予防と訂正：人はどうすれば事実を正しく捉えられるのか

松田 美慧[1,2,a)]　鈴木 悠[1]　藤田 彬[1]　吉岡 克成[1,3,4]　笠間 貴弘[1]

概要：偽誤情報の拡散は，ユーザが事実と反する信念を抱くことを助長し，長期的な影響を及ぼす．ユーザに向けた対策には，検証済みの事実を偽誤情報の拡散前に周知するプレバンキングと拡散後に周知するデバンキングがある．しかし，対策の効果はソーシャルメディアの推薦アルゴリズムなどの影響により制限される懸念がある．本研究では，ウェブアンケート（N=723）を実施し，事実に関する情報を見るタイミングと内容を操作し，対策の効果が制限される状況下における対策の効果検証を行った．結果，デバンキングの実施後にユーザの認識が統計的有意差をもって向上し，特に事実情報が偽誤情報に含まれる情報を明確に否定している場合に，効果が高かった．また両対策は負の感情を抑えることを明らかにした．

キーワード：偽誤情報，ユーザ調査，デバンキング，プレバンキング

Pre-bunking and Debunking: How Can People Get the Facts Right?

Misato Matsuda[1,2,a)]　Haruka Suzuki[1]　Akira Fujita[1]　Katsunari Yoshioka[1,3,4]
Takahiro Kasama[1]

Abstract: The spread of false information can have long-term effects by encouraging users to hold beliefs that contradict the facts. User-focused interventions include pre-bunking, which provides factual information before falsehoods spread, and debunking, which provides it afterward. However, there are concerns that the effectiveness of countermeasures may be limited by factors such as social media recommendation algorithms. In this study, we conducted a web survey (N=723) and manipulated the timing and content of factual information viewed to verify the effectiveness of countermeasures under conditions where their effectiveness is limited. The results showed that there was a statistically significant change in users' perceptions after debunking, and the effectiveness of the countermeasures was particularly high when the factual information clearly denied the information contained in the false information. We also found that both countermeasures suppress negative emotions.

Keywords: False information, User surveys, Debunking, Prebunking

1. はじめに

偽誤情報は，真偽検証済みの事実と反する信念の形成を助長し，長期的な悪影響と訂正の困難さを生む．偽誤情報による信念への影響を低減するために，信念の適正化を目的としたユーザに向けた対策の「デバンキング」と「プレバンキング」の研究が行われてきた．デバンキングとは，既に広まった偽誤情報に対して，真偽検証で事実と判定された情報（以下，「事実情報」）を提供する事後対策である．一方，プレバンキングとは，ユーザが偽誤情報に接触する前に偽誤情報への耐性を構築する事前対策である．具体的には，リテラシー教育の実施や，流布される可能性のある偽誤情報の予測と関連する事実情報を用いた警告などがある．

デバンキングとプレバンキングがユーザの信念の適正化

[1] 国立研究開発法人情報通信研究機構 National Institute of Information and Communications Technology
[2] 横浜国立大学大学院環境情報学府 Graduate School of Environment and Information Sciences, Yokohama National University
[3] 横浜国立大学大学院先端科学高等研究院 Institute of Advanced Sciences, Yokohama National University
[4] 横浜国立大学大学院環境情報研究院 Faculty of Environment and Information Sciences, Yokohama National University
[a)] m.matsuda@nict.go.jp

に有用であることは複数の研究で示されている．しかし，いずれも実験環境で検証された効果であり，実際のソーシャルメディア環境の動的要素が考慮されていない [1]．動的要素には，投稿に対するリプライ，フォロワーの影響，推奨アルゴリズムなどがある．実際にユーザが目にするタイムラインは，これらの影響により事実情報と偽誤情報が混在した状況となる．このような状況下では，3 つの理由から両対策の効果が制限される可能性がある．

第一に，**ユーザが偽誤情報と事実情報を見る順番の制御は不可能**である．ユーザのタイムラインに情報を表示するタイミングは，推薦アルゴリズムの影響下にあり，ソーシャルメディアで事実情報を発信する対策者はそれを制御できない．そのため，仮に対策者が事後的な訂正を目的とする発信をした場合も，ユーザのタイムラインに偽誤情報よりも先に事実情報が表示されれば，事実情報は予防的役割を担うことになり，その逆も考えうる．よって両対策の効果の測定では，単一の事実情報に対し，ユーザがそれを見る順番による影響を検証する必要がある．

第二に，**多様な偽誤情報に対して単一な事実情報では部分的な対応**になりやすい．偽誤情報は，拡散する過程でユーザの加筆・削除により内容が変容していき，分化あるいは事実情報が混在する場合が生じる．デバンキングは対策リソースが限られるため，分化した各偽誤情報に対応した事実情報を発信することは困難である．プレバンキングでは，偽誤情報に事実情報が混在している場合，その一部の事実情報が偽誤情報への正確さの認識を高め，対策の効果が発揮されない可能性がある．このような場合を想定し，偽誤情報と事実情報の内容が部分的に対応する状況での効果を検証する必要がある．

第三に，**偽誤情報は強い感情を喚起**する．強い感情は，偽誤情報を誤って信じる可能性を高める [2]．デバンキングは，偽誤情報により生じた怒りを緩和する [3]．プレバンキングでは，偽誤情報による感情的な操作の手口を事前に周知し，偽誤情報への抵抗力を高める [4]．一方，実環境ではその効果が弱まる可能性が示されている [5]．よって，実環境に近い状況下で，両対策が強い感情の抑制に効果があるかを検証する必要がある．

