・光の当たり方は？
・髪の毛の質感は？
・布の質感は？

しわや滑らさに注意

生成AIの悪用の危険性：不自然な画像でも効果あり

ご近所お兄さん

A銀行が倒産するの、マジだわ…

2025年XX月XX日・5.6M 閲覧

コメント 130　いいね3.5k　引用10k

普通OLりのまる

○○役員のパワハラ常態化しすぎ

2025年XX月XX日・7.8M 閲覧

コメント 561　いいね5.5k　引用7k

ナラティブ（物語）の付与などにより、事実の様な印象を与えることは可能

生成AIで顕在化したリスク

- **悪用：詐欺や世論操作などを目的とした利用事例を複数確認** [総務省 (2024)]
  - ✓ 2024年2月 香港の多国籍企業で計2億香港ドル(約38億円)規模の詐欺事件[CNN (2024)]
  - ✓ 2022年9月 日本で静岡の水害を撮影したとする偽画像がTwitter上で拡散[JFC (2022)]
- **ハルシネーション：事実に反する情報の生成** [総務省 (2024)][JITERA(2024)]
  - ✓ 2022年11月 Metaの科学分野LLM「Galactica」が架空の人物や理論を生成→3週間で運用停止
  - ✓ 2023年2月　Googleが公開した会話型AIサービス「Bard」で多数のハルシネーションを確認
  - ✓ 2023年6月　ChatGPTに個人的な偽誤情報を生成されたMark Walters氏はOpenAIを提訴
- **バイアスの再生成：既存の情報に含まれれるデータの偏りを増幅** [総務省 (2024)]
  - ✓ 2022年4月　OpenAIの画像生成AI「DALL-E2」で性別・人種への偏りを確認 [日経ビジネス (2023)]

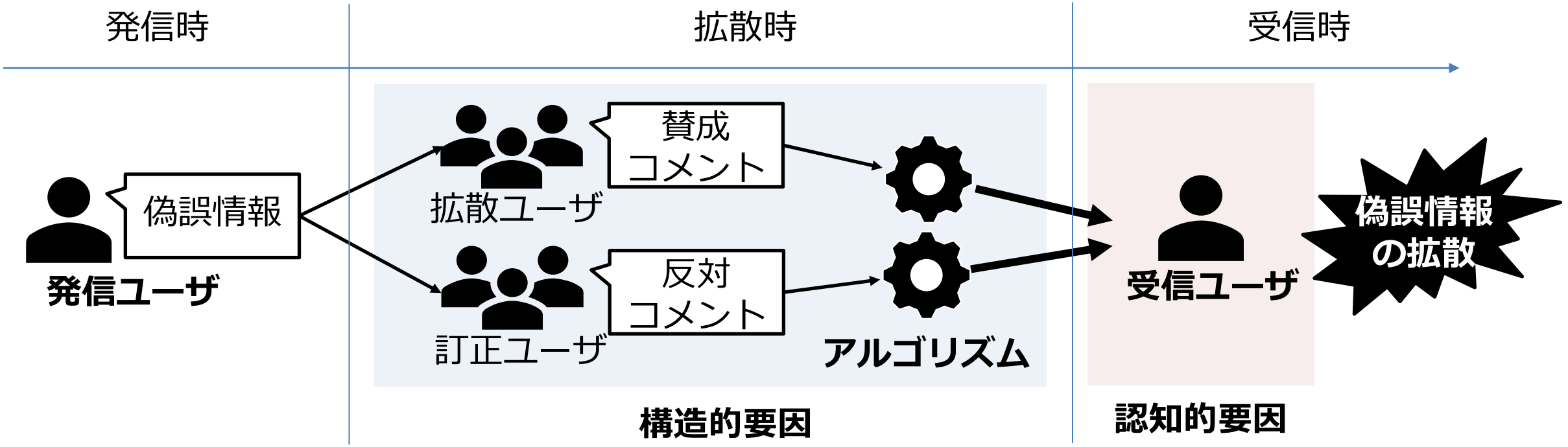
様々なリスクが顕在化しているが，生成AIの利用を止められない

➡　**生成AIとの共生時代へ**

目次

- そもそも偽誤情報の拡散要因は何か？
- 真偽検証は自動化できるのか？
- 真偽検証の自動化により偽誤情報の拡散は解決するのか？
- 生成AIとどのように共生するか？
- 将来の対策検討は？

