

3 フェーズ2 災害発生時における現場対応・支援、避難活動

3 Phase 2 Reacting to Natural Disasters

3-1 偽・誤情報への対策技術によるレジリエンス構築

3-1 Building Resilience Through Technologies to Combat Disinformation and Misinformation

松田 美慧 鈴木 悠

MATSUDA Misato and SUZUKI Haruka

デジタル空間でのコミュニケーションは、社会を支える重要な情報インフラとして定着している。一方で、ソーシャルメディア上で拡散する偽・誤情報は、社会の分断や混乱を招き、深刻な社会問題となっている。生成AIの発展により偽・誤情報は更に高度化・大量化しており、対策には、情報や通信の信頼性を確保するデジタル空間のトラスト形成と、個人・社会が偽・誤情報に騙されず拡散しないためのレジリエンス構築の両立が重要である。本稿では、ソーシャルメディアユーザが偽・誤情報を信じる・拡散する要因を明らかにする研究と、それらの要因への対策および偽・誤情報の検出支援を通じたトラスト形成に関する研究開発成果を紹介する。

Digital communication has become an essential information infrastructure that supports modern society. However, the rapid spread of misinformation and disinformation on social media has emerged as a serious societal challenge, contributing to social polarization, public confusion, and manipulation of public opinion. Recent advances in generative artificial intelligence have further increased the scale and sophistication of false information, making effective countermeasures increasingly difficult. Addressing this challenge requires both the establishment of digital trust, which ensures the reliability of information and communication, and the enhancement of societal and individual resilience so that users can recognize, resist, and avoid spreading false information. This paper introduces our research on identifying the factors that influence why social media users believe and disseminate misinformation and disinformation. It also presents countermeasures targeting these factors, together with technologies that support the detection of false information to promote trustworthy digital information environments.

1 まえがき

ソーシャルメディアの普及により、デジタル空間は従来のマスメディアによる一方向の情報発信だけでなく、個人が双方向に情報を発信・閲覧できるようになった。デジタル空間でのコミュニケーションは、災害発生時や感染症流行時などの生命の危機に関わる状況下においても活発に行われ、社会生活における情報インフラとして機能している。一方で、ソーシャルメディアで流通する偽・誤情報^{*1}がデジタル空間の健全性を妨げていることが問題となっている。2016年のアメリカ合衆国大統領選挙で拡散された偽・誤情報は社会の分断・対立を助長し、世界情勢の不安

定化や経済安全保障のリスクを高めた。近年では、生成AIの急速な発展により高精度なディープフェイク動画・画像の生成が可能となり、AIが情報収集してプロファイリングを行い、パーソナライズした偽情報を作成・配信する一連の仕組みが構築されつつある。

偽・誤情報の問題に対処するためには、「デジタル空間のトラスト形成」と「社会・個人のレジリエンス構築」の両輪での対策を実施することが重要である。デジタ

*1 偽情報 (Disinformation) とは、他者を欺く意図をもって発信された虚偽または誤解を招く情報を指し、誤情報 (Misinformation) とは意図せずに発信された虚偽または誤解を招く情報である。本稿では両者を指す際に「偽・誤情報」と表現する。

ル空間におけるトラストは、通信をする相手や通信経路、やり取りされる情報の信憑性を保証することで実現する。しかし、生成AIの普及により偽・誤情報の流通量が増大し、ファクトチェックや訂正が追いつかない状況にある中で、真偽の判断は個人に委ねられている。人間の意思決定には、科学的・客観的な事実よりも、個人の感情や世界観などによる信念への訴えかけに強く影響を受ける特性がある。偽・誤情報はこの特性を悪用し、個人の感情や世界観に訴えるナラティブ（物語）を騙ることで、意図する世論や意思決定へ誘導する。このため、偽・誤情報が流通するデジタル空間においては、健全性を確保するためのトラストを形成するとともに、偽・誤情報が紛れ込んでくることを前提とした「社会や個人が騙されない・信じない・拡散しない免疫力（レジリエンス）」を持つことが重要である。

このような背景を踏まえ、サイバーセキュリティ研究室では人間中心社会を想定した「ヒューマン・センタード・セキュリティ技術」の一環として、偽・誤情報によるユーザへの悪影響を低減するための研究開発を実施している。社会・個人のレジリエンス構築に寄与することを目的に、ソーシャルメディアユーザが偽・誤情報を信じる／拡散する要因を解明し（2.1, 2.2）、それらの要因に対処するための有効策について研究している（3.1, 3.2）。また、デジタル空間のトラスト形成に寄与することを目的に、偽・誤情報の検出を支援するための研究開発を行っている（3.3）。サイバーセキュリティ研究室では、偽・誤情報は人間の認知的な脆弱性を狙うサイバー攻撃であるという認識のもと、社会を形成する個人の認知を守るためのコグニティブセキュリティの観点から偽・誤情報対策に貢献している。