そこで本研究では，**RQ1[閲覧順]: 偽誤情報と事実情報を見る順番がユーザの認識に影響を及ぼすか？，RQ2[事柄の対応]: 偽誤情報と事実情報に記述された個別的な事柄の対応がユーザの認識に影響を及ぼすか？，RQ3[感情]: 情報から喚起された感情が対策効果に影響を及ぼすか？** を検証する．実験では，医療・健康に関する偽誤情報または事実情報の事例を元に，X（旧 Twitter）の投稿，発信者のプロフィール，コメントの画面を模した画像を提示し，提示する順番を入れ替えて反応を測定した．偽誤情報-事実情報の順で提示する場合をデバンキング，事実情報-偽誤情報の順で提示する場合をプレバンキングと想定する．その結果，提示順が情報の正確さに対する認識に影響を与えることが示された．本論文の貢献を以下に示す．

- 事実情報と偽誤情報の閲覧順は，情報に対する正確さの認識に影響を及ぼし，特にデバンキングで対策の効果が大きいことを明らかにした．
- 事実情報は偽誤情報の内容との対応がなくとも対策効果が確認された．一方，事実情報にない偽誤情報は不正確な認識を広げる可能性を示し，事実情報の逐次発信の重要性を明らかにした．
- いずれの閲覧順においても負の一次感情（恐れ，嫌悪，悲しみ）は軽減するが，閲覧順によって軽減する感情が異なることを示した．

2. 関連研究

2.1 デバンキング

デバンキングとは，偽誤情報による誤った信念が形成された後に，その誤解に対処する事後対策である．思考に影響を与え続ける偽誤情報の影響を減らすためには，誤りの指摘よりもデバンキングが有効であると考えられている [6]．デバンキングでは，誤りであることの説明と共に，事実に関する情報の提供（トピック反論）や，誤解を招くために使用された修辞的な戦術の暴露（テクニック反論）が用いられる [7]．しかし，ユーザがデバンキングを自身の信念や意見への反論と受け取った場合には，逆にその信念が強化されるという「バックファイヤ効果」が確認されている [8]．加えてユーザの信念と事実が異なる場合には，ユーザが訂正する情報を受け取ることを選択的に回避し，訂正が進まない [9]．本研究では，トピック反論の手法を用いて，検証を行う．

2.2 プレバンキング

プレバンキングとは，ユーザが偽誤情報を見抜き，偽誤情報に対する認知的抵抗力を高めるための事前対策である．プレバンキングは，「予防接種理論」に基づき，(1) 信念や態度に対する攻撃の警告，(2) 偽誤情報拡散の手口（トロール，扇情的な表現など）に対する先制的な反論という 2 つの主要な要素で構成される [7]．プレバンキングでは，世論や人間の心理を操作するために拡散者が用いる操作テクニックを説明することでユーザに見抜く力を与えるという教育的側面があることから，メッセージに批判などが含まれず共感を得やすいとされている [10]．またゲーム，動画，テキストなど多様な介入媒体で即時効果が確認されており，政策的には若年層向け教育や広報キャンペーンへの応用が期待されている．一方，信頼できる情報源や受け手の属性による効果差にも注意が必要である．自己効力感や反論意欲といった動機づけへの効果は必ずしも伴わず，長期的な持続性（記憶減衰）が課題となる [11]．またリアルなソーシャルメディア環境下では，その対策効果が弱まる

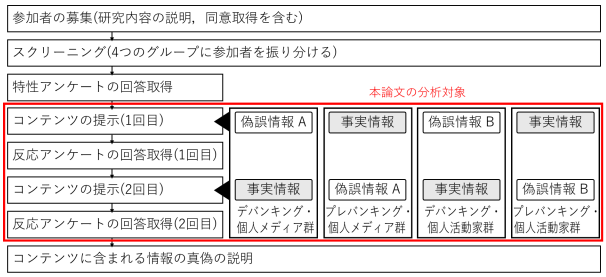


図 1　本実験の流れと分析対象範囲

可能性が示されている [5]．本研究では，トピック反論と対極的な対策を行うため，プレバンキングの意味を事実情報の先行的な発信と限定し，検証を行う．

3.　実験手法

3.1　実験手法の概要

事実情報と偽誤情報が混在した状況下で，情報を受け取る順番が与える影響を検証する．ウェブアンケートにおいてコンテンツを提示する順番をコントロールした状態で，過去に X で確認された事例をもとに作成した画像を参加者に提示し，参加者の心理状態や属性，リテラシースキルに関する情報を取得する．回答完了時間は 30 分程度と想定する．質問は，(1) 個人的な特性を問う質問（以下「特性アンケート」），(2) 表示されたコンテンツに対する反応を問う質問（以下「反応アンケート」）の 2 つから構成される．本論文ではコンテンツに対する反応に注目するため，特性アンケートへの回答は分析の対象外とする．ユーザ調査の流れと本論文での分析対象範囲を図 1 に示す．