偽誤情報拡散のエコシステムと要因

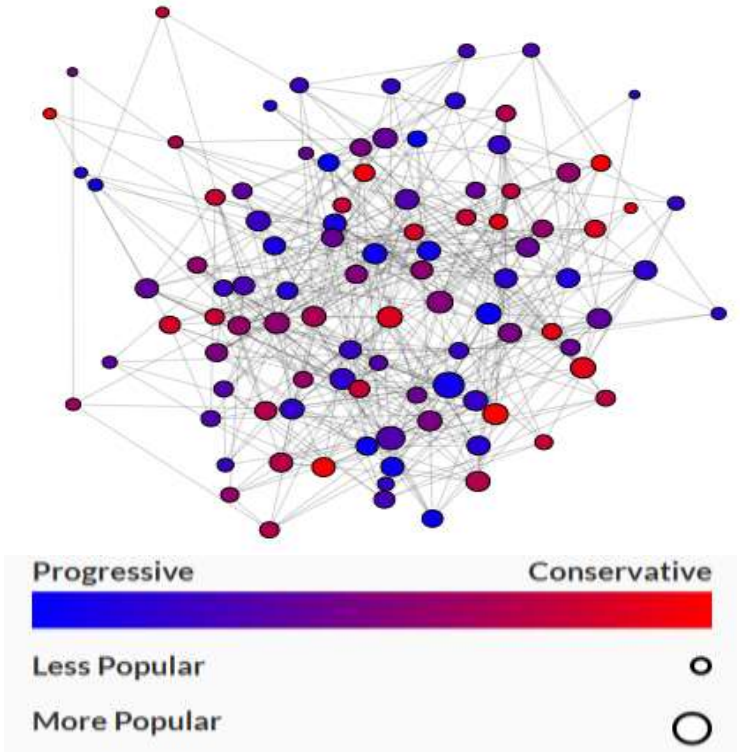


偽誤情報の拡散は，構造的要因と認知的要因により発生

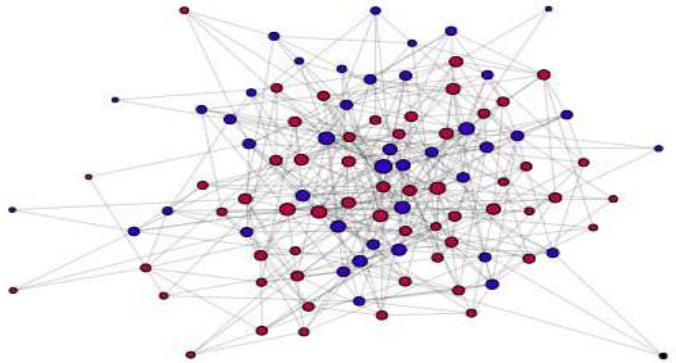
- ✓構造的側面：受信ユーザを取り巻く環境　例：サービスの特性、法規制など
- ✓認知的側面：受信ユーザの情報処理能力　例：認知バイアス、感情喚起など

構造的要因：便利さに最適化されたアルゴリズム

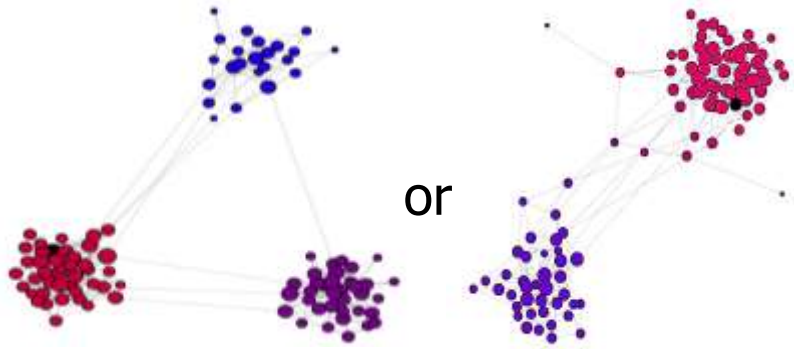
- フィルターバブル：パーソナライズ機能がユーザにとって好ましい情報だけを表示
- エコーチェンバー：フォロー機能などが同じ意見を持つ人だけと交流を促進
  - フォロー機能付きSNSを使うだけで分断が起きる [Hao(2017)][Sasahara(2023)]



- 意見の許容範囲: Medium
- 社会的な影響　: Strong
- フォローの解除: Never



- 意見の許容範囲: Medium
- 社会的な影響　: Strong
- フォローの解除: Often



構造的要因：SNSの安全は誰の責任か？

事業者の基本姿勢

- **自由な言論空間の保持**
  - ✓2019年：Mark Zuckerberg（Meta CFO）がGeorgetown大学の講演にて

    何が信用できるかを決めるのは技術企業ではなく、人々であるべきだ。
    I believe people should decide what is credible, not tech companies.
  - ✓2025年：Metaはファクトチェック機能を終了し，コミュニティノートへ移行

- 悪質コンテンツの責任は利用者にあり，プラットフォームは広場にすぎない
  - ✓2018年：Jack Dorsey (Twitterの当時CEO)がCNNの取材にて

    私たちのような会社が真実の裁定者になるのは危険だと思う。
    I do think it would be dangerous for a company like ours … to be arbiters of truth.

SNS運用事業者などに対する規制の動向

2020年以降，規制が進み，事業者の責任義務が一部規定される

積極的

消極的

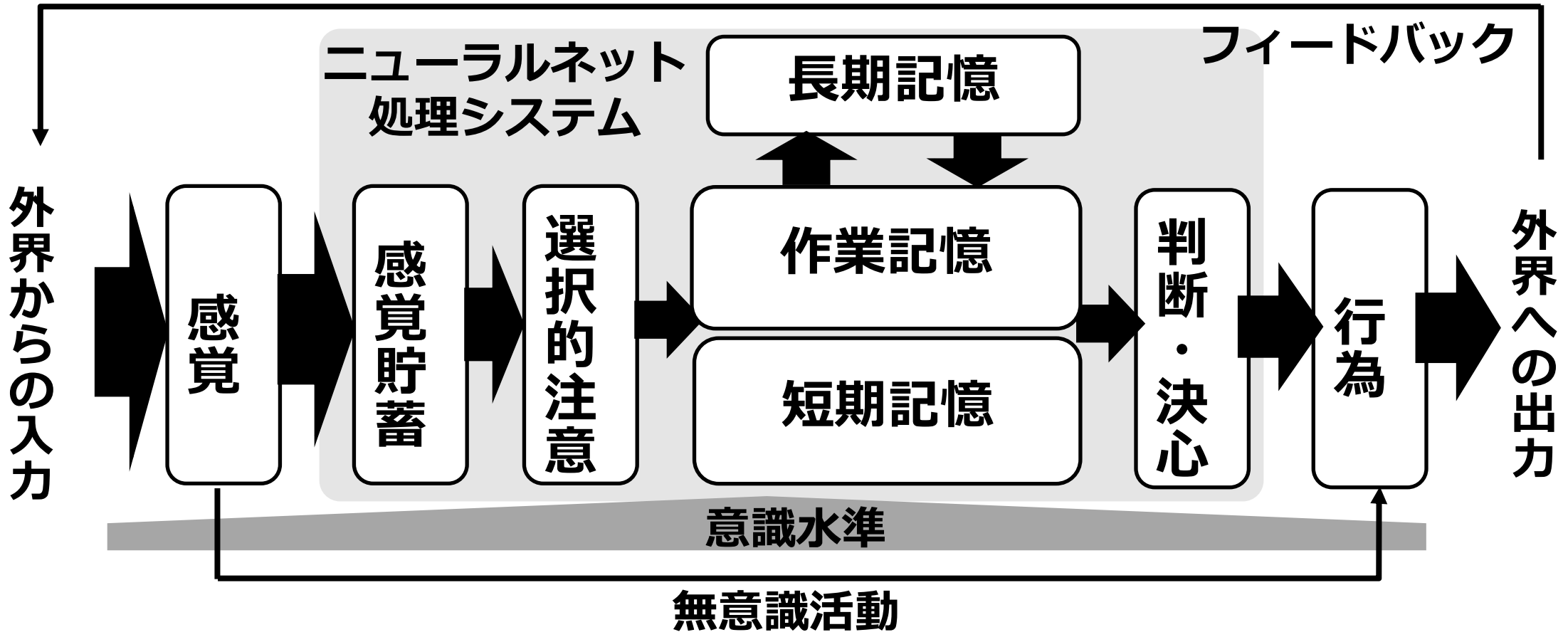
地域	主な法制度	対象事業者	規制内容
EU	**Digital Services Act (DSA)** （2024年2月 全面適用）	全IT企業	◎違法・偽情報の削除対応 ◎リスク評価と軽減策 ◎違法コンテンツへの対応報告 ◎データアクセス提供
中国	**・サイバーセキュリティ法** **・インターネット情報サービス ディープフェイク管理規定**	全サービス提供者	◎コンテンツ検閲（国家指導） ◎AIコンテンツの識別とログの保存 ◎利用者の事前認証
インド	**2021年情報技術規則（IT Rules 2021）**	SNS・メッセンジャー事業者	◎苦情への対応（24時間以内） ◎特定のコンテンツへの年齢確認
日本	**・プロバイダ責任制限法** **・情報流通プラットフォーム対処法** （2025年4月施行）	SNS事業者・大規模プラットフォーム	◎プロバイダの責任の制限 (放置は賠償責任) ◎違法コンテンツの削除対応 ◎発信者情報開示への協力 ◎削除結果の報告 ※偽・誤情報は削除の対象外
米国	**・通信品位法(CDS) Section230** (2023年6月改正) **・各州法（例：CAADCA）**	プラットフォーム	⚠ 明示的義務は少ない（現時点） ・コンテンツ仲介における免責 ※AIが作成したコンテンツは免責の対象外