2 偽・誤情報を信じる／拡散する要因の解明

偽・誤情報を信じる、あるいは拡散する要因に関する当研究室の研究を紹介する。先行研究では、偽・誤情報を閲覧したときにはユーザの感情的反応や直感的思考との関連が強まることが実証されており、これらは感情的な投稿コンテンツやプラットフォーム事業者が実装している推薦アルゴリズムによって助長される。また、偽・誤情報を信じやすいユーザの特徴には、性別や年齢といった人口属性とともに、リテラシの低さや情報源への信頼の高さなどが指摘されている。このうち、当研究室では感情的反応に焦点をあて、投稿コンテンツに含まれる感情表現がユーザの感情的反応を促し、対策効果に影響を与えることを明らかにした。

2.1 偽・誤情報を信じる要因

偽・誤情報の信じやすさに影響を及ぼす要因の検証：松田らは、偽・誤情報の信じやすさに影響する要因を明らかにするため、口調（冷静／感情的）を変えた偽・誤情報を異なるユーザ群に提示する実験を行い、先行研究より抽出した109個の要因を比較した。その結果、偽・誤情報の口調の違いが信じやすさに影響を及ぼす要因は、性別ごとに異なることが明らかになった。表1に、要因と信じやすさとの間の相関係数 r と、予測性能実験で強い関連が示された要因の一覧を示す。予測性能実験では、相関係数が高い15個の特徴量（TOP15）を用いた全状態探索を行い、独立したリッジ回帰の交差検証誤差が最小となるモデルで利用されていた特徴量を要因として抽出している。

冷静な口調の偽・誤情報では、男性群・女性群ともに、偽・誤情報の信じやすさと「肯定的なコメントへの嫌悪」との間に負の相関が確認された。この傾向は予測性能実験でも同様にみられた。よって、偽・誤情

表1 各要因と信じやすさとの間の相関係数 r とTOP15の全状態探索により強い関連が示された要因の一覧

要因	冷静な口調-女性群		冷静な口調-男性群		感情的な口調-女性群		感情的な口調-男性群	
	相関 r	TOP15	相関 r	TOP15	相関 r	TOP15	相関 r	TOP15
習慣：「雑誌・書籍」から健康情報を取得している	0.181	✓(+)	-0.079		0.022		0.056	
習慣：「動画投稿サイト」を利用する時間が長い	-0.155		-0.210		-0.284	✓(-)	-0.143	
能力：「メディアリテラシ」が高い	-0.187	✓(-)	-0.297		-0.195		-0.087	
信頼：「テレビへの信頼度」が高い	0.049		0.332	✓(+)	0.271		0.154	
信頼：「X への信頼度」が高い	0.070		0.242		0.349	✓(+)	0.178	
価値観：真偽検証を行う優先順位が高い話題は「芸能・文化」である	-0.207	✓(-)	0.162		-0.067		0.235	✓(+)
価値観：真偽検証を行う優先順位が高い話題は「スポーツ」である	-0.044		0.249	✓(+)	-0.121		0.153	
感情：「肯定的なコメントへの怒り」が強い	-0.198		-0.218	✓(+)	0.058		0.064	
感情：「肯定的なコメントへの嫌悪」が強い	-0.194	✓(-)	-0.386	✓(-)	0.072		-0.139	
感情：「否定的なコメントへの驚き」が強い	-0.057		-0.025		0.202		0.311	✓(+)
認知：「投稿内容」を根拠に真偽を判断した	0.224	✓(+)	0.009		0.318		0.181	
認知：「統計情報（いいねの数など）」を根拠に真偽を判断した	0.180		0.096		0.352	✓(+)	0.188	
認知：「自分の経験・知識」を根拠に真偽を判断した	0.239	✓(+)	-0.027		-0.045		0.098	

表中の✓は予測性能実験において信じやすさとの関連が示されたことを示す。(+)は交差検証誤差が最小となる予測モデルにおいて各要因の重み係数が正の値であったことを示す。(-)は負の値であったことを示す。

報に対する肯定的なコメントに嫌悪を抱いたユーザは、偽・誤情報を信じにくくなることが示された。加えて、女性群では「メディアリテラシ」との間にも負の相関が確認された。冷静な口調として提示した実験刺激は、フリーランスの記者から発信されたニュース記事のような投稿文であったことから、ユーザは情報源に対する批判的思考を働かせたことで信じにくくなったことが考えられる。一方、感情的な口調の偽・誤情報では、「Xへの信頼度」や「動画投稿サイトの利用習慣」などのメディアとの関わりに関する要因がより多く示された。ただし回答傾向は冷静な口調の偽・誤情報よりも複雑だったため、TOP15以外のより多くの特徴量による検証が有用である。