3.2　参加者の募集とスクリーニング

2025 年 2 月 10 日〜12 日の 2 日間に渡り，調査会社である Cross Marketing を介し，参加者を募集した．対象は，20 代から 60 代までの日本在住者である．参加者の募集時には，5.2 節で示す方法で研究内容の説明ならびに同意取得を行い，その後，性別ならびに年代が均等になるようにスクリーニングを実施した．最終的に回答を取得した人数は 800 名で，このうち有効回答をした人数は 723 名であった．有効回答の判定には四分位偏差法を用い，回答時間が極端に長いあるいは短い参加者の回答を外れ値として除外した．有効回答を行った 723 名について，回答者の性別ならびに年代の分布に偏りがないことを確認した．723 名の参加者の属性情報を表 1 に示す．

表 1　群ごとの有効回答数

	デバンキング・個人メディア群		デバンキング・個人活動家群		プレバンキング・個人メディア群		プレバンキング・個人活動家群		
	男性	女性	男性	女性	男性	女性	男性	女性	小計
20-29 歳	15	20	20	19	19	16	18	19	146
30-39 歳	17	19	20	19	19	16	19	20	149
40-49 歳	19	15	19	17	20	17	17	18	142
50-59 歳	19	17	17	18	19	19	18	20	147
60-69 歳	19	18	19	17	17	15	19	15	139
小計	89	89	95	90	94	83	91	92	723

3.3　コンテンツの提示

3.3.1　提示するコンテンツの作成方法

3.3.1.1　偽誤情報ならびに事実情報の事例の抽出

リアルなソーシャルメディア環境に近づけるため，本実験で使用するコンテンツには，過去にソーシャルメディアで拡散した事例を用いる．偽誤情報の事例は，日本のファクトチェック支援団体である認定 NPO 法人ファクトチェック・イニシアティブ（FIJ）が運用する「ファクトチェック・ナビ」[12] より取得した．当該サイトには，2019 年 6 月より 983 件のファクトチェックの結果が掲載されている．真偽判断の根拠には，各ファクトチェック機関による真偽検証の結果を用いる．FIJ の真偽のレーティング基準は 7 段階あるが，本研究では「ミスリード/不正確/誤り/虚偽」の 4 段階に該当する情報を偽誤情報と定義する．

本研究では，実験を実施する時期（2025 年 2 月）から半年前までに X で拡散された偽誤情報を当該サイトから抽出した．期間の制限は，ユーザの反応が話題の時勢に左右されやすいため，時勢に即した話題を抽出する目的で設けた．模擬するソーシャルメディア環境を X とした理由は，X のニュース目的での利用率が YouTube，Instagram について国内で 3 番目に高く，全目的に対するニュース目的での利用率が約 53.13%と最も高いためである [13]．また対象とするニュースカテゴリは，Innnovation Nippon2024 の報告書 [14] にて，ファクトチェックの優先度が高いとされる「災害」「医療健康」を中心とする．このほか，特定の個人や団体に対する不利益が生じる危険のある話題を排除し，X で関連する事実情報やリプライが十分に確保できることを確認した上で，8 つの候補を抽出した．

3.3.1.2　コンテンツで取り扱う話題の選定

選定では，3 名の研究者が独立して各候補を評価し，5 つの評価軸ごとに評価の高いものが高い配点となるように 0〜4 点を付与し，平均点が高い話題を選んだ．評価軸ごとの評価が同程度の場合は同点とする．評価軸は，「公共性/有害性/公平性/不確実性/感情性」の 5 つで，詳細を表 2 に示す．これらは，事前調査において，ファクトチェック機関が検証対象を選定する際に重視する基準として抽出した観点である．事前調査では，国際ファクトチェックネットワーク [15] に加盟しているファクトチェッカー 191 団体のうち，ランダムサンプリングされた 40 団体について，各団体の公式サイトで検証対象を選定する手法を説明する記述を収集し，再帰的テーマティック分析に基づくコーディングを行った．これらの選定過程を経て，実験で取り扱う話題は，医療・健康カテゴリの「鳥インフルエンザは鳥から人間に感染しない」とした．

3.3.1.3　X の画面を模した画像の作成

実験刺激は，投稿内容，発信者のプロフィール情報，コメントの 3 種類ある．投稿内容には，2 種類の偽誤情報と 1 種類の事実情報を用意する．ソーシャルメディアでのユー

表 2　話題の選定に用いた評価軸

評価軸	説明
公共性	一般的に多くの人からの関心の集めやすさ
有害性	情報が広まることによる社会への悪影響の大きさ
公平性	偏見的もしくは一方的な主張の程度
不確実性	一般的に多くの人が事実判断に迷うような疑わしさ・事実らしさ
感情性	一般的に主張を読んだ人の感情を掻き立てる程度

ザの反応は，発信者の信頼性や属性などに影響を受けることが知られている．そこで偽誤情報の 1 つは個人メディア，他方は個人活動家からの発信を想定し，それぞれ「偽誤情報 A」と「偽誤情報 B」と呼称する．個人メディアとは，無所属の記者を語る人物がメディアとして運用するアカウントである．個人活動家とは，特定の分野に対する強い主張をもった人物が運用するアカウントである．偽誤情報の発信者を個人とした理由は，真偽検証の対象のほとんどは一般ユーザの発信であり，偽誤情報の発信源となりやすいためである．これに対し，事実情報の発信者は，情報の信頼性が高いとされる大手メディアを想定する．