構造的要因のまとめ

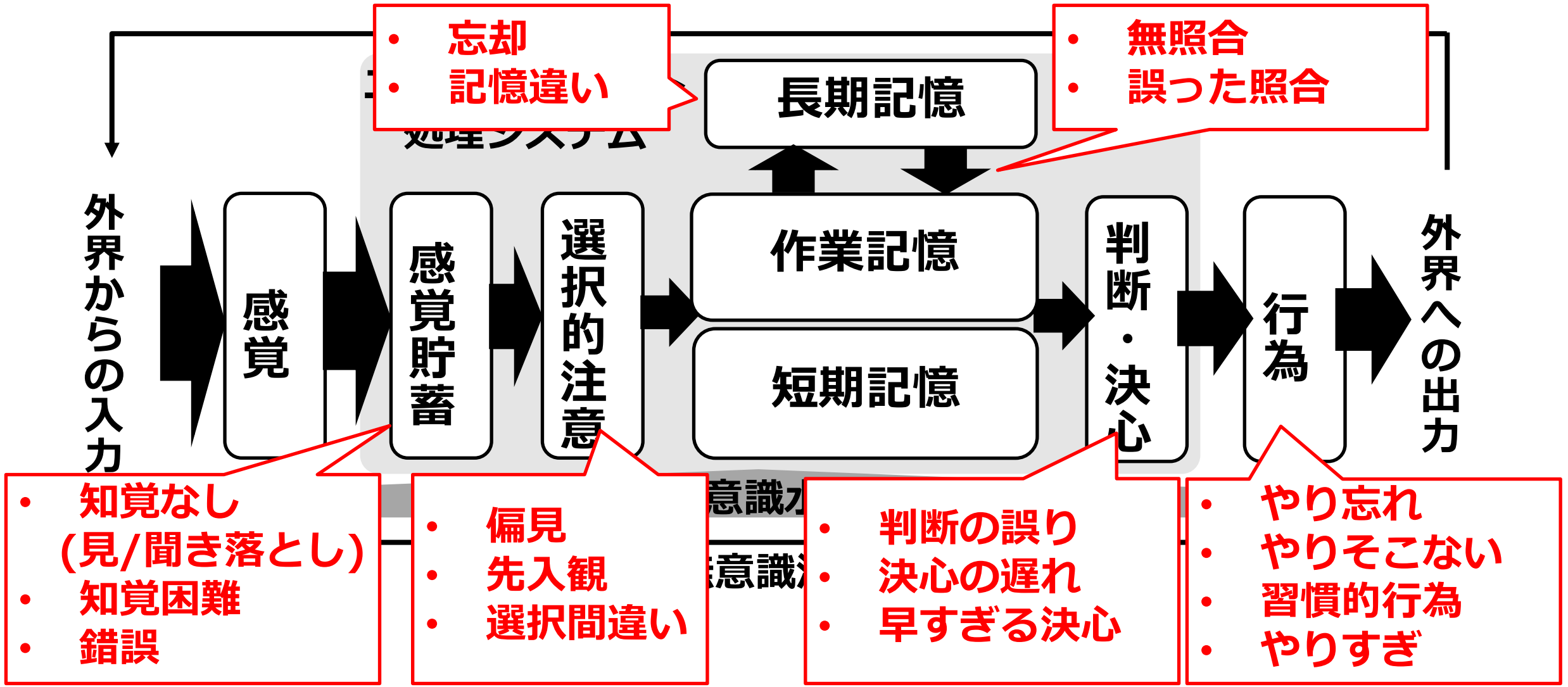
- アルゴリズムの特性
  - ✓社会分断を促進する特性　例：エコーチェンバー，フィルターバブルなど
- プラットフォーマの義務責任
  - ✓基本姿勢は，自由，利益追求の保持
  - ✓法規制により一部の地域でプラットフォーマに対する規制が始まる
  - ✓ただし政治動向に大きく左右される

