社会選好を考慮したSNS上の誹謗中傷投稿の拡散を抑制する介入戦略

林　央祐†　山本　修平††

† 筑波大学 情報学群 知識情報・図書館学類　〒 305–8550 茨城県つくば市春日 1 丁目 2 番地
†† 筑波大学 図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1 丁目 2 番地
E-mail: †s2210263@u.tsukuba.ac.jp, ††syamamoto@slis.tsukuba.ac.jp

あらまし　SNS 上の誹謗中傷拡散が深刻な社会問題となる中，法や技術的手法による対策が数多く行われているが，SNS 上の誹謗中傷の本質的な原因はその同調や拡散であるため，同調や拡散を抑制するアプローチも重要である．本研究は，行動経済学に基づき，SNS 上での誹謗中傷投稿への拡散を抑制するため，社会選好に応じて個別化された介入戦略の有効性を検討した．SNS 利用者に，4 種類の介入メッセージを提示する実験を行い，行動意向の変化を分析した．その結果，特定の社会選好タイプと介入メッセージの組み合わせによる誹謗中傷の有意な抑制効果はあまり確認されなかったが，特定の条件下では通常投稿の拡散まで抑制してしまう副作用が観察された．また，利得局面よりも損失局面の方が，介入メッセージの個別化に適している可能性が示されたため．社会選好は個別化の有力な手段となり得るが，メッセージとの関連性を考慮すべきであり，今後より効果的な分析をおこなうためには社会選好以外の個人の属性や傾向を取り入れる必要がある．

キーワード　誹謗中傷，SNS，ナッジ，行動変容，社会選好（Social Value Orientation）

1 はじめに

SNS での誹謗中傷：近年，SNS（ソーシャル・ネットワーキング・サービス）の普及により，時間や場所を問わず誰もがコミュニケーションを取ることが可能となった．一方で，SNS の持つ匿名性や拡散性を悪用した誹謗中傷が増加し，深刻な社会問題となっている．誹謗中傷に関する相談件数は年々増加しており，2024 年度には違法・有害情報相談センターへの相談件数が 6,403 件に達した [1]．また，インターネット利用者の 60.6% が誹謗中傷を目撃した経験があると報告されており [2]，問題の深刻さが窺える．誹謗中傷の被害は一般ユーザにも及ぶが，特に著名人に対するものは，メディアでの注目度が高く，被害者への心理的影響も大きい [3]．このような被害を防ぐための対策が急務となっている．炎上において実際に書き込みに参加する人は，インターネット利用者のうちごく一部にすぎない．むしろ，「安易に再投稿・拡散する人」の増加により誹謗中傷が集積し集団攻撃となることが示されている [4]．したがって，効果的な対策には，主体的な投稿者だけでなく，誹謗中傷に同調する人を減らすことが重要である．

効果的な誹謗中傷対策の難しさ：現在，誹謗中傷対策として法的手法と技術的手法が用いられている．法的手法では，プロバイダ責任制限法の改正により発信者情報開示請求の手続きが簡略化され，加害者特定が迅速化した [5]．技術的手法では，自然言語処理や機械学習による誹謗中傷の自動検出システムの開発が進められている [6]．しかし，法的手法では被害者の負担が依然として大きく，匿名性の高いプラットフォームでの加害者特定が困難になることもある．また，自動検出では文脈依存の表現や間接的な攻撃の判別精度が不十分であり，検出できなかったり，誤検出により正当な批判まで削除したりするリスクが指摘されている．このように，法律や技術のみに依存した対策には限界がある．

ナッジ：本研究ではこの問題に対し，行動経済学で注目されているナッジ理論に着目する．ナッジ理論は，個人の自由を損なわずに望ましい行動を促進する手法であり，意思決定に影響を与える環境や情報提示を工夫することで行動を自然に誘導する．ナッジは法律や罰則と異なり強制力を伴わないため，心理的抵抗が少ない点が特徴である．近年，オンラインハラスメントの文脈でもナッジの有効性が実証されつつあり，正木ら [7] の研究で検証されている．誹謗中傷の抑制においても，損失回避を意識させるメッセージ（例：「あなたの投稿が法的責任を問われる可能性があります」）や利他性に基づくメッセージ（例：「この投稿は相手を深く傷つけます」）の有効性が示唆されている [8]．このように介入メッセージにより誹謗中傷への同調行動を抑制できる可能性がある一方で，その効果は受け手の個人特性によって大きく異なることが指摘されている [9]．

社会選好：こうした個人差は，利他性や社会規範への感受性といった心理的特性に関連していると考えられる．先行研究でも，このような個人差が攻撃性や同調行為に影響することが明らかにされている．Hilbig ら [10] は，社会的価値志向性（Social Value Orientation: SVO）が攻撃性や反社会的行動と密接に関連することを実証的に示し，SNS での誹謗中傷やその同調への影響も例外ではないと考えられる．こうした特性は，行動経済学において社会選好として概念化されている．社会選好とは，自己の物質的利益だけでなく他者や集団の利益，あるいは分配の公平性を考慮する選好を指し，伝統的な経済学の自己利益最大化の仮定を超えて人間行動を説明する重要な概念である．社

会選好の測定には，SVO スライダー（Social Value Orientation Slider）が広く用いられている [11]．この手法は，自己と他者への資源配分課題を通じて社会選好を定量的に評価し，利他主義，協調主義，個人主義，競争主義などにタイプ分けすることができる．利他主義は他者の利益の最大化を，協調主義は自己と他者の利益の合計の最大化を，個人主義は自己の利益のみを，競争主義は他者との利益差の最大化を志向する．このように社会選好には多様な傾向が存在し，それぞれが攻撃性や社会規範への反応において異なる特徴を持つ．したがって，誹謗中傷への同調を抑制する介入メッセージの効果も，個々人の社会選好の傾向を考慮することでより高まると考えられる．

本研究：本研究は，SNS 上での誹謗中傷への同調や拡散を抑制するため，ナッジ理論を活用し，社会選好に応じて個別化された介入戦略の有効性を検討する．具体的には，実験参加者へのメッセージ提示実験を通じて，利己性と利他性，社会規範を強調するメッセージの効果を評価する．社会規範を提示するメッセージについては，多数派への協調を促すものと少数派からの離脱を促すものに区別し，比較分析を行う．実験参加者の社会選好を測定するため，資源配分課題を提示し，自己と他者への配分パターンから個々の社会選好のタイプを特定する．その上で，社会選好の傾向ごとに各介入メッセージを提示し，誹謗中傷への同調をどの程度抑制できるかを評価する．加えて，先行研究 [8] で問題とされた，介入策が誹謗中傷以外の投稿に対しても同調を抑制する副作用をどの程度低減できるかについても評価する．以上により，異なるメッセージが各傾向の受け手に与える影響を明らかにし，SNS 上での誹謗中傷を効果的に抑制するアプローチを構築する．