各刺激は，3.3.1.2 項に示した方法で選定された話題を取り扱った投稿とその発信者の情報のうち，個人の特定につながる情報や URL を架空の名称に置換し，リアクション数やフォロワー数，フォロー数を架空の数値にしたものである．数値は，ユーザが情報の信頼性を判断する根拠となる場合がある．そのため，偽誤情報 AB 間での結果の違いを生まないように偽誤情報 AB 間では同数となるように数値を設定した．

投稿内容: 投稿内容とは，タイムラインに表示される投稿自体を指す．画像の上部には発信したアカウントのプロフィール画像，表示名，ID が記載され，中央部には 140 文字以内のテキスト，下部には投稿日時，閲覧数，コメント数，いいね数，引用数，ブックマークされた数が記載される．真偽検証では，取り扱う話題に関連する 4 つの事柄について真偽が明らかになっている．投稿内容には，4 つの事柄のいずれかが複数記載され，事柄によっては記載がない場合や真偽が異なる場合を含む．コンテンツごとの記載の有無と，情報の真偽を表 3 に纏める．実験で使用した画像を図 2 に例示する．図 2 内の赤線ならびに赤字は，事柄 1〜4 との対応箇所を表し，実験時には非表示とする．

発信者のプロフィール情報: 発信者のプロフィール情報とは，X においてアカウントの特徴を説明するために表示される情報を指す．画像の上部には発信したアカウントのプロフィール画像，表示名と ID，中央部にはアカウントの発信の目的などを説明するテキスト，下部にはアカウントを作成した日付，フォロー数，フォロワー数が記載される．ユーザが情報の真偽対する認識を構築する際には，発信者の過去の発言を参考にする場合もあるが，本実験ではそのような情報は取り扱わない．実験で使用したプロフィール情報を図 3 に例示する．

投稿内容に対するコメント: 投稿内容への第三者からの返信やリツイートを指し，1 つの投稿内容あたり 10 個用意した．このうちの 5 個は，投稿内容を支持する「肯定コメント」であり，他方は「否定コメント」である．各コメントには，アカウント名にアカウントの保有者の年代や嗜好について記載し，アカウントの属性に偏りがないように工夫した．またコメントの内容は，虚偽の情報を含む場合もあり，真偽の入り混じった環境を実現するため，ノイズとしての役割を担う．コメントを図 4 に例示する．

なお，発信者の属性と内容の多様性を確保するため，一部のコメントは LLM を利用して作成した．利用したモデルは OpenAI 社の「GPT-4o」である．プロンプトではシミュレーションを行う研究目的であることを明示し，「鳥インフルエンザは人に感染しない」という偽誤情報に対し，肯定または否定的なコメントを 140 文字以内で 20 件ずつ出力するように指示した．最終的なコメントは，実験上の投稿者の属性と発言内容の自然さ，内容の重複のなさを人間の作業者が考慮し，採用の可否を判断した．

3.3.2　コンテンツの掲示方法

実験では，X のタイムライン上で一般ユーザが情報を偶然目にする状態を想定する．1 人の参加者は 1 つの話題について真偽の相反する主張を交互に閲覧する．偽誤情報 A →事実情報の順でコンテンツを閲覧したグループを**デバンキング・個人メディア群**，偽誤情報 B →事実情報の順を**デバンキング・個人活動家群**，事実情報→偽誤情報 A の順を**プレバンキング・個人メディア群**，事実情報→偽誤情報 B の順を**プレバンキング・個人活動家群**と呼称する．コンテンツの閲覧時には，投稿内容，プロフィール情報，コメントの順で画像が表示される．参加者は任意のタイミングで，各情報を繰り返し閲覧できる．

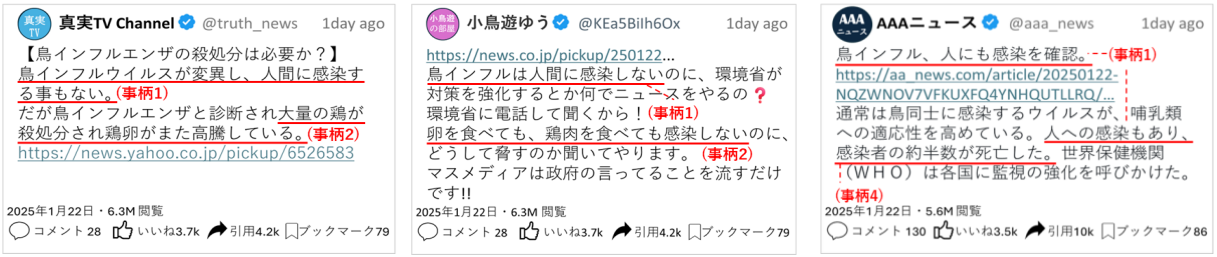
3.4　反応アンケートの回答取得

反応アンケートでは，直前に閲覧したコンテンツに対する**思考的反応，行動的反応，感情的反応**の 3 種類の反応について尋ねる．思考的反応では，4 つの事柄に対する考えとその考えを抱いた根拠への依存度を尋ねた．行動的反応では，投稿内容ならびにコメントを他者に共有する意思の有無とその共有方法を，7 つの名義尺度にて取得した．感情的反応では，プルチックの感情の輪で示される 8 つの一次感情ごとに，喚起された感情の程度を 0〜4 の数値にて取得した．各設問の形式はマトリクス設問で，1 つの設問文につき，複数の事柄への反応に関する回答を取得している．