NEXT▶認知的要因

認知的要因の前提知識：人間の情報処理モデル



認知的要因：ヒューマンエラー



認知的要因：人間は偽物を見破れるのか？

- **平均識別率：約50~70%（ランダムよりも少し高いくらい）**

文献情報	対象	真陰性 (偽物を偽物)	真陽性 (本物を本物)	真陰性 ＋真陽性
Bray et al. (2023)	画像（人間の顔（白人男女））	**52%**	**68%**	**59.7%**
Lu et al. (2023)	画像（人物，風景，動物，植物等）	**56%**	**67%**	**61.3%**
Korshunov et al. (2020)	動画（人物）	**62% (24.5%~88%)**	**82%**	**約70%**
Bond et al. (2006)	テキスト（英語）※社会学的分析	**54%**	**61%**	**54.0%**

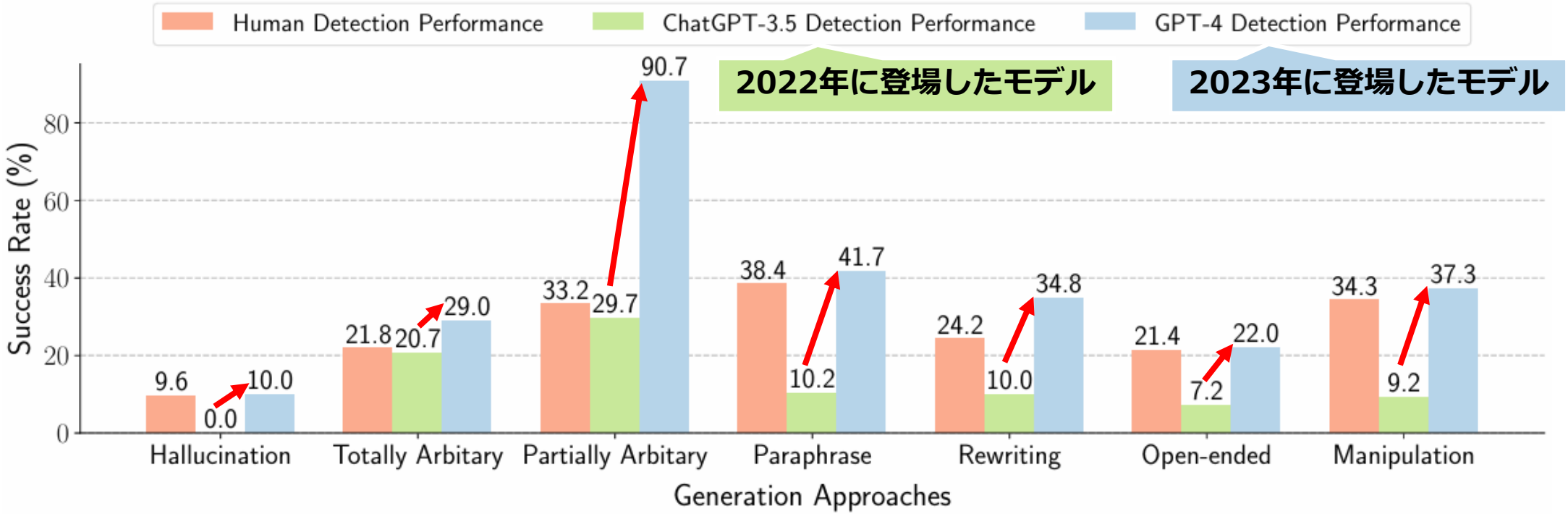
- **偽物に気付くことは本物に気付くよりも難しい：約50~60% ＜ 約60~80%**
- **質の良い偽物の識別率は24.5%** (Korshunov et al. (2020))

➡**人間は偽物を識別できない前提での対策の検討が求められる**

生成AIで真偽ができるか？

- 性能はモデルに依存…現在は真偽判別性能は人間と同等以上
  - ✓例：LLMでは，1年間で性能は最大約60%↑[Chen(2023)]

➡<u>将来的には，偽誤情報の真偽検証の自動化は可能になるはず</u>！ただ鼬ごっこは続く



将来にわたって残留する課題：人間の認知バイアス

- **選択的回避**：信念と異なる情報を避ける
- **正常性バイアス**：都合の悪い情報を無視あるいは過小評価する
- **確証バイアス**：自分に都合のいい情報ばかりを収集して反証をしない
- **心理的リアクタンス**：指摘されるとやりたくなくなる
- **確実性効果**：不確実なものより確実なものを好む
- **認知的不協和**：矛盾する認知において不都合な認知を変え，正当化する
- **誤情報持続効果**：訂正後も誤情報を信じ続ける（信じ切っている場合も）
- **根本的な帰属の誤り**：性格など内的要因を重視し，状況要因を軽視する
- **真実(性)錯覚効果**：同じ情報に繰り返し接触し，情報処理の流暢性が高まる
- **バックファイア効果**：対立する意見を受け入れず，自分の考えをより強固にする