2 関 連 研 究

本研究は，社会選好に応じた介入メッセージを提示することで，誹謗中傷への同調を効果的に抑制できるかを明らかにすることを目的としている．この章では，法的・技術的手法による SNS 上の誹謗中傷対策に関する研究，ナッジによる介入メッセージの効果に関する研究，社会選好の測定方法とその活用に関する研究を概観する．最後に，本研究の位置付けについて述べる．

2.1 法的・技術的手法での SNS 上の誹謗中傷対策に関する研究

SNS 上での誹謗中傷対策として，法的手法と技術的手法が発展してきている． 法的手法では，プロバイダ責任制限法の改正により発信者情報開示請求の手続きが簡略化され，加害者特定の迅速化が実現されている [5]．従来，被害者は裁判所を通じてプロバイダと投稿者の両方に情報開示請求を行う必要があり，複数回の訴訟手続きを要していた．2021 年の法改正により一度の非訟手続きで発信者情報の開示が可能となり，手続きの簡略化と迅速化が達成された．また，侮辱罪の法定刑引き上げにより誹謗中傷に対する抑止力が強化され，法的枠組みの整備が進展している．これらの法的整備により，被害者が加害者に対して責任を追及する環境が改善されつつある．技術的手法では，自然言語処理や機械学習を用いた誹謗中傷の自動検出システムの開発が進展している．西谷ら [6] は，BERT 等の深層学習モデルを用いて日本語の誹謗中傷表現を高精度に検出する手法を提案し，実用的な精度を達成した．この手法では，大規模なテキストデータセットを学習することで文脈を考慮した誹謗中傷の判別が可能となっている．Zhang ら [12] は，畳み込みニューラルネットワーク（CNN）と注意機構を組み合わせたモデルを提案し，X（旧 Twitter）[1]上の攻撃的な言語を高精度で検出することに成功した．Davidson ら [13] は，ヘイトスピーチと単なる攻撃的な言語を区別するための分類器を開発し，両者の違いを機械学習で識別する手法を確立した．これらの技術により，大量の投稿から有害コンテンツを迅速に識別することが可能になり，プラットフォーム運営者による効率的なモデレーションが実現されつつある．さらに，深層学習の進展により，文脈依存の表現や新たなスラング表現への対応も継続的に改善が進められている．

2.2 ナッジによる介入メッセージの効果に関する研究

ナッジ理論は，個人の自由を損なわずに望ましい行動を促進する手法として，オンライン上の行動変容にも応用されている．Thaler ら [14] が提唱したナッジ理論の枠組みでは，損失回避や社会規範の活用が行動変容に効果的であるとされている．この理論は，法律や罰則と異なり強制力を伴わないため，心理的抵抗が少ない点が特徴である．正木ら [7] は，SNS 上での攻撃的コメント投稿を抑制するために，投稿前に注意喚起メッセージを表示するナッジ介入を実施した．その結果，注意喚起メッセージを受け取ったユーザは，攻撃的コメントの投稿率が有意に低下することが示された．また，Munger ら [15] は，Twitter 上で人種差別的な発言をしたユーザに対し，影響力のあるアカウントから警告メッセージを送信する実験を行った．その結果，警告を受けたユーザは，その後の差別的発言を有意に減少させることが実証された．この研究は，社会的影響力を持つ存在からのメッセージが行動変容に効果的であることを示している．誹謗中傷への同調抑制に関して，村田ら [8] は，利己性メッセージ（例：「あなたの投稿が法的責任を問われる可能性があります」）や利他性に基づくメッセージ（例：「この投稿は相手を深く傷つけます」），また社会規範メッセージ（例：「たった 1.5% の人しか誹謗中傷に加担していません」）の３つのメッセージの有効性を検証した．実験の結果，利己性・利他性のメッセージが誹謗中傷への同調を一定程度抑制する効果があることが示されたが，社会規範メッセージには効果がなかったことが示された．また，介入メッセージが誹謗中傷以外の投稿に対しても同調を抑制してしまう副作用も確認された．これらの研究により，ナッジを用いた介入メッセージは誹謗中傷自体の抑制や，誹謗中傷への同調に対して一定の効果を持つものの，その効果を最大化し，副作用を最小化するためには，副作用の原因の特定や効果を高める更なる工夫を見出すことが求められる．

1：X,https://x.com/?lang=ja

2.3 社会選好の測定方法とその活用に関する研究

社会選好は，自己の物質的利益だけでなく他者や集団の利益，あるいは分配の公平性を考慮する選好として，行動経済学において重要な概念として確立されている．社会選好の測定には複数の手法が開発されてきたが，中でも SVO スライダー（Social Value Orientation Slider）が広く用いられている．Murphy ら [11] は，SVO スライダーを開発し，自己と他者への資源配分課題を通じて個人の社会選好を定量的に評価する手法を確立した．SVO スライダーは，簡便かつ高い信頼性を持つ測定手法として多くの研究で採用されている．SVO スライダーを用いた研究では，社会選好が様々な社会的行動の予測に有効であることが示されている．Murphy ら [11] の研究では，SVO スライダーで測定された社会選好が協力行動や向社会的行動を予測することが実証された．特に，協調的志向性を持つ個人は，資源分配ゲームにおいてより公平な分配を選択し，他者との協力を重視する傾向が確認された．さらに，Pletzer ら [16] は，SVO スライダーを用いたメタ分析により，社会選好が社会的ジレンマにおける協力行動の強力な予測因子であることを明らかにした．分析の結果，協調的な社会選好を持つ個人は，個人主義的または競争的な社会選好を持つ個人と比較して，有意に高い協力行動を示すことが確認された．一方で，社会選好は協力行動だけでなく，攻撃性や反社会的行動とも関連することが示されている．Hilbig ら [10] は，SVO スライダーで測定した個人主義的・競争的な社会選好を持つ個人が攻撃的・反社会的な行動傾向を示しやすいことを実証した．彼らの研究では，社会選好と攻撃性の間に有意な関連が見られ，特に個人主義的・競争主義的志向性を持つ個人は，他者を犠牲にしてでも相対的優位を得ようとする行動パターンを示すことが明らかになった．これらの研究により，SVO スライダーで測定される社会選好は協力行動から攻撃性まで幅広い社会的行動を予測する重要な個人特性であることが確認されている．