2.2 偽・誤情報を拡散する要因

偽・誤情報の感情表現がユーザの共有意思に及ぼす影響の検証：鈴木ら [1] は、偽・誤情報の感情表現に着目し、偽・誤情報に対するユーザ感情が信憑性判断（信じる／信じない）よりも偽・誤情報を共有する意思に影響を与えることを明らかにした。2016年のアメリカ合衆国大統領選挙で拡散した偽・誤情報は、社会不安と人間の認知バイアスを利用して極端な怒りと疑念を引き起こそうとしていた。先行研究では、人間は情報の真偽に関わらず、感情的な反応を引き起こす情報を伝える可能性があるという考えが示されている。そのため、偽・誤情報が表現する強い怒りがユーザの共有行動を促進しているという仮説を設定し、偽・誤情報を含むテキスト投稿刺激を作成して検証した。

実験結果を分析したところ、偽・誤情報刺激に対するユーザ感情の強さと信憑性判断の両要因とも偽・誤情報の共有意思に有意に影響を及ぼしていたが($p < .05$)、その傾向は刺激ごとに異なっていた。偽・誤情報刺激のうち、やや強い～強い怒りの回答が多かった場合は、感情の強さの方が信憑性判断よりも共有意思へ大きな影響を及ぼしていた。一方、やや強い～強い悲しみの回答が多かった偽・誤情報刺激の場合は、信憑性判断の方が感情の強さよりも共有意思への影響が大きかった。さらに、テキスト投稿刺激に対してユーザが認識した感情の種類による違いを分析したところ、信じる」と回答した割合が多かった感情は悲しみ、怒り、喜び、期待であり、共有すると回答した割合が多かった感情は喜びと怒りだった。したがって、偽・誤情報は強い怒りを表現することによって、ユーザが強い怒りから信憑性に関わらず共有する場合と、怒りから信じて共有する場合の2パターンで促進されている可能性が示された。

3 偽・誤情報対策の研究開発

2で実証された要因を踏まえ、当研究室が実施するレジリエンスの構築に向けた研究開発を紹介する。偽・誤情報への対策はユーザレベルとシステムレベルに大きく分けられる。ソーシャルメディアユーザに向けたユーザレベルの対策では、リテラシ教育や事実の事前共有のような予防的な対策と、訂正やナッジによる事後的な対策が行われる。一方、システムレベルの対策には、プラットフォームが実施するコンテンツの削除やアカウントの凍結措置、行政組織による法規制の整備などがある。このうち当研究室では、事実の共有、ナッジ、モダリティの検出技術への応用といった研究開発を実施している。

3.1 事実の共有による認識への影響の軽減

偽・誤情報対策における予防と訂正効果の比較検証：松田ら [2] は、事実情報をユーザに見せる適切なタイミングを明らかにするため、ユーザ実験を行い、偽・誤情報と事実情報を見せる順番を変えたユーザ群間で対策効果を比較した。情報を閲覧する順番が人間の判断に影響することは、社会心理学分野を中心に実証されており、最初に閲覧した情報がその後の判断に強く影響する「初頭効果」と、最後に閲覧した情報がより強く影響する「親近効果」が存在する。本研究では、偽・誤情報が拡散しているプラットフォームで対策者が事実情報を発信することを想定し、初頭効果を利用した

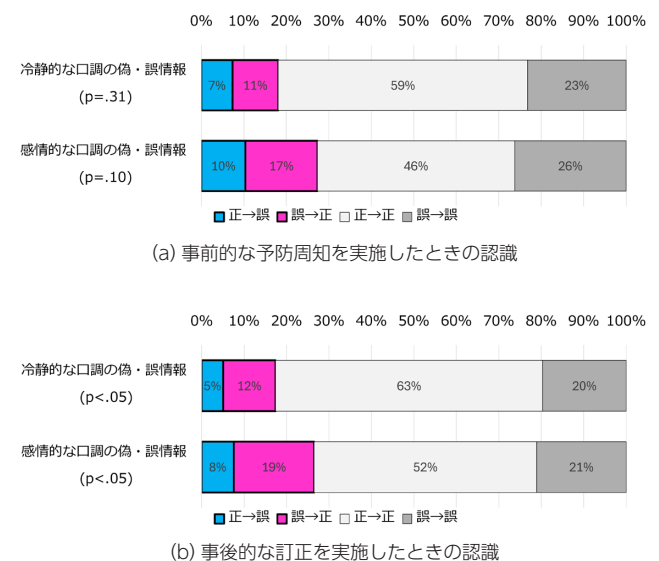


図1 偽・誤情報と事実情報を閲覧したユーザの偽・誤情報に対する認識の変化

「正」は事実関係が正しく認識されている場合を指し、「誤」はその逆を意味する。図中の数値は、正誤の割合を群ごとに百分率で表現したものである。 p とは有意差検定で算出された p 値を指し、0.05以下の場合に有意差があると判断する。