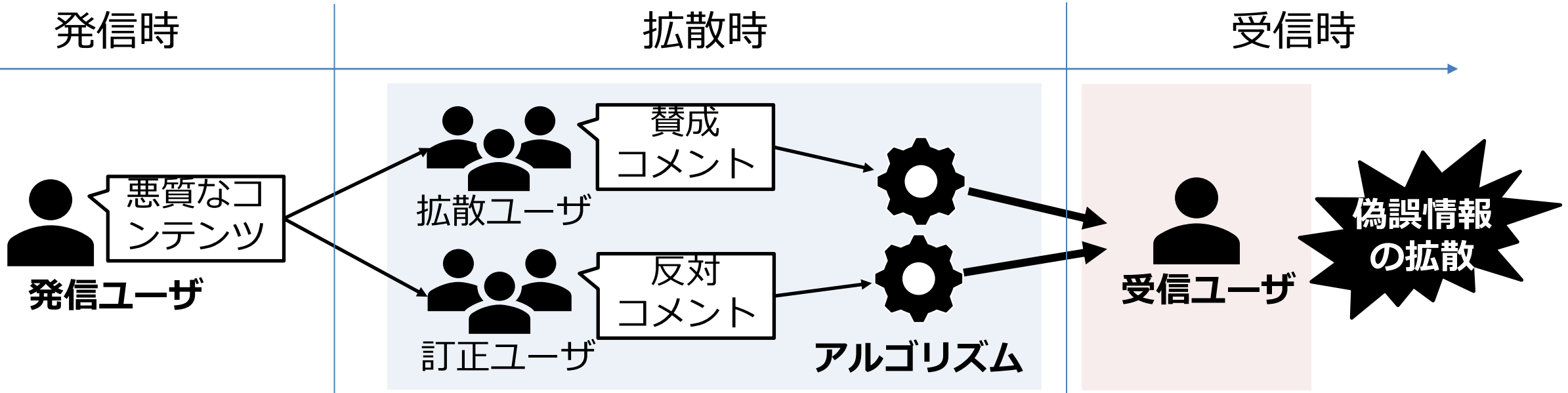
➡**真実を提示しても，信念に反する真実が受け入れられない**

認知バイアス以外の影響因子

- **感情的バイアス**：怒り・恐怖に敏感．特に義憤（道徳的怒り）は炎上しやすい
- **情報源の信頼度**：投稿者の知名度，専門性や人口属性の類似性で判断される
- **人口属性**　　：性別・年齢は拡散行動に影響．ただ正負は議論が割れている
- **リテラシー**　　：メディアリテラシ，情報リテラシ，批判的思考が注目される
- **文化的文脈**　　：国や地域によって関心を持たれやすい話題が異なる
- **政治的信念**　　：支持する党派や政治家の主張との整合性が政治話題には影響
- **SNS利用習慣**　：SNSを利用する目的や利用時間などを含む

➡SNSの情報に対する拡散行動や真偽判断には，多数の要因が複雑に影響する

偽誤情報拡散のエコシステム with 生成AI



生成AI

偽誤情報を生成
➡対策リソースのひっ迫

生成AI

・生成AIボットと人間の見極め困難
・敵対的コメントで真偽判断能力×
➡世論操作の容易化

[Lu et al. (2025)]

生成AI

・バイアス回避の提案
・真偽検証の自動化
➡対策への貢献

主な対策と課題

- 主な対策
  - ✓ユーザ　　：リテラシー教育→内容の注意深い確認，通報
  - ✓プラットフォーマー：投稿の削除，アカウントの凍結措置
  - ✓ファクトチェッカー：事実確認の実施，結果の公開
  - ✓官公庁組織　　：不審な投稿に対する通報の受付，捜査活動，法的措置
    （アカウントの情報開示請求を可能に）
- 課題
  - ✓訂正情報や注意喚起情報の効果が薄い＝人は信じたいものを信じる
  - ✓思想・言論の自由との両立＝社会的な悪影響の推定が必要だが困難
  - ✓対策の実行の判断基準が曖昧で，対策の実行までに時間を要する
  - ➡**対人心理への影響を考慮**した対策の検討が必要

ファクトチェックの課題：検証対象の偏り

- **圧倒的なリソース不足**
  - ✓日本では、1年間に合計3315件（事業者から2123件、一般電話から1192件）の疑義言説が確認されたが、実際に検証されたのは245件 [FIJ (2024)]
- **検証方法や結果を第三者が検証する体制になっていない**
  - ✓**検証対象の選定手法の詳細が十分に開示されていない**
  - ✓**異なるファクトチェッカーが同じ主張を選ぶことはほとんどない**
    - PolitiFactの1065記事とFact Checkerの240記事を比較した結果、両サイトで検証の対象となった記事は70件であった [Lim (2018)]
- **多くのファクトチェッカーは検証対象を手作業で選定**

  基準は主観的であったり、時勢の変化に左右されたりすることが多い [Guo (2022)]

➡**自動検出が可能な特徴量が必要**

対人心理への影響推定のための自動検出手法の提案

- **心理への影響を推定する特徴量として，モダリティ(modality)の活用を提案**
- **モダリティとは，言語学において話者の心理的態度を表す**

例：新型コロナウィルス対策として、すぐに緑茶を飲んでください！！

新型コロナ対策として緑茶を飲む ＋ すぐに，ください，！！

命題内容（客観的な叙述の素材） **モダリティ（主観的な態度）**

【モダリティに注目した理由】

- 受信者は発信者の意図や熱量に共鳴すると考えられるため
- 投稿文から得られる特徴は拡散前の初期段階で利用できるため

➡モダリティだけを見て偽誤情報を予測できそうかを検証

モダリティの分類

	大分類	中分類	小分類		言語学的分類	例	定義
対事的	命題的	認知的	強	**断言**	断定・説明	のだ、わけだ、ことだ、からだ、つまり	ただ事実を述べているのではなく，揺るぎない主張・論理として強調している箇所
			弱	**推測**	推量・判断・可能性・蓋然性	だろう、にちがいない、はずだ、かもしれない、まい，でしょう、思う，きっと，たぶん	自分の意見を意見だとわかる形で強調して記述している箇所
		証拠的	強	**経験**	様態	（～し）そうだ	文脈から自分の経験を根拠として記述している箇所 ※断定的な言い方であっても，根拠としての活用であれば様態とする.
			弱	**伝聞**	伝聞	らしい、ようだ、そうだ、によると	自分は当事者ではないが，他社の経験をそのまま聞き伝えている箇所
	事象的	束縛的	強	**強制**	義務・命令・禁止	べきだ、ほうがよい、（ものだ、ことだ、）なければならない、なさい、てはいけない、こと	他者に対して行動を示唆，強要する箇所
			弱	**誘導**	許可・勧誘・依頼・誘い	てもいい、ましょう、てください、てくれ、していただけませんか，ほしい	他者に対して行動をお願いする箇所
		力動的		**意志**	意志	つもりだ，したい	ー
				能力	能力	できる	ー
対人的				**対人**	確認・強調	ね、よ，！	話口調に関する箇所
モダリティなし				**なし**			モダリティがない＝命題だけの状態
				URL			