2.4 本研究の位置付け

誹謗中傷投稿そのものを技術的に防ぐ試みは，今後の技術の進歩に伴いより効果的なものになるだろう．一方で，SNS 上の誹謗中傷被害は，誹謗中傷投稿が多くの人間に拡散され，被害者の目に届いてしまうことが原因であり，この同調等による拡散行動を抑制することも効果的である．また，ナッジ理論を用いた介入メッセージがオンライン上の誹謗中傷，それらの同調に対しても有効であることが様々な研究で示されている [8] [15]，一方で，従来のナッジは一般的に効果があるメッセージを全員に一律で提示するものであり，個人の特性に応じてメッセージを出し分けることで効果を高める研究は十分に行われていない．先行研究から，他者との協力行動や社会的行動が社会選好に応じて異なることが示されており [10] [11] [16]，誹謗中傷への同調行動も社会選好に応じて異なる可能性が考えられる．本研究は，攻撃性や同調行動と関連する社会選好に着目し，個人の社会選好の傾向に応じた介入メッセージの効果を検証することで，より効果的な誹謗中傷対策の実現を目指す．

3 社会選好を考慮したメッセージ提示実験

3.1 実験の概要

本研究では，社会選好に応じて個別化された介入戦略が，SNS 上での誹謗中傷への同調や拡散をどの程度抑制できるか評価する．本研究では 2025 年 10 月 27 日，28 日，11 月 10 日の 3 日間にわたり質問紙調査を実施した．実験参加者には，SNS 投稿を閲覧し介入メッセージを見る前の反応と，介入メッセージを見た後の反応として適切な選択肢を回答してもらい，両者を前後比較することで介入効果を測定した．本章の構成は以下の通りである．3.2 節で実験参加者の募集方法と参加条件，及び 4 つの実験グループへの割り当てについて説明する．3.3 節では実験で提示した介入メッセージを述べる．3.4 節では実験の具体的な流れを説明し，3.5 節で収集したデータの概要をまとめる．本研究は，筑波大学図書館情報メディア系研究倫理委員会の承認を受けて実施された（承認番号：第 25―93 号）．

3.2 実験参加者とグループ分け

3.2.1 実験参加希望者の募集

本研究の実験参加者は，オンラインアンケートサイト「Freeasy」[2]を通じて募集した．Freeasy は，調査対象者のリクルーティングとアンケート配信を行うマーケティングリサーチプラットフォームである．募集に際して設定した条件は，(1) 18 歳以上，(2) 直近 1 か月以内に SNS（実験で使用する X）で「いいね」や「リポスト・引用ポスト」を利用した経験がある，(3) 実験で提示する X 上の投稿内容（野球・サッカー）に興味・関心がある，の 3 つである．

条件 (2)，(3) は，SNS やテーマへの関心が低い参加者が含まれることで適切な分析結果が得られないことを防ぐために設定した．特に条件 (3) でスポーツ，具体的には野球とサッカーをテーマに指定した理由は以下の 2 点である．第一に，Williams らの研究 [17] において，スポーツに関する誹謗中傷は他分野の誹謗中傷と比べて試合等のイベント直後に急速に拡散されやすいことが示されており，誹謗中傷の拡散抑制を目的とする本研究に適したテーマであると判断したためである．第二に，総務省の調査結果 [18] から，野球とサッカーは日本における観戦者割合が最も高い上位 2 つのスポーツであり，多くの実験参加者が関心を持つテーマであると考えたためである．

これらの条件を満たす参加者を絞り込むため，Freeasy 上でスクリーニング調査を事前に実施した．その結果，計 15,000 名の応募者の中から条件を満たす 400 名の参加希望者を確保した．

3.2.2 個人の社会選好の測定

実験参加者の個人の社会選好を測定するために，上記のスクリーニング条件を満たした 400 人に社会選好に関するアンケートを実施した．本実験で実際に参加希望者に回答してもらった社会選好測定法を表 1，2 に示す．本実験では SVO スライダー [11] を参考にした．スクリーニングで提示する 6 問と本実験中に提示する 6 問でそれぞれ異なるシナリオを提示する．前

2：Freeasy, https://freeasy-survey.com/

半では，実験参加者自身と見ず知らずの第三者を想定し，利得を得られる金額配分の選択肢の中から実験参加者が適切だと思うものを選択する．後半では，想定する人物は変更せず，前半の 6 問の選択肢すべてから一律 10,000 円を引いた損失を被る金額配分の選択肢の中から，適切だと思うものを選択する．本来の SVO スライダーでは，社会選好を測定する際には利得の金額配分のみを用いるが，本実験で使用する介入メッセージは損失回避に働きかけるものであるため，利得局面で考える SVO スライダーよりも損失の金額配分を用いた SVO スライダーを使用することで，より介入メッセージが効果的に作用する個人の傾向を測定できると考えたため，二種類の SVO スライダーで測定する．スクリーニング時点で利得局面に関する社会選好測定を行うのは，個人の社会選好の傾向を測定する目的と，実験参加者の社会選好に偏りがないことを確認する目的があるためである．

また，表から実験参加者が選択した金額配分から，個人の社会選好の傾向を測定できる SVO スコアを算出する．自己への配分の平均を A_s，他者への平均を A_o として，SVO スコアは，以下の式で計算される．

$$SVO_p = \arctan\left(\frac{A_o - 5000}{A_s - 5000}\right),\ SVO_n = \arctan\left(\frac{A_o + 5000}{A_s + 5000}\right)$$

また，SVO_p と SVO_n は，それぞれ利得局面，損失局面によるスコアになっている．このスコアは，$-16.26° \sim 61.39°$ の範囲で算出される．

3.3 実験グループの割り当て

本実験では，実験参加者を 4 つの介入群にランダムに割り当てた．各介入群には異なるタイプの介入メッセージを提示し，個人の社会選好の傾向と介入メッセージの組み合わせによって誹謗中傷の拡散行動がどのように変化するかを分析した．本研究では個人の社会選好に応じて効果的な誹謗中傷を抑制する効果を持つ介入メッセージがどのようなものか明らかにすることを目的としているため，実験参加者の社会選好に極端な偏りがないことを確認したうえで，提示する介入メッセージの種類ごとに社会選好を考慮せずランダムな 100 人に振り分けた．

3.4 メッセージ提示実験

3.4.1 実験参加者に提示した投稿内容

本実験では，介入群に関わらず，実験参加者全員に対して X 上で実際に投稿された野球・サッカーのテーマそれぞれの誹謗中傷を含む投稿と，誹謗中傷を含まない投稿の計 4 種類を提示した．誹謗中傷を含む投稿に対しては，介入メッセージによって拡散行動を抑制したい狙いがあるが，誹謗中傷を含まない投稿に対しては，介入メッセージによって拡散行動を抑制してしまう副作用の効果を検証するために用いた．図 1 に実験参加者に提示した各投稿を示す．野球についての投稿は 2025 年 8 月に，サッカーについての投稿は 2024 年 10 月にそれぞれ X 上で実際に投稿されたニュース記事に返信された誹謗中傷を含む投稿と含まない投稿を選択した．誹謗中傷の定義として，国際大学グローバル・コミュニケーション・センターが定めた 9 つの定義 [19] を用いた．中でも，経験率が高かった「侮辱・攻撃」要素を含みかついいねやリポストが，対応する誹謗中傷なし投稿とできるだけ同数程度になっていることに注意して設定した．選択した計 4 つの投稿を，SNS を使用する際実際に閲覧するであろう X の画面を模した形式で提示した．