事前的な予防周知と、親近効果を利用した事後的な訂正の効果を実践的に検証した。

分析の結果、事実と偽・誤情報が混在する環境下で効果的な介入タイミングは、事前よりも事後であることが示された。図1に偽・誤情報と事実情報を閲覧した前後で事実関係への認識が変化あるいは継続した割合と、有意差検定で算出されたp値を示す。予防周知と訂正ともに認識が変化する割合は約20～30%程度で、事実関係を正しく認識する方向に修正される割合が認識を誤る割合を上回った。しかし予防周知では有意差が確認されず、対策効果は事後的な訂正のみにあることが明らかになった。また予防の実施時には偽・誤情報に対する悲しみや驚き、訂正の実施時には事実情報に対する恐れや嫌悪、驚きが微減した。

3.2 偽・誤情報の拡散抑制対策

偽・誤情報の共有時に熟慮を促す情動調節ナッジの提案：鈴木ら [3][4] は、偽・誤情報の強い怒りが促進するユーザの共有行動を減らすため、偽・誤情報の感情情報にユーザの注意を向けさせて熟慮を促す情動調節ナッジを提案した。情動調節とは、自発的に感情の強弱を調節する内発的プロセスのことである。人間は感情が強過ぎると衝動的になりやすく、感情のバランスをとることが思慮深い判断につながる。この情動調節は、現在の感情と望ましい状態との間に不一致があるときに生じる。このため、偽・誤情報をユーザが共有しようとしたときに、①偽・誤情報の感情情報(円グラフ)により感情の不一致に気付かせ、②熟慮を促す情動調節メッセージをポップアップ画面で表示するナッジを作成し(図2)、その有効性を既存の対策と比較評価した。

先行研究を基に情動調節メッセージを作成して検証した結果、客観的・分析的視点で見る「視点取得」と

フォロワーへの思いやりを想像する「気晴らし」のメッセージによって、偽・誤情報刺激に対するユーザの感情が弱まる傾向がみられた($p < .05$)。偽・誤情報の強い怒りによる共有を減らす効果を検証するため、情動調節ナッジ(視点取得/気晴らし)とソーシャルメディアに実装されているナッジを模した既存ナッジを比較評価したところ、情動調節ナッジ(気晴らし)が既存ナッジよりも効果が有意に高いことが示された($p < .05$)。また、情動調節ナッジ(気晴らし)と事前に既存の教材を読む教育を比較評価した結果、偽・誤情報刺激の共有を減らす効果に有意な差はなかった。ただし、情動調節ナッジは偽・誤情報の強い怒りに対するユーザ感情が弱まることで信憑性判断の影響が高まり共有が減ったのに対し、教育は教育効果により共有が減るものの、偽・誤情報が表現する強い怒りによって熟慮が妨げられた場合には教育効果が十分に発揮されず怒りのままに共有していた可能性がみられた。対策によって効果が発揮されるプロセスが異なることから、教育の補完策として情動調節ナッジを併用することが有用である。

3.3 偽・誤情報の検出

モダリティに着目した偽・誤情報の検出手法の提案：松田ら [5] は、多様な話題への汎化性が高い検出技術の開発への応用を見据え、偽・誤情報の投稿文に含まれるモダリティに着目し、偽・誤情報の特性とエンゲージ数がモダリティとどのように関係するかを検証した。モダリティとは、主に副詞や助動詞で構成される話者の心理表現を指し、従来の検出実験ではクリーニング処理により除去される場合もある。ただしモダリティは名詞や動詞で構成される命題への依存度が低いことから、汎用性の高い特徴量となる可能性がある。本研究では、Xで流通した偽・誤情報に含まれるモダリティの頻度や割合を算出し、相関分析や有意差検定による検証を実施した。