結果1-1：[推測]のモダリティについて有意差を確認

- 対象：投稿日時が早い50個の返信投稿に含まるモダリティの割合
- 結果：推測のモダリティについて，真偽判定の結果がFalseとMisleadingの間に有意差が確認された
- 考察：
  - ✓「～かも」など推量的な表現は，**事実の誤った情報と，事実を述べているが誤解を招く表現を区別する**
  - ✓確信を得ない表現が新たな推測を生みながら拡大していると考えられる

表 6: Q1: 情報の真偽と [推測] との多重比較の結果

ラベル	統計量 (p 値)
False:Inaccurate	0.875 (0.818)
False:Misleading	2.860 (0.022)
False:Unknown-Evidence	0.166 (0.999)
Inaccurate:Misleading	2.062 (0.166)
Inaccurate:Unknown-Evidence	0.373 (0.982)
Misleading:Unknown-Evidence	2.286 (0.101)

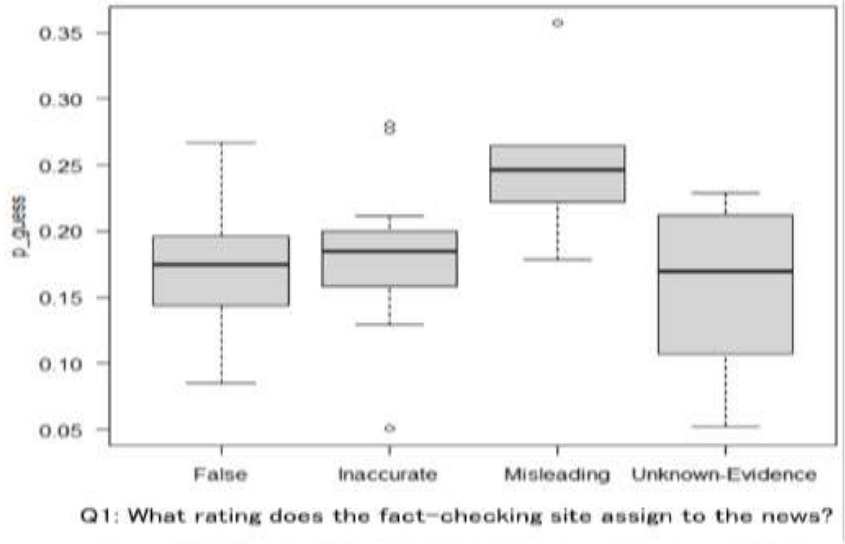


図 2: [推測] の出現割合の Q1 ラベルごとの分布

結果1-2：[意志]のモダリティと発信の目的との有意差

- 対象：投稿日時が早い50個の返信投稿に含まるモダリティの割合
- 結果：意志のモダリティについて，発信の目的がPartisanとPropagandaの間に有意差が確認された
- 考察：
  - ✓PartisanとPropagandaの違いは，誘導の有無
  - ✓～したいなどの**発信者の願望を含む表現**は，社会など他者に対する願望を含むことから**誘導の有無を分けるのに有用**である可能性あり

表 7: Q5: 発信の目的と [意志] との多重比較の結果

ラベル	統計量 (p 値)
2.Partisan:3.Propaganda	2.466 (0.036)
2.Partisan:4.No purpose/Unknown	0.689 (0.770)
3.Propaganda:4.No purpose/Unknown	2.193 (0.072)

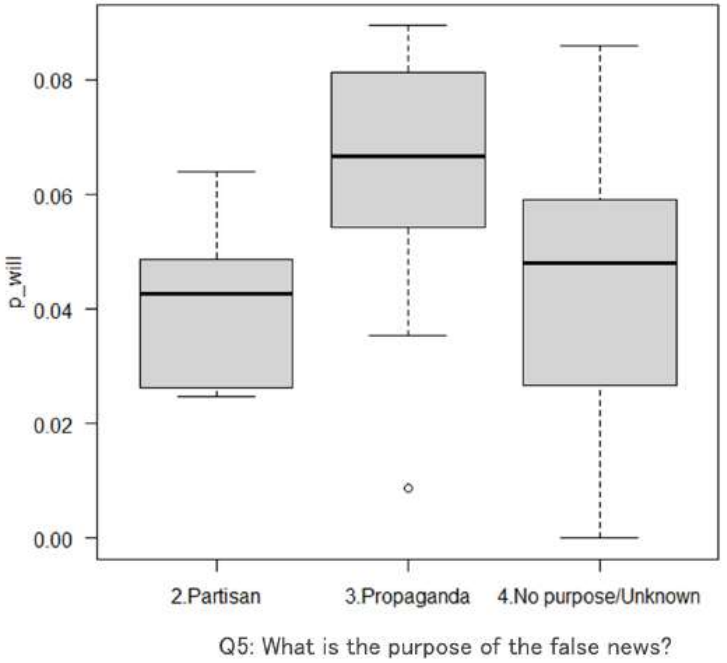


図 3: [意志] の出現割合の Q5 ラベルごとの分布

結果2:リアクション数とモダリティとの相関分析

- 対象：投稿日時が早い50個の返信投稿に含まるモダリティの出現頻度と割合
- 結果：**返信，いいね，引用，ブックマークされた数との相関**を確認
- 考察：
  - 命令や禁止を呼び掛ける表現は，いいねとブックマークを得にくい
  - 他者の話を伝え聞いている表現は，引用とブックマークの数を得やすい　など

表 8: モダリティと返信の数との相関

ラベル	First50_cnt	First50_per
断言	0.230 (0.120)	-0.156 (0.294)
推測	0.275 (0.061)	-0.079 (0.598)
経験	0.194 (0.192)	0.088 (0.558)
伝聞	0.193 (0.194)	0.017 (0.909)
強制	0.099 (0.509)	-0.087 (0.561)
誘導	0.486 (0.001)	0.356 (0.014)
意志	0.391 (0.007)	0.204 (0.169)
能力	0.230 (0.119)	0.151 (0.311)
対人	0.364 (0.012)	0.092 (0.540)
引用	-0.120 (0.421)	-0.335 (0.021)