3.4.2 実験参加者に提示した介入メッセージの内容

本節では，実験に使用した介入メッセージの作成方法について説明する．先行研究 [8] では，自分自身の損失を強調する利己性メッセージや，他人の損失を強調する利他性メッセージが，誹謗中傷の拡散行動を効果的に抑制できることが報告されている．一方で，先行研究では社会規範メッセージは有意な抑制効果が確認されていないことも報告されている．本研究で改めて介入メッセージを提示する理由は 2 つある．まず，利己性メッセージと利他性メッセージは，誹謗中傷拡散を効果的に抑制できる一方で，誹謗中傷がない通常の投稿の拡散まで抑制してしまうことが分かっている．この原因として，個人の特性を考えずにメッセージを提示したことにより効果に個人差が出たと考えられる．よって，利己性メッセージと利他性メッセージにあたる利己損失強調メッセージと利他損失強調メッセージは本研究でも採用する．次に，社会規範メッセージは有意な抑制効果がなかったことが分かっている．この原因として，先行研究でも課題として挙げられていたが，実験参加者に社会規範を意識させられなかったことが考えられる．つまり社会規範メッセージを効果のある介入メッセージにするためには，より社会規範を意識させる工夫が必要である．よって，本研究では先行研究の社会規範メッセージを参考にしつつ，より多様な人に社会規範を意識させるために 2 つのメッセージを作成した．自身が多数派に属することを強く意識させる多数派帰属強調メッセージと，自身が少数派でありそこから脱却することを強く意識させる少数派脱却メッセージに細分化した．以上の知見を踏まえて，本研究では社会選好を考慮した効果的な介入戦略を構築するために，以下 4 種類のメッセージ介入群の効果を検証する．

利己損失強調群：自分自身の損失を強調するメッセージを提示する群．自身の利得を最大化／損失を最小化しようとする利己的な傾向が強い個人に効果的であると考えられる．よって，本研究では次のようなメッセージを作成した．「SNS での誹謗中傷への関与は，**あなたの評判を下げ，法的リスクを高める可能性があります**.」太字のように，個人の損失に関する具体的な内容を盛り込むことで，損失回避傾向が強い個人に効果的に働きかけることを狙っている．

利他損失強調群：他者の損失を強調するメッセージを提示する群．他者の利得を最大化／損失を最小化しようとする利他的な傾向が強い個人に効果的であると考えられる．よって，本研究では次のようなメッセージを作成した．「あなたの誹謗中傷への関与は，**アスリートの心理的健康に影響を与える可能性があります**.」太字のように，第三者の損失に関する具体的な内容を盛り込むことで，損失回避傾向が強い個人に効果的に働きかけることを狙っている．

多数派帰属強調群：社会規範メッセージを細分化したうちの一つで，自身が多数派に属することを強く意識させるメッセー

	選択肢 1	選択肢 2	選択肢 3	選択肢 4	選択肢 5	選択肢 6	選択肢 7	選択肢 8	選択肢 9
Q1	10,000	9,400	8,800	8,100	7,500	6,900	6,300	5,600	5,000
	5,000	5,600	6,300	6,900	7,500	8,100	8,800	9,400	10,000
Q2	8,500	8,700	8,900	9,100	9,300	9,400	9,600	9,800	10,000
	1,500	1,900	2,400	2,800	3,300	3,700	4,100	4,600	5,000
Q3	5,000	5,400	5,900	6,300	6,800	7,200	7,600	8,100	8,500
	10,000	9,800	9,600	9,400	9,300	9,100	8,900	8,700	8,500
Q4	5,000	5,400	5,900	6,300	6,800	7,200	7,600	8,100	8,500
	10,000	8,900	7,900	6,800	5,800	4,700	3,600	2,600	1,500
Q5	8,500	8,500	8,500	8,500	8,500	8,500	8,500	8,500	8,500
	8,500	7,600	6,800	5,900	5,000	4,100	3,300	2,400	1,500
Q6	10,000	9,800	9,600	9,400	9,300	9,100	8,900	8,700	8,500
	5,000	5,400	5,900	6,300	6,800	7,200	7,600	8,100	8,500

表 1: SVO スライダー測定法の質問と選択肢（利得局面）．各セルの上の行が自分の利得，下の行が他者の利得を示す．実験参加者は見ず知らずの第三者を想定し，金額配分（Q1〜Q6）の選択肢（1〜9）の中から実験参加者が最も好ましい組み合わせを選択する．

	選択肢 1	選択肢 2	選択肢 3	選択肢 4	選択肢 5	選択肢 6	選択肢 7	選択肢 8	選択肢 9
Q1	0	-600	-1,200	-1,900	-2,500	-3,100	-3,700	-4,400	-5,000
	-5,000	-4,400	-3,700	-3,100	-2,500	-1,900	-1,200	-600	0
Q2	-1,500	-1,300	-1,100	-900	-700	-600	-400	-200	0
	-8,500	-8,100	-7,600	-7,200	-6,700	-6,300	-5,900	-5,400	-5,000
Q3	-5,000	-4,600	-4,100	-3,700	-3,200	-2,800	-2,400	-1,900	-1,500
	0	-200	-400	-600	-700	-900	-1,100	-1,300	-1,500
Q4	-5,000	-4,600	-4,100	-3,700	-3,200	-2,800	-2,400	-1,900	-1,500
	0	-1,100	-2,100	-3,200	-4,200	-5,300	-6,400	-7,400	-8,500
Q5	-1,500	-1,500	-1,500	-1,500	-1,500	-1,500	-1,500	-1,500	-1,500
	-1,500	-2,400	-3,200	-4,100	-5,000	-5,900	-6,700	-7,600	-8,500
Q6	0	-200	-400	-600	-700	-900	-1,100	-1,300	-1,500
	-5,000	-4,600	-4,100	-3,700	-3,200	-2,800	-2,400	-1,900	-1,500

表 2: SVO スライダー測定法の質問と選択肢（損失局面）．各セルの上の行が自分の損失，下の行が他者の損失を示す．実験参加者は見ず知らずの第三者を想定し，金額配分（Q1〜Q6）の選択肢（1〜9）の中から実験参加者が最も好ましい組み合わせを選択する．

(a) テーマ：野球，誹謗中傷あり　(b) テーマ：野球，誹謗中傷なし　(c) テーマ：サッカー，誹謗中傷あり　(d) テーマ：サッカー，誹謗中傷なし

図 1: 実験で提示した 4 種類の投稿．「野球」か「サッカー」をテーマとする投稿に対して，誹謗中傷が含まれる投稿と，含まれない投稿の合計 4 つの組み合わせを用意した．