検証の結果、モダリティの頻度と割合は、偽・誤情報の特性(ファクトチェックの結果や発信意図など)とリアクション数との間でそれぞれ相関があることが示された。推量的な表現(例：～かもしれない、～はずだ etc.)の構成頻度はミスリードな投稿と事実関係が誤った投稿との間で有意な差があり、ファクトチェックにおける真偽ラベルの自動付与を支援する可能性があることが示された。また他者からの伝聞を示す表現(例：～らしい、ようだ etc.)の出現頻度は、引用数とブックマーク数との間で正の相関(各 $r=0.34$ 、 0.40)があることが示され、投稿の拡散を助長する可能性が示された。これにより偽・誤情報の検出にモダリティを応用することの有益性が示された。

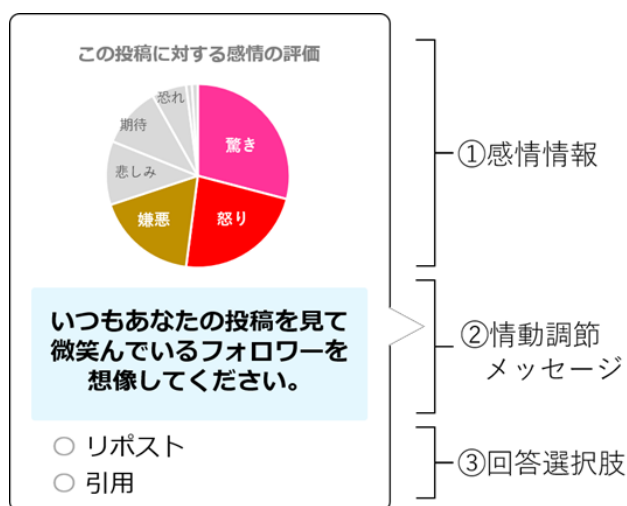


図2 情動調節ナッジ(気晴らし)のデザイン

4 おわりに

本稿では、サイバーセキュリティ研究室におけるデジタル空間のレジリエンス構築とトラスト形成に向けた研究開発の成果を一部紹介した。現代社会において、デジタル空間への依存度は高く、デジタル空間で流通する情報が人間の認知へ与える影響は大きい。偽・誤情報が認知へ与える影響とその要因を検証した実験では、人間は入力情報への態度を決定するにあたり、自身の経験や価値観に基づく判断をする場合もあれば、怒りや嫌悪などの感情的な反応が優位に働く場合もあることが示された。対策の検討では、ユーザへの介入策の効果を高める方法として、事後対策的な事実共有と事中共策に該当する情動調節ナッジの重要性を確かめた。またファクトチェックなどの対策の起点となる検出技術の開発に向け、モダリティを検出へ応用する有益性を示した。サイバーセキュリティ研究室は、偽・誤情報による社会と個人への悪影響を低減するため、デジタル空間における情報システムを対象としたサイバーセキュリティとともに、人間の心理特性を考慮したコグニティブセキュリティの観点からの研究開発に今後も取り組む。

【参考文献】

- 1 Haruka Nakajima Suzuki and Midori Inaba, "Psychological Study on Judgment and Sharing of Online Disinformation," Proceedings of the 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), pp.1558–1563, 2023.
- 2 松田 美慧, 鈴木 悠, 藤田 彬, 吉岡 克成, 笠間 貴弘, "予防と訂正: 人間の現実認識を歪める攻撃への対策効果の検証," 情報処理学会論文誌, vol.67, no.9, pp.1–13, 2026.
- 3 Haruka Nakajima Suzuki and Midori Inaba, "Digital Nudges Using Emotion Regulation to Reduce Online Disinformation Sharing," Journal of Information Processing, vol.33, pp.733–743, 2025.
- 4 Haruka Nakajima Suzuki and Midori Inaba, "Nudges to Reduce the Spread of Online Disinformation: A Comparison with the Educational Effect," Proceedings of the 20th International Workshop on Security (IWSEC 2025), vol.16208, pp.513–532, 2026.
- 5 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を検討," 情報処理学会論文誌, vol.66, no.3, pp.1–16, 2025.



鈴木 悠 (すずき はるか)

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
専門研究員
博士 (情報学)
ディスインフォメーション、コグニティブ
セキュリティ

【受賞歴】

2025年 CSS2025奨励賞
2022年 2021年度ISSスクエア シンポジウム
ベストポスター賞
2022年 2021年度ISSスクエア シンポジウム
ISSスクエア賞



松田 美慧 (まつだ みさと)

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
博士 (情報学)
ソーシャルメディア、ユーザブルセキュリティ

【受賞歴】

2025年 CSS2025奨励賞
2023年 CSS2023奨励賞
2023年 2022年度ISSスクエア シンポジウム
ベストポスター賞
2023年 2022年度ISSスクエア シンポジウム
ISSスクエア賞