反応アンケートは 1 つ目と 2 つ目のコンテンツを閲覧した後に行うが，その設問文は 2 回とも同じである．1 つ目のコンテンツが表示される直前には，「いくつかの情報や SNS へ投稿された意見」が提示されると説明し，2 つ目の直前には，「先ほどと異なる情報や SNS へ投稿された意見」が提示されると説明した．なお，コンテンツを閲覧する際

表 3　各事柄の投稿内容への記載の有無と真偽の一覧

事柄	内容	真偽	偽誤情報 A（個人メディア）	偽誤情報 B（個人活動家）	事実情報
事柄 1	鳥インフルエンザは鳥から人間に一切感染しない	偽	記載有（肯定：偽）	記載有（肯定：偽）	記載有（否定：真）
事柄 2	卵や鶏肉を食べても鳥インフルエンザウイルスに感染しない	真	言及有（感染に関する情報なし）	記載有（肯定：真）	記載なし
事柄 3	鳥インフルエンザウイルスがヒトからヒトに感染することはない	真	記載なし	記載なし	記載なし
事柄 4	人に感染すると，重症化することもあり，致死率は 50 %以上である	真	記載なし	記載なし	記載有（肯定：真）



(a) 偽誤情報 A（個人メディア）の投稿内容　(b) 偽誤情報 B（個人活動家）の投稿内容　(c) 事実情報の投稿内容

図 2　投稿内容

には，その真偽に対する説明は行わない．

3.5　コンテンツに含まれる情報の真偽の説明

全ての回答を終えた参加者に対し，実験中に掲示したコンテンツに含まれた情報の真偽を説明する文章を提示し，内容を理解したかについて尋ねた．真偽を説明する文章は，ファクトチェック機関が行った真偽検証の結果に基づいて，実験設計者が作成した．文中では，厚生労働省など公的機関からの公開情報を引用し，事柄 1〜4 に対する真偽を紹介した．文章の内容を理解したと回答した参加者は，全回答者の 94.5%であった．

4.　実験結果

4.1　ユーザの正確さへの認識に対する影響

RQ1[閲覧順]，**RQ2[事柄の対応]** を明らかにするため，2 つ目のコンテンツを閲覧する前後で，ユーザの考えが変化したかを検証する．検証には，思考的反応を取得する質問のうち，4 つの事柄に対する考えを尋ねる質問への回答を用いる．考えを尋ねる設問文は，「投稿やコメント，発信者の情報を見て，あなたはどのように考えましたか.」であり，回答は「1 確実に間違っている」から「6 確実に正しい」の 6 段階のリッカート尺度である．

4.1.1　思考的反応が変化する割合

1 つ目と 2 つ目のコンテンツを閲覧した同一の参加者の回答をクロス集計した結果を表 4 に示す．2 つ目のコンテンツを閲覧する以外はほぼ同条件下で回答したにも関わらず，表 4 より約 20〜30%の参加者の考えが変化したことが確認された．このうち，「誤→正」に変化した割合が「正→誤」よりも大きくなった事柄は，16 個中 6 個であり，個人メディア群の方がその個数が多くなった．

4.1.2　個人内の思考的反応の比較

より詳細な分析として，2 つのコンテンツを閲覧した前後の考えに，統計的に有意な差があるかを検証する．検証

表 4　思考的反応の正誤の割合 (%)

		変化なし		変化あり		増減率
		正→正	誤→誤	(ア) 正→誤	(イ) 誤→正	(イ)-(ア)
デバンキング・個人メディア群	事柄 1	62.92	19.66	5.06	12.36	+7.30*
	事柄 2	35.39	41.57	13.48	9.55	-3.93
	事柄 3	22.47	56.74	14.04	6.74	-7.30*
	事柄 4	29.21	51.12	6.74	12.92	+6.18
デバンキング・個人活動家群	事柄 1	52.43	21.08	7.57	18.92	+11.35*
	事柄 2	35.68	40.00	16.22	8.11	-8.11*
	事柄 3	23.24	53.51	15.68	7.57	-8.11*
	事柄 4	29.19	48.65	11.89	10.27	-1.62
プレバンキング・個人メディア群	事柄 1	58.76	23.16	7.34	10.73	+3.39
	事柄 2	48.02	37.29	9.04	5.65	-3.39*
	事柄 3	30.51	50.28	12.43	6.78	-5.65*
	事柄 4	24.29	58.76	6.78	10.17	+3.39
プレバンキング・個人活動家群	事柄 1	46.45	26.23	10.38	16.94	+6.56
	事柄 2	38.25	38.25	12.02	11.48	-0.54
	事柄 3	31.69	48.63	10.38	9.29	-1.09
	事柄 4	21.86	57.38	12.02	8.74	-3.28

「正」は真偽検証で事実と判定された結果と思考的反応への回答の真偽が一致する場合を指し，「誤」はその逆を意味する．回答の選択肢は 6 段階あるが，「1 確実に間違っている」から「3 どちらかというと間違っている」までと，「4 どちらかというと正しい」から「6 確実に正しい」までの 3 段階ずつを 1 つのカテゴリとして，正誤を判断する．*表 5 で有意差が確認されたことを示す．