ジを提示する群．社会的同調傾向が強い個人に効果的であると考えられる．よって，本研究では次のようなメッセージを作成した．「**98.5%の人は誹謗中傷に関与していない**という研究報告があり，あなたの言動が社会から逸脱していないかを考慮す

る必要があります.」太字のように，多数派の行動に関する具体的な内容を盛り込むことで，社会的同調傾向が強く，社会への帰属意識が高い個人に効果的に働きかけることを狙っている．

少数派脱却強調群： 社会規範メッセージを細分化したうちの一つで，自身が少数派でありそこから脱却することを強く意識させるメッセージを提示する群．社会的同調傾向が強い個人に効果的であると考えられる．よって，本研究では次のようなメッセージを作成した．「**たった 1.5%の人しか誹謗中傷に関与していない**という研究報告があり，あなたの言動がその一部の少数派になってしまっていませんか.」太字のように，少数派の行動に関する具体的な内容を盛り込むことで，社会的同調傾向が強く，自身が少数派であることを避けようとする個人に効果的に働きかけることを狙っている．

これら 4 種類の介入メッセージ用いて，個人の社会選好の傾向に合わせた誹謗中傷拡散を効果的に抑制できる介入メッセージがどのようなものかを検証した．

3.5 実験の流れ

本実験は以下の手順で実施した．Freeasy では年齢や性別などの基本的な属性情報を収集できるため，本実験の質問項目ではこれらの情報は収集せず，以下の 6 つのステップで実施した．

1. 同意確認： Freeasy 上で募集した参加希望者の中から選ばれた，実験の分析効果を高めるためのスクリーニングを通過した実験参加者に対して，社会選好の傾向に応じた介入メッセージの効果の検証という本研究の目的についてや．60 分程度の所要時間，いつでも休憩や中断を行えることを説明し，同意をした方のみ次に手順に進んだ．
2. 社会選好の測定： SVO スライダーを参考にした社会選好に関するアンケートを実施する．スクリーニング時点で社会選好に偏りがないことを確認するために一度利得分配の 6 問は測定しているため，本実験中では上記で説明した損失の分配を行う 6 問を提示した．
3. 誹謗中傷を含む／含まない投稿の提示： 野球とサッカーの各テーマにおいて，誹謗中傷を含む投稿と含まない投稿の計 4 種類の投稿を表示した．投稿内容は図 1 の通りである．
4. 反応計測（1 回目）： 提示された各投稿に対する行動意向を測定した．ここでは，いいね，リポスト，引用ポストの 3 種類の拡散行動への行動意向を 0% ~ 100% の範囲で，20% 刻みで回答してもらった．
5. 介入メッセージの提示： 上記の説明の通り，実験参加者が所属する介入群に応じた介入メッセージを提示した．
6. 反応計測（2 回目）： 介入メッセージの提示後，1 回目の反応計測と同様に，いいね，リポスト，引用ポストの 3 種類の拡散行動への行動意向を 0% ~ 100% の範囲で，20% 刻みで回答してもらった．

表 3: 介入群ごとの分析対象となる実験参加者統計

統計量	利己損失	利他損失	多数派帰属	少数派脱却
男性人数	41	43	31	33
女性人数	10	16	15	14
平均年齢	51.1	48.2	43.5	45.2
標準偏差	9.6	12.5	9.0	10.3
参加者数	51	59	46	47

4 分析結果

4.1 実験データの概要

4.1.1 介入メッセージ提示前後の行動意向に関する分布

介入メッセージ効果を示すヒートマップを図 2 に示す．行動意向すべてにおいて，介入メッセージ提示前後で行動意向の変化がある程度見られるものの，介入前後でともに「0%」を選択する参加者が多かったことが分かる．全てのアンケート項目に対して「0%」と回答している実験参加者は，行動意向の変化が計測できないことから，今回の実験参加者の対象外であるため，除外して以降の分析を行った．

4.1.2 実験参加者の性別／年齢分布

分析対象となる実験参加者の年齢，性別の分布を図 3 に示す．分析対象となる人数は合計 203 人で，その内男性が 148 人，女性が 55 人であった．表 3 に各介入群ごとの実験参加者数と平均年齢，年齢の標準偏差を示す．各介入ごとで実験参加者数は概ね 50 人程度であった．

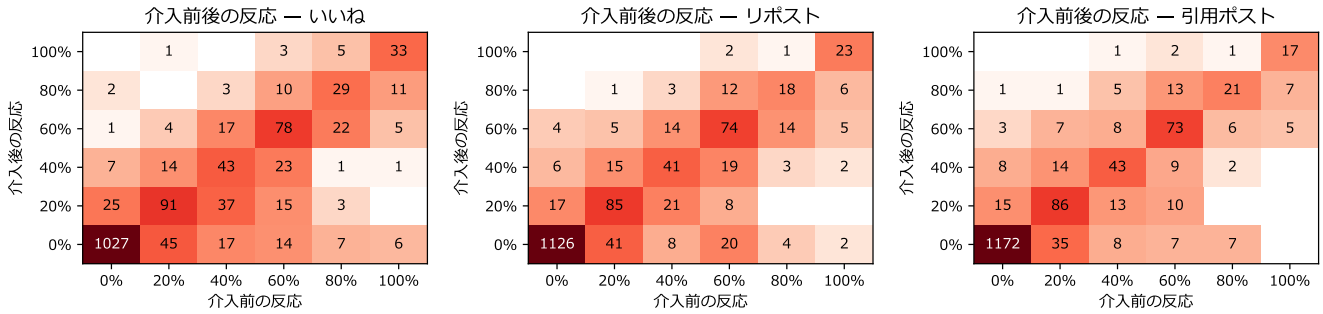
4.1.3 SVO スコアの分布

SVO スコアの分布を図 4 に示す．この図では，値全体を 25 分割したヒストグラムとなっており，SVO_p は刻み幅 3.11°，SVO_n は刻み幅 2.82° となっている．横軸は SVO スコア，縦軸はその値をとった参加者数を示しており，値が大きいほど利他性を強く示す．SVO_p と SVO_n の中央値がそれぞれ 32.45°，28.44° と，SVO_n の方がやや低くなった．これは，損失回避傾向の方が利得配分傾向よりも利己的な傾向を示す参加者が多いことを示している．グラフ内の縦線は中央値を示しており，中央値を基準に参加者の社会選好を利己的（PROSELF）と利他的（PROSOCIAL）に分類した．