表 9: モダリティといいねの数との相関

ラベル	First50_cnt	First50_per
断言	0.177 (0.233)	0.101 (0.501)
推測	0.223 (0.131)	-0.043 (0.772)
経験	0.267 (0.070)	0.199 (0.180)
伝聞	0.230 (0.120)	0.155 (0.297)
強制	-0.188 (0.205)	-0.305 (0.037)
誘導	0.235 (0.112)	0.161 (0.281)
意志	0.175 (0.238)	0.109 (0.467)
能力	-0.052 (0.727)	-0.077 (0.609)
対人	0.144 (0.333)	-0.025 (0.869)
引用	0.070 (0.640)	-0.076 (0.610)

表 10: モダリティと引用された数との相関

ラベル	First50_cnt	First50_per
断言	0.078 (0.602)	-0.192 (0.195)
推測	0.133 (0.372)	-0.165 (0.268)
経験	0.168 (0.260)	0.098 (0.511)
伝聞	0.340 (0.020)	0.249 (0.092)
強制	-0.110 (0.460)	-0.273 (0.063)
誘導	0.216 (0.145)	0.097 (0.516)
意志	0.181 (0.224)	0.083 (0.581)
能力	0.094 (0.528)	0.051 (0.734)
対人	0.406 (0.005)	0.220 (0.137)
引用	0.060 (0.690)	-0.103 (0.491)

表 11: モダリティとブックマークされた数との相関

ラベル	First50_cnt	First50_per
断言	-0.010 (0.949)	-0.119 (0.427)
推測	0.125 (0.403)	-0.109 (0.466)
経験	0.138 (0.353)	0.084 (0.574)
伝聞	0.397 (0.006)	0.370 (0.011)
強制	-0.234 (0.114)	-0.351 (0.016)
誘導	0.004 (0.976)	-0.079 (0.597)
意志	0.079 (0.596)	0.086 (0.567)
能力	-0.127 (0.394)	-0.140 (0.349)
対人	0.226 (0.126)	0.114 (0.446)
引用	0.274 (0.062)	0.162 (0.277)

まとめ

- 背景：
  - ✓ソーシャルメディアでの偽誤情報の拡散は社会的脅威である
  - ✓生成AIの進化は止まらない．品質の安定化は時間の問題
- 本論：
  - ✓そもそも偽誤情報の拡散要因は何か？ ▶構造的要因と認知的要因
  - ✓真偽検証は自動化できるのか？ ▶将来的には可能
  - ✓真偽検証の自動化で解決するのか？ ▶認知バイアスなどで困難
  - ✓生成AIとどのように共生するか？ ▶対策への導入で対抗
  - ✓将来の対策検討は？ ▶対人心理に配慮した自動推定手法
    提案：モダリティは有用な可能性あり

参考文献

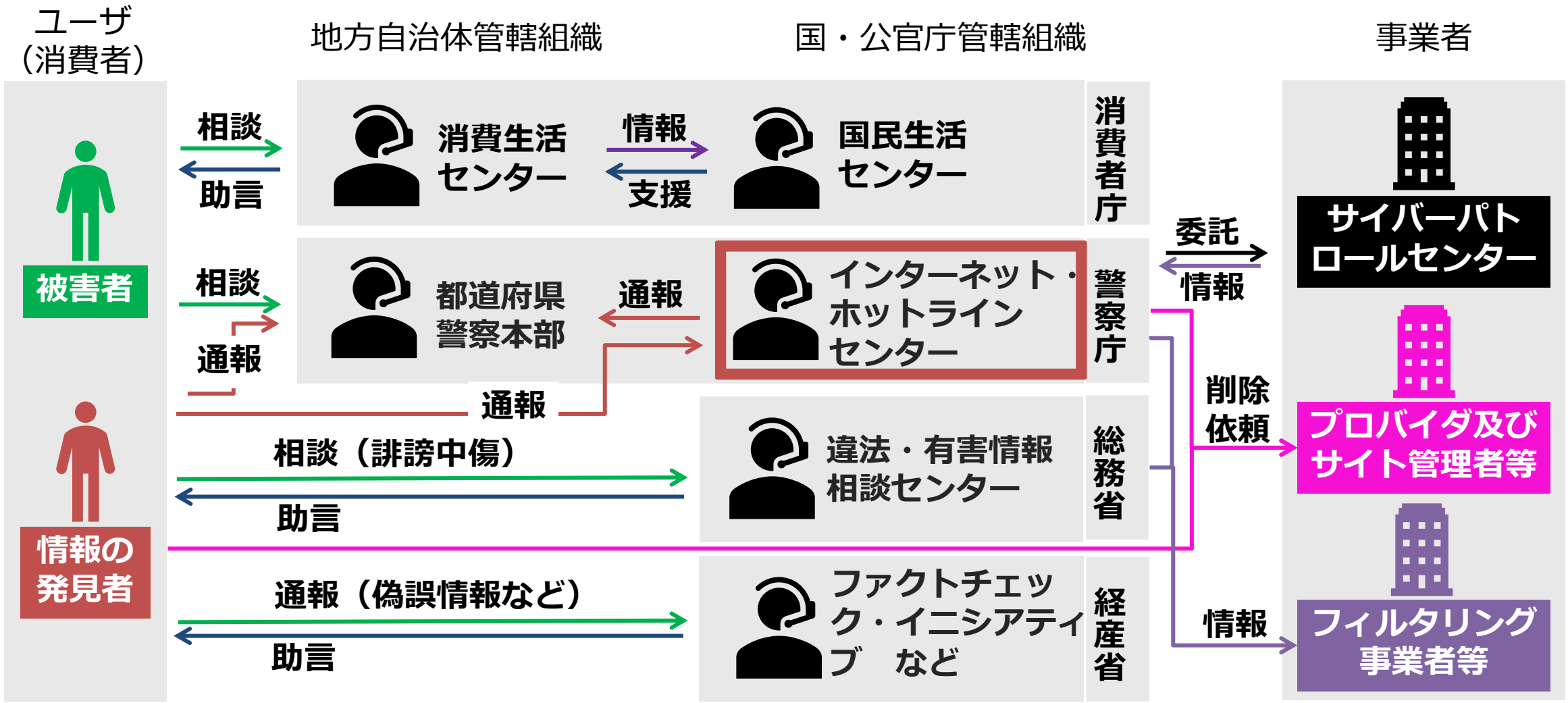
- P9　生成AIで顕在化したリスク
  - ✓ **総務省(2024)**　総務省，令和6年版 情報通信白書，生成AIが抱える課題，https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r06/html/nd141100.html
  - ✓ **CNN(2024)**　CNN，会計担当が３８億円を詐欺グループに送金、ビデオ会議のＣＦＯは偽物　香港，https://www.cnn.co.jp/world/35214839.html
  - ✓ **JFC(2022)**　日本ファクトチェックセンター，ドローンで撮影された静岡県の災害画像？ AIディープフェイクの見分け方【ファクトチェック】，https://www.factcheckcenter.jp/fact-check/disasters/shizuoka-disaster-drone-captured-ai-generated-fake-image/
  - ✓ **JITERA(2024)**　JITERA, https://jitera.com/ja/insights/44908
  - ✓ **日経ビジネス(2023)**　日経ビジネス電子版，AIが助長する差別・偏見　Appleにも襲いかかったリスクの本質，https://business.nikkei.com/atcl/gen/19/00548/050100003/
- P12　構造的要因：便利さに最適化されたアルゴリズム
  - ✓ **Sasahara(2021)** Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2021) Social Infuence and unfollowing accelerate the emergence of echo chambers. Journal of Computational Social Science, 4, 381–402.
  - ✓ **EchoDemo(2017)** Hao, P., Menczer, F., & Sasahara, K. (2017) EchoDemo. https://osome.iu.edu/demos/echo/
- P13　構造的要因：SNSの安全は誰の責任か？
  - ✓ **Meta (2019)** Mark Zuckerberg Stands for Voice and Free Expression, https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression/
  - ✓ **Twitter (2018)** Twitter Shouldn't be 'Arbiters of Truth,' Says CEO Jack Dorsey – TheWrap
- P18　認知的要因：人間は偽物を見破れるのか？
  - ✓ **Bray et al. (2023)**　Sergi D Bray, Shane D Johnson, Bennett Kleinberg, Testing human ability to detect ‘deepfake’ images of human faces, Journal of Cybersecurity, Volume 9, Issue 1, 2023, tyad011, https://doi.org/10.1093/cybsec/tyad011
  - ✓ **Lu et al. (2023)**　Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. 2023. Seeing is not always believing: benchmarking human and model perception of AI-generated images. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 1105, 25435–25447.
  - ✓ **Korshunov et al. (2020)**　P. Korshunov, S. Marcel, Deepfake detection: Humans vs. machines, arXiv preprint arXiv:2009.03155, 2020.
  - ✓ **Bond et al. (2006)**　Bond, C. F., & DePaulo, B. M, Accuracy of Deception Judgments. Personality and Social Psychology Review, 10(3), 214-234, 2006
  - ✓ Chen (2023) Canyu Chen, Kai Shu：Can LLM-Generated Misinformation Be Detected?, arXiv:2309.13788, 2023.
- P24ファクトチェックの課題：検証対象の偏り
  - ✓ [FIJ (2024)] 日本ファクトチェックイニシアティブ, https://drive.google.com/file/d/1miyGy31FlNsR1_Q7Xgmb9neJ6I2XSq3q/view
  - ✓ [2] Zhijiang Guo, Michael Schlichtkrull, Andreas Vlachos : A Survey on Automated Fact-Checking, Transactions of the Association for Computational Linguistics, Vol.10, p.178-206 (2022).
  - ✓ [3] Lim, C. (2018). Checking how fact-checkers check. Research & Politics, 5(3). https://doi.org/10.1177/2053168018786848