には，対応のある 2 群間の比較をする手法のうち，ノンパラメトリックなデータの分析をするための一般的な手法である Wilcoxon 符号付順位和検定を採用する．検定の結果を表 5 に示す．表 5 内の灰色でマーキングされた箇所は，p 値が 0.05 以下となり，有意差が確認されたことを表す．表 5 より，デバンキング・個人メディア群とプレバンキング・個人メディア群でそれぞれ 2 つずつ，デバンキング・個人活動家群で 3 つ有意差が確認され，プレバンキング・個人活動家群では有意差が確認されなかった．これらは，事柄 1〜3 に偏って存在し，偽誤情報に記述がある場合と事実情報に記述がない場合に有意差が確認された．

4.1.3　群間の思考的反応の比較

2 つ目のコンテンツを閲覧した後の 4 つの事柄への思考的反応を群間で比較した．これは，対応のない 3 群以上の

(a) 偽誤情報 A のプロフィール情報　(b) 偽誤情報 B のプロフィール情報　(c) 事実情報のプロフィール情報

図 3　発信者のプロフィール情報

(a) 偽誤情報 AB への肯定　(b) 偽誤情報 AB への否定　(c) 事実情報への肯定　(d) 事実情報への否定

図 4　コメント

表 5　介入前後での有意差検定の結果

	事柄 1		事柄 2		事柄 3		事柄 4	
	統計量	p 値	統計量	p 値	統計量	p 値	統計量	p 値
デバンキング・個人メディア群	2241.5	0.0044	2220	0.1708	1929	0.0427	896.5	0.0521
デバンキング・個人活動家群	2718.5	0.0009	2068	0.0127	1940.5	0.0208	1563	0.9312
プレバンキング・個人メディア群	1824	0.3071	1384.5	0.0162	1740.5	0.025	1117.5	0.8879
プレバンキング・個人活動家群	2645	0.1037	2321	0.1719	1871.5	0.2011	1878	0.0769

表 6　事柄 4 に対する介入後の思考的反応の多重比較の結果

比較群	統計量	p 値
デバンキング・個人メディア群 vs プレバンキング・個人メディア群	1.946	0.209
デバンキング・個人メディア群 vs デバンキング・個人活動家群	0.146	0.999
デバンキング・個人メディア群 vs プレバンキング・個人活動家群	2.899	0.020
プレバンキング・個人メディア群 vs デバンキング・個人活動家群	1.875	0.239
プレバンキング・個人メディア群 vs プレバンキング・個人活動家群	0.933	0.787
デバンキング・個人活動家群 vs プレバンキング・個人活動家群	2.854	0.022

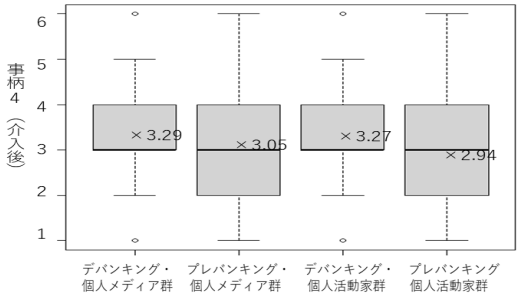


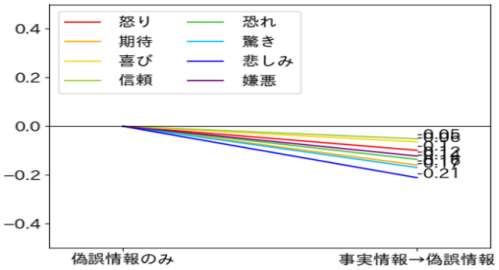
図 5　事柄 4 に対する介入後の思考的反応の箱ひげ図

間での比較であるため，Kruskal-Wallis 検定を実施し，有意差が確認できた場合は Steel Dwass の多重比較により有意差のある群のペアを特定する．Kruskal-Wallis 検定の結果，事柄 4 に対して $p = 0.006(<0.05)$ となり，有意差が確認された．事柄 4 に対する多重比較の結果を表 6 に示し，箱ひげ図を図 5 に示す．図 5 の太線は中央値，×は平均値を示す．表 6 より，デバンキングを実施した 2 つの群がプレバンキング・個人活動家群との間に有意差を持っていることが確認された．図 5 より，プレバンキング・個人活動家群は他方に比べ，第一四分位数が低く，平均値も低いことが分かる．よって，デバンキング実施時の方が介入後に誤った考えを持ちにくくなっていた．

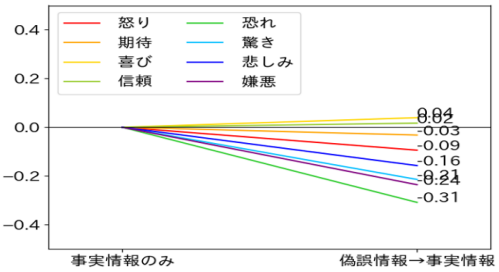
4.2　ユーザの感情に対する影響

RQ3[感情] を明らかにするため，投稿内容に対する感情的反応を比較する．感情的反応を問う設問文は，「投稿の内容に対する，いまのあなたの感情の度合いをお答えください．」である．回答は選択式で，感情が喚起されていない場合は「0 なし」，何らかの感情が喚起されている場合は「1 弱い」～「4 強い」の 5 段階のリッカート尺度である．