4.2 メッセージ提示実験の分析

4.2.1 重回帰分析の式と各変数の説明

本研究では，個人の社会選好に対して介入メッセージがどのような介入効果をもたらすかを分析するため，重回帰分析モデルを構築した．重回帰分析の式は以下の通りである．



（a）いいねの分布　（b）リポストの分布　（c）引用ポストの分布

図 2: 介入前後の行動意向別の人数．行方向にに介入前の行動意向を，列方向に介入後の行動意向を示し，対応する人数をセル内に示している．

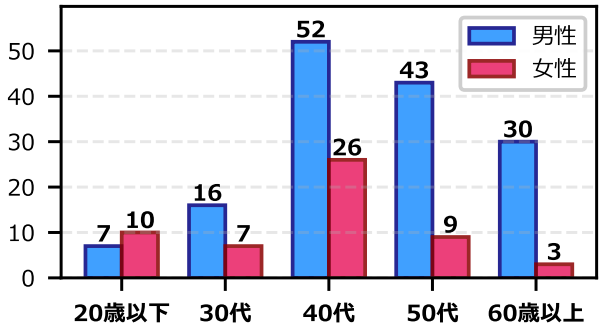
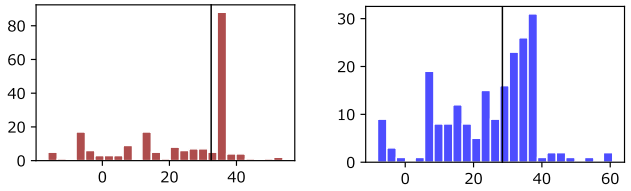


図 3: 実験参加者の性別・年齢分布



（a）SVO_p（利得配分傾向）の分布　（b）SVO_n（損失回避傾向）の分布

図 4: 除外後の SVO スライダーによって測定した SVO_p と SVO_n の分布．それぞれの中央値を，縦軸と平行に引かれている黒の実線で示している．

$$
\begin{aligned}
y_{i,j} = & \beta_0 \\
& + \beta_1 \cdot \mathrm{PROSELF}_x \\
& + \beta_2 \cdot \mathrm{PROSOCIAL}_x \\
& + \beta_3 \cdot \text{利己損失強調} \\
& + \beta_4 \cdot \text{利他損失強調} \\
& + \beta_5 \cdot \text{多数派帰属強調} \\
& + \beta_6 \cdot \text{少数派脱却強調} \\
& + \beta_7 \cdot \mathrm{PROSELF}_x \times \text{利己損失強調} \\
& + \beta_8 \cdot \mathrm{PROSELF}_x \times \text{利他損失強調} \\
& + \beta_9 \cdot \mathrm{PROSELF}_x \times \text{多数派帰属強調} \\
& + \beta_{10} \cdot \mathrm{PROSELF}_x \times \text{少数派脱却強調} \\
& + \beta_{11} \cdot \mathrm{PROSOCIAL}_x \times \text{利己損失強調} \\
& + \beta_{12} \cdot \mathrm{PROSOCIAL}_x \times \text{利他損失強調} \\
& + \beta_{13} \cdot \mathrm{PROSOCIAL}_x \times \text{多数派帰属強調} \\
& + \beta_{14} \cdot \mathrm{PROSOCIAL}_x \times \text{少数派脱却強調} \\
& + \epsilon
\end{aligned}
\tag{1}
$$

目的変数：本研究の目的変数 $y_{i,j}$ は，介入前後の反応の差分を離散化して定義する．具体的には，まず以下の式で結果の差分を算出する．

$$
y_{i,j} = \begin{cases} 1 & \text{介入前の行動意向} > \text{介入後の行動意向} \\ 0 & \text{介入前の行動意向} = \text{介入後の行動意向} \\ -1 & \text{介入前の行動意向} < \text{介入後の行動意向} \end{cases} \tag{2}
$$

ここで，i は投稿テーマ（野球・誹謗中傷あり／なし，サッカー・誹謗中傷あり／なし）種類を表す．また，j は反応の種類（いいね，リポスト，引用ポスト）を表す．$y_{i,j}$ は {−1, 0, 1} の 3 種類の値をとることになり，1 は介入メッセージによって介入後の行動意向が介入前に比べて低下したことを示し，-1 は行動意向が上昇したことを示す．0 は行動意向に介入前後で変化がなかったことを示す．本研究では計 12 種類（投稿テーマ 4 種類×行動意向 3 種類）の目的変数 $y_{i,j}$ に対して重回帰分析を行う．誹謗中傷を含む投稿（$y_{1,1} \sim y_{1,3}$，$y_{3,1} \sim y_{3,3}$）は説明変数の回帰係数が正の値となっている場合に，誹謗中傷拡散の抑制効果があることを示す．一方で，誹謗中傷を含まない投稿（$y_{2,1} \sim y_{2,3}$，$y_{4,1} \sim y_{4,3}$）の回帰係数が正の値となっている場合に，副作用があることを示す．

説明変数：各説明変数については以下の通り．

- **$\mathrm{PROSELF}_x$，$\mathrm{PROSOCIAL}_x$：**個人の社会選好に基づく分類を示すダミー変数であり，利己的（PROSELF）な場合は 1，利他的（PROSOCIAL）な場合は 1，それ以外の場合は 0 として定義する．社会選好の分類は，前節で示した SVO スコアの中央値を基準に行う．x は利得配分局面 p または損失回避局面 n を表す．
- **利己損失強調，利他損失強調，多数派帰属強調，少数派脱却強調：**各介入群を示すダミー変数であり，介入群に属する場合は 1，それ以外の場合は 0 として定義する．
- **交互作用項：**個人の社会選好に基づく分類と各介入群の交互作用を示す変数であり，例えば $\mathrm{PROSELF}_x$× **利己損失強調**は，利己的な参加者が利己損失強調介入群に属する場合に 1，それ以外の場合は 0 として定義する．

以上により，本研究では各目的変数 $y_{i,j}$ に対して，個人の社会選好，各介入群，および個人の社会選好と各介入群の交互作用が行動意向の差分に与える影響を分析した．

4.2.2 重回帰分析の結果

重回帰分析の結果を表 4 ~ 7 に示す．表は投稿テーマごとに SVO_p，SVO_n をそれぞれ説明変数に用いた場合の分析結果である．各表には，説明変数ごとに回帰係数，標準誤差を示している．

表 4 に示すように，テーマが野球で誹謗中傷を含む投稿に対して SVO_p を説明変数に用いた場合，「いいね」において PROSOCIAL$_p$ が回帰係数 0.1023 であり，$p < 0.01$ で有意な抑制効果が見られた．また，利他損失強調 ×PROSOCIAL$_p$ の交互作用項で回帰係数が 0.1748 であり，$p < 0.05$ で有意な抑制効果が見られた．

表 5 に示すように，テーマがサッカーで誹謗中傷を含む投稿に対して SVO_p を説明変数に用いた場合，「引用ポスト」において利他損失強調が回帰係数 0.0729 であり，$p < 0.05$ で有意な抑制効果が見られた．また，誹謗中傷を含まない投稿に対しても，「引用ポスト」において利他損失強調が回帰係数 0.1015 であり，$p < 0.01$ で有意な副作用が確認された．さらに，誹謗中傷を含まない投稿の「引用ポスト」において，利己損失強調 ×PROSOCIAL$_p$ の交互作用項が回帰係数-0.1155 であり，$p < 0.05$ で有意な効果が見られた．