補足資料：用語定義の曖昧さ（フェイクニュースを例に）

提唱者（発行年）	真実性	誤解を与える可能性	意図の有無	政治性	報道機関の装い
Allcott and Gentzkow (2017)	○(完全な虚偽)	○	○	―	―
DiFranzo and Gloria-Garcia (2017)	○（偽）	○	―	―	―
Horne and Adalı (2017)	―	○	○	―	―
McNair (2017)	○（偽）	―	○	○	―
Mustafaraj and Metaxas (2017)	○	○	―	―	○
Pennycook and Rand (2017)	○	―	○	―	○
Lazer et al. (2017)	―	―	×(誤情報)	―	○
Lazer et al. (2018)	―	―	○	―	○
Nelson and Taneja (2018)	○	○	―	―	○
Guess et al. (2018)	―	―	×(誤情報)	○	―
Bakir and Mcstay (2018)	○(完全な虚偽)	○	○	―	―
Tandoc et al. (2018)	○(事実性の低さ)	―	○	―	○
Axel Gelfert(2018)	○	○	○	―	―

※事実性の検証には，**事実検証ができること，社会的な影響があるまたは予想されること**などの基準が加わる

悪質な投稿全般に関わる課題
＝研究用の正解データを共有し合う組織間連携が不十分



補足資料：事実はどう規定されるべきか？

- 真偽は二値では決まらない．多くは**混在状態**.
- ファクトチェッカーが用いる真実を判断する基準

正確	事実の誤りはなく、重要な要素が欠けていない。
ほぼ正確	一部は不正確だが、主要な部分・根幹に誤りはない。
ミスリード	一見事実と異なることは言っていないが、釣り見出しや重要な事実の欠落などにより、誤解の余地が大きい。
不正確	正確な部分と不正確な部分が混じっていて、全体として正確性が欠如している。
根拠不明	誤りと証明できないが、証拠・根拠がないか非常に乏しい。
誤り	全て、もしくは根幹部分に事実の誤りがある。
虚偽	全て、もしくは根幹部分に事実の誤りがあり、事実でないと知りながら伝えた疑いが濃厚である。
判定留保	真偽を証明することが困難。誤りの可能性が強くはないが、否定もできない。
検証対象外	意見や主観的な認識・評価に関することであり、真偽を証明・解明できる事柄ではない。

補足資料：ソーシャルメディアで確認される悪質な投稿

- ソーシャルメディアはユーザにとって身近なコミュニケーション手段である
- ソーシャルメディアでの悪質な投稿の拡散が社会的脅威になっている