偽誤情報に対する感情の推移を図 6(a) に示し，事実情報に対する感情の推移を図 6(b) に示す．図 6(a) より，偽誤情報のみを見た場合に比べ，事実情報を見た後に偽誤情報を見た場合の方が，投稿の内容に対する一次感情が低くなる傾向にあることが分かった．特に悲しみの感情が軽減される傾向にあり，これは偽誤情報 A を閲覧した場合により強く表れた．図 6(b) より，事実情報についても投稿に対す



(a) 偽誤情報に対する感情



(b) 事実情報に対する感情

図 6 投稿内容に対する感情的反応の平均値の推移

(a) デバンキングを実施した 2 つの群が 1 つ目のコンテンツを閲覧した直後の回答と，プレバンキングを実施した 2 つの群が 2 つ目のコンテンツを閲覧した直後の回答の平均値の差分によって表す．(b) プレバンキングを実施した 2 つの群が 1 つ目のコンテンツを閲覧した直後の回答と，デバンキングを実施した 2 つの群が 2 つ目のコンテンツを閲覧した直後の回答の平均値の差分によって表す．

る一次感情が低くなる傾向が見られた．事実情報に対しては，偽誤情報を見た後の方が恐れや嫌悪，驚きといった一次感情が緩和される傾向が確認された．

5. 議論

5.1 有用な介入的対策

RQ1 に対する閲覧順が正確さの認識に与える影響の検証では，個人内と群間の両方において，**デバンキングの方がプレバンキングよりも対策としての効果が高い**ことが示された．個人内においての正確さの認識の比較では，デバンキングは事実情報と一致する認識を抱く割合を向上させた（表 4）のに対し，プレバンキングは事実情報と異なる認識を抱く割合が高めるという逆効果が確認された．群間においての正確さの認識の比較においても，デバンキングの方がプレバンキングよりも効果が高かったこと（図 5）から，事実情報を偽誤情報の後に見た方が事実関係を認識できると考察される．ただし，デバンキングの効果が高くなった要因として，本実験が短期的な効果の取得に留まっていることに加え，プレバンキングが十分に機能しなかった可能性が考えられる．プレバンキングはユーザが事実情報を事実として認識し，偽誤情報を閲覧した際に偽誤情報への警戒心を高めることで機能する．しかし，本実験では事実情報を先に閲覧した参加者が，その内容を事実とは認識せず，警戒心が高まらない状態で偽誤情報を見た可能性が考えられる．

RQ2 に対する事柄の対応による影響の検証では，事実情報と偽誤情報の内容に**対応が取れていることよりも事実情報が発信されていることの方が対策の効果がある**ことが示された．個人内においての正確さの認識の比較では，事実情報と一致する認識を抱く割合を向上させたのは事柄 1 のみであった（表 4）．事柄 1 は，事実情報と偽誤情報の両方に記載があるため，両者に対応が取れる状態である．一方で，事柄 2〜4 は事実情報と偽誤情報が非対応だった．事柄 2 や事柄 3 では，偽誤情報に対応する事実情報がなかったことから，ユーザの正確さの認識は変化しないと考えられたが，実際には逆効果がみられた（表 4）．これは，対応する事実情報がないことから，偽誤情報に基づいて事実と異なる信念の形成が助長されたと考察される．事柄 4 の事実情報のみに記載がある場合は，個人メディアから発信された偽誤情報を受けても，事実情報を信じる割合が高まったことから，偽誤情報と事実情報の内容が非対応である場合も事実情報を発信することには効果があると考察される．よって，事実情報の発信時には検証済みの情報から逐次発信し，事実情報が流通していない時間をできるだけ短くすることが必要と結論付けられる．

RQ3 に対する感情への影響の検証では，**両対策において負の感情の喚起を抑える傾向**が見られ，対策の実施に有用な効果があると考察される．プレバンキングとデバンキングでは，抑制された感情の種類が異なり，プレバンキングでは偽誤情報に対する悲しみの感情，デバンキングでは事実情報に対する恐れや嫌悪の感情が軽減している．特に事実情報に対する強い感情は，ユーザの確固たる信念を形成し，訂正の受け入れを困難にする．そのため，デバンキングにおける事実情報に対する嫌悪や恐れの軽減は，ユーザが事実情報を受け入れるために機能すると考察される．

5.2 倫理的配慮

実験の計画段階に，実験や分析の手法について所属組織のパーソナルデータ取扱研究開発業務審議委員会にて審査を仰ぎ，「低リスク（注意あり）」との判定を受け，データの取得，保管，偽誤情報の訂正に関する対策を実施した．データの取得では，参加者を募集する際に，研究の目的や人権擁護上の配慮，受ける可能性のある負担及び不利益とそれを最小化するための対策などについて参加者へ文面にて提示し，提示した情報への理解と自由意志による参加への同意を得た．募集時のタイトルは，「真偽が不確かな情報への反応プロセスの解明とその対策に関する研究」とし，偽誤情報への関心が高い参加者が偏在しないように，偽誤情報やフェイクニュースといった用語は用いなかった．