表 6 に示すように，テーマが野球で誹謗中傷を含む投稿に対して SVO_n を説明変数に用いた場合，「いいね」において PROSOCIAL$_n$ が回帰係数 0.1161 であり，$p < 0.01$ で有意な抑制効果が見られた．また，「リポスト」においても PROSOCIAL$_n$ が回帰係数 0.0693 であり，$p < 0.05$ で有意な抑制効果が確認された．さらに，「引用ポスト」においても PROSOCIAL$_n$ が回帰係数 0.1056 であり，$p < 0.01$ で有意な抑制効果が見られた．一方で，誹謗中傷を含まない投稿の「いいね」において，利他損失強調 ×PROSOCIAL$_n$ の交互作用項が回帰係数 0.2101 であり，$p < 0.01$ で有意な副作用が確認された．

表 7 に示すように，テーマがサッカーで誹謗中傷を含む投稿に対して SVO_n を説明変数に用いた場合，「いいね」において PROSOCIAL$_n$ が回帰係数 0.0721 であり，$p < 0.05$ で有意な抑制効果が見られた．また，「リポスト」において PROSOCIAL$_n$ が回帰係数 0.0849 であり，$p < 0.01$ で有意な抑制効果が確認された．一方で，誹謗中傷を含まない投稿に対しては，「リポスト」において PROSOCIAL$_n$ が回帰係数 0.0791 であり，$p < 0.01$ で有意な副作用が見られた．さらに，誹謗中傷を含まない投稿の「引用ポスト」において，利他損失強調 ×PROSOCIAL$_n$ の交互作用項が回帰係数 0.1166 であり，$p < 0.05$ で有意な副作用が確認された．

5 考　　察

5.1 社会選好が持つ誹謗中傷拡散の抑制効果

本研究では，SNS 上の誹謗中傷投稿への同調・拡散を抑制するために，個人の社会選好（Social Value Orientation; SVO）を考慮した介入戦略の有効性を検証した．特に，利得局面で測定される SVO_p と損失局面で測定される SVO_n の差異に着目し，社会選好が誹謗中傷拡散行動に与える影響を分析した．

分析結果では，利得局面の社会選好タイプ（PROSOCIAL$_p$）において，テーマが野球で誹謗中傷を含む投稿に対する「いいね」に有意な抑制効果が確認された．具体的には，PROSOCIAL$_p$ が回帰係数 0.1023（$p < 0.01$）となり，利他的傾向を持つ参加者ほど誹謗中傷投稿への拡散意向が低下する傾向が示された．さらに，利他損失強調メッセージと PROSOCIAL$_p$ の交互作用項においても回帰係数 0.1748（$p < 0.05$）と有意な効果が見られた．

一方で，損失局面で分類した社会選好タイプ（PROSOCIAL$_n$）は，より広範に抑制効果を示した．テーマが野球で誹謗中傷を含む投稿に対して，「いいね」では回帰係数 0.1161（$p < 0.01$），「リポスト」では 0.0693（$p < 0.05$），「引用ポスト」では 0.1056（$p < 0.01$）と複数の行動意向において有意な抑制効果が確認された．また，テーマがサッカーの場合でも，「いいね」（0.0721, $p < 0.05$）や「リポスト」（0.0849, $p < 0.01$）において抑制効果が見られた．

このように，利得局面よりも損失局面で測定した社会選好の方が誹謗中傷拡散抑制に寄与していたことから，本研究で提示した介入メッセージが損失回避を強調する内容であった点が影響している可能性がある．損失局面の SVO による分類は，誹謗中傷拡散抑制においてより適切な個別化指標となり得ることが示唆された．

しかしながら，本研究で導入した社会選好タイプと介入メッセージの交互作用項については，誹謗中傷投稿に対する有意な抑制効果がほとんど確認されなかった．交互作用項が有意となったのは誹謗中傷を含まない投稿に対する副作用としての結果であり，期待された「社会選好に応じたメッセージの出し分けによる誹謗中傷抑制」は十分に実証されなかった．

抑制効果が見られなかった要因として，第一にサンプルサイズの制約による統計的検出力不足が考えられる．本研究では分析対象者が 203 名に絞られており，社会選好タイプと介入群の組み合わせごとの人数が少なくなったことで，交互作用効果を検出することが困難であった可能性がある．第二に，社会規範メッセージのように短い文面で受け手に規範意識を十分伝えることが難しく，個別化しても効果が現れにくかったことが考えられる．以上より，社会選好は個別化介入の手がかりとなり得る一方で，メッセージ設計との適合性を慎重に検討する必要がある．

5.2 誹謗中傷を含まない投稿の拡散を抑制する副作用

本研究では，誹謗中傷投稿への拡散抑制を目的とした介入メッセージが，誹謗中傷を含まない通常投稿に対しても拡散意向を低下させる副作用を引き起こす可能性についても検討した．先行研究 [8] でも指摘されていたように，介入メッセージが誹謗中傷投稿のみを選択的に抑制することは難しく，通常投稿への拡散行動まで抑制してしまうことが課題となる．

分析結果では，誹謗中傷を含まない投稿に対しても有意な効果が複数確認された．例えば，テーマがサッカーの場合，「引用ポスト」において利他損失強調メッセージが回帰係数 0.1015（$p < 0.01$）となり，誹謗中傷を含まない投稿の拡散意向を抑制する副作用が見られた．また，損失局面の社会選好に基づく交

互作用項でも，誹謗中傷を含まない投稿の「いいね」において利他損失強調 ×PROSOCIAL$_n$ が回帰係数 0.2101（$p < 0.01$）となり，有意な副作用が確認された．さらに，テーマがサッカーで誹謗中傷を含まない投稿に対しても，「リポスト」で PROSOCIAL$_n$ が 0.0791（$p < 0.01$）と有意であり，利他的傾向を持つ参加者ほど通常投稿への拡散も抑制される可能性が示された．

これらの結果から，介入メッセージは誹謗中傷拡散を抑制するだけでなく，SNS 上の一般的な情報共有行動まで萎縮させるリスクを持つことが明らかになった．特に利他性や損失回避を強調するメッセージは，受け手に対して「拡散行動そのものが望ましくない」という印象を与える可能性がある．

また，本研究の実験デザインは先行研究と異なり，介入前後の行動意向を測定し差分を目的変数とする形式を採用した．この設計の利点として，介入メッセージが参加者の意向にどのような変化を与えたかを直接的に捉えられる点が挙げられる．介入前の意向を基準とすることで，個人差を考慮したより詳細な分析が可能となった．