データの保管に関しては，分析を行う際などの使用時以外は研究所の特定のメンバーだけがアクセスできるファイル・サーバに暗号化して保存する．加えて，性格特性や信

条などの要配慮個人情報を含む機微な情報の取り扱うことから，分析および分析結果の公表において，特定の個人が「偽誤情報に騙されやすい」あるいは「偽誤情報を拡散しやすい」という評価を行うなど，参加者個人が不利益を受けるような分析及び個人の取り扱いを行わないこととする．

5.3 制約と課題

本研究で行った実験は，ソーシャルメディアで拡散された実際の投稿や真偽の入り混じったコメントの導入により，複雑な実環境により近づけている．しかし，画像や動画といったマルチモーダルな投稿形式や他言語への対応を未実施である．特にユーザは画像や動画を根拠に情報への認識を構築する場合があることから，マルチモーダルな投稿形式における効果の検証が課題である．また実際のソーシャルメディアで，一般ユーザのタイムラインに表示される情報は，推薦アルゴリズムの働きによりユーザの興味や趣向に最適化されており，ユーザは自分の興味関心に合わせ，能動的な行動を伴う場合がある．ユーザは類似する情報を繰り返し閲覧することで強い信念を形成するが，検索などのユーザの活動は観測していない．加えて，本実験の介入によって見られた対策の効果は短期的な効果である可能性がある．よってユーザ活動の測定と，長期的な観測は今後の課題である．

6. おわりに

事実情報と偽誤情報が混在した状況下におけるデバンキングとプレバンキングの効果を検証し，正確さの認識に対してはデバンキングの効果が高く，感情に対しては両対策共に効果があることを明らかにした．本研究では，X で実際に拡散した偽誤情報，プロフィール情報，コメントの 3 種類の情報を提示するウェブアンケートを実施し，両対策の効果を検証した．その結果，偽誤情報と事実情報に記載された内容が直接的に対応づいている場合には，介入前後の正確さの認識に有意な差が確認され，デバンキングの効果があることが示された．一方，偽誤情報に対する事実情報がない場合には誤解が広がり，事実情報だけを発信した場合には一定の効果がみられたことから，検証済みの事実は積極的に発信することが有用である．本研究で得られた知見は，プレバンキングとデバンキングが対策としての効果の違いを示すものであり，両対策を併用することで互いの課題を補い合うことが望ましい．今後の課題には，画像や動画を含むマルチモーダルな投稿による検証，長期的な影響の観測，他言語対応などがある．

参考文献

[1] Walter, N. and Tukachinsky, R.: A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It?, *Communication Research*, Vol. 47, No. 2, pp. 155–177 (2020).

[2] Featherstone, J. D. and Zhang, J.: Feeling angry: the effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude, *Journal of Health Communication*, Vol. 25, No. 9, pp. 692–702 (online), DOI: 10.1080/10810730.2020.1838671 (2020). PMID: 33103600.

[3] Nakajima Suzuki, H. and Inaba, M.: Psychological Study on Judgment and Sharing of Online Disinformation, *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1558–1563 (online), DOI: 10.1109/COMPSAC57700.2023.00240 (2023).

[4] Basol, M., Roozenbeek, J. and van der Linden, S.: Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News, *Journal of Cognition*, Vol. 3, No. 1, pp. 1–9 (2020).

[5] Wang, S. Y. N., Phillips, S. C., Carley, K. M., Lin, H. and Pennycook, G.: Limited effectiveness of psychological inoculation against misinformation in a social media feed, *PNAS NExus*, Vol. 4, No. 6, p. pgaf172 (2025).

[6] Lewandowsky, S. and et al.: The Debunking Handbook 2020 (2020). available from `https://sks.to/db2020`.

[7] Roozenbeek, J., Culloty, E. and Suiter, J.: Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions, *European Psychologist*, Vol. 28, No. 3, p. 189–205 (2023).

[8] Nyhan, B. and Reifler, J.: When Corrections Fail: The Persistence of Political Misperceptions, *Polit Behaver*, Vol. 32, p. 303–330 (2010).

[9] Tanaka, Y. and et al.: Who Does Not Benefit from Fact-checking Websites? A Psychological Characteristic Predicts the Selective Avoidance of Clicking Uncongenial Facts, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023).

[10] Harjani, T. and et al.: A Practical Guide to Prebunking Misinformation (2022).

[11] Maertens, R. and et al.: Psychological booster shots targeting memory increase long-term resistance against misinformation, *Nature Communications*, Vol. 16, No. 2062, pp. 1–17 (2025).

[12] ファクトチェック・イニシアティブ: FactCheck Navi, ファクトチェック・イニシアティブ (online), available from ⟨`https://navi.fij.info/`⟩ (accessed 2025-08-21).

[13] 総務省：情報通信白書令和 7 年版データ集，総務省（オンライン），入手先 ⟨`https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r07/html/datashu.html`⟩（参照 2025-08-21）.

[14] Innovation Nippon 2024: 偽・誤情報、ファクトチェック、教育啓発に関する調査研究報告書，国際大学グローバル・コミュニケーション・センター（オンライン），入手先 ⟨`https://www.glocom.ac.jp/wp-content/uploads/2024/04/IN2024_report_fakenews_full.pdf`⟩（参照 2025-08-21）.

[15] The International Fact-Checking Network(IFCN): Signatories of the IFCN Code of Principles, , available from ⟨`https://ifcncodeofprinciples.poynter.org/signatories`⟩ (accessed 2025-08-21).