一方で欠点として，意向変化を $\{-1, 0, 1\}$ に離散化したことで情報が粗くなり，微細な変化が捨象される可能性がある．また，介入前の意向自体が低い参加者が多く，0%回答者を除外する必要が生じたため，サンプルサイズが縮小した．さらに，「意向の変化」を測定することは実際の行動とは必ずしも一致しないため，現実の SNS 環境での拡散抑制効果を推定するには限界が残る．

以上より，本研究は介入メッセージによる誹謗中傷拡散抑制の可能性を示すと同時に，通常投稿への副作用や実験デザイン上の課題を明らかにした．今後は誹謗中傷投稿のみを選択的に抑制できる介入方法の設計や，実際の行動データを用いた検証が求められる．

5.3 本研究の限界と課題

サンプルサイズの制約：本研究ではスクリーニングで条件を満たした実験参加希望者に本実験に参加してもらった．しかし，参加希望者 15,000 名に対してスクリーニング条件を満たした実験参加者が 400 名で，その中で適切な回答ができた参加者が 203 名とさらにサンプルサイズが小さくなったため，統計的検出力が不足している可能性がある．スポーツという興味が限定的であるテーマを扱ったのは，興味の有無がはっきりしているテーマの方がより適切な参加者を募集できると考えたためであるが，ここまで絞られてしまった原因は，参加者を募集したクラウドソーシングサービスでのスポーツへの興味関心が低かったからだと思われる．考察にも挙げたが，今回の実験設計でより効果的な分析結果を得るためには，より多くの参加者を集め，説明変数を増やすことで様々な要因を考慮できる必要がある．

若年層のサンプル不足：本研究の実験参加者は若年層が少なく，特に 18 歳から 20 歳の参加者が非常に少なかった．若年層は SNS 利用率が高く，誹謗中傷投稿の拡散行動にも大きく関連があると考えられる．Ungaretti らによると [20]，若年層ほど加害経験が多いという研究結果もあるため，今後同様の実験を行う際は若年層の参加者をより多く集めることが重要であると考えられる．

誹謗中傷の定義：先行研究では，誹謗中傷を特定の単語を含むかどうかで定義していたが，本研究ではより多様な誹謗中傷を実験に使用するため，誹謗中傷の 9 つの定義 [19] を用いて投稿を選択・設計し，反応測定を行うという実験を行った．しかし，今回の実験での誹謗中傷の捉え方が参加者にとって一様だったかは不明である．誹謗中傷の認識は個人差が大きく，文化的背景や価値観によっても異なるため，今後の研究では，より明確な定義や具体的な例を提示することで，参加者の理解を統一する必要がある．例えば，誹謗中傷の定義として，本研究でも投稿を設定することに使用した 9 つの定義を参加者に事前に説明するなどが考えられるが，詳細に説明しすぎると実験の自然性の欠如や本来のプラットフォームでの行動との乖離が顕著になる可能性もあるため，そのバランスを考慮する必要がある．

5.4 今後の展望

異なる目的変数の検証：介入前後の反応の差分を離散化したもののみを目的変数とするのではなく，介入前の反応を考慮しない最終的に投稿を拡散したかどうかの結果だけに着目した別の目的変数の設定などを試し，今回の実験手法で得られた結果と比較することなどが考えられる．特に，結果だけに着目した場合は誹謗中傷投稿の拡散を有意に抑制できる組み合わせが存在するかを再度検証する必要があると考えている．

参加者の増加・テーマの多様化：より多くの参加者を集めることで，年齢や性別，年収や生活の満足度など誹謗中傷に関連性がありそうな要素を説明変数として分析に使用できるようになる．また，誹謗中傷に関するテーマをスポーツ以外にも多様化する工夫も今後行っていきたい．例えば，政治，芸能，社会問題など，スポーツ以外の誹謗中傷が発生しやすいテーマを取り入れることが考えられる．

個人の社会選好以外の個人傾向との比較：本研究では，個人の社会選好の傾向に基づいて介入メッセージの効果を検証したが，今後は他の個人傾向（例えば，性格特性，価値観，過去の SNS 利用経験など）と比較し，どの個人傾向が誹謗中傷拡散行動に最も影響を与えるかを明らかにすることを検討している．

6 おわりに

本研究は，SNS 上の誹謗中傷拡散を抑制するため，個人の社会選好（SVO）を考慮したメッセージ介入の有効性を検証することを目的として実施したオンライン実験である．利己損失強調，利他損失強調，多数派帰属強調，少数派脱却強調の 4 種類の介入を用い，SVO_p および SVO_n の両面から効果を分析した．

主な知見は以下のとおりである．今回の介入前後の反応の差分を離散化したものを目的変数とした分析では，介入群のみ，社会選好タイプのみを考慮する場合には有意な抑制効果を持つものが一部見られた．また，社会選好タイプは損失局面で分類した方がより効果的に介入メッセージの出し分けができた．一方で誹謗中傷拡散を抑制する有意な効果を持つ介入メッセージ

と個人の社会選好傾向の組み合わせはなく，有意だった組み合わせは誹謗中傷を含まない通常投稿の拡散まで抑制してしまう副作用としての効果だった．

これらの結果から，社会選好は個別化の有力な手がかりになりうる値ではある一方で，メッセージ内容との適合度を十分に検討したうえでの個別化指標であることが分かった．しかし現状の実験設計では相性の良い組合わせの場合であっても抑制効果が副作用としてのみ確認されたため，今後はより効果的な介入メッセージの設計や異なる要素の導入を検討する必要がある．

本研究はナッジメッセージを SVO による個別化を利用して出し分けることで効果的な誹謗中傷拡散の抑制が可能かを検証しメッセージとの適合度が高い社会選好タイプは有意な抑制効果を持ち，個人の社会選好と介入メッセージの組み合わせによって有意な抑制効果を示すものが存在するということが分かったため，いくつかの課題を残してはいるものの SNS 上の有害情報拡散抑制に向けた個別化介入の可能性を示す第一歩となった．今後も多様な個人の属性や個人差指標を活用した介入手法の開発と検証を進めたい．

謝　　辞

本研究の一部は，JSPS 科研費 24K03046，24K23877 の助成を受けたものです．ここに記して謝意を示します．

文　　献

[1] 総務省 2025. 令和 7 年版情報通信白書 第 1 部第 2 章第 3 節 インターネット上の偽・誤情報等への対応, 2025. https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r07/html/nd123100.html（参照 2026-01-05）.

[2] 総務省 2023. 令和 5 年度インターネット上の違法・有害情報対応相談業務等請負業務報告書（概要版）. https://www.soumu.go.jp/main_content/000946765.pdf（参照 2026-01-05）.

[3] Masanori Takano, Fumiaki Taka, Chiki Ogiue, and Natsuki Nagata. Online harassment of japanese celebrities and influencers. *Frontiers in psychology*, Vol. 15, p. 1386146, 2024.

[4] Ruth Jeong, Megan Gilbertson, Logan N Riffle, and Michelle K Demaray. Participant role behavior in cyberbullying: An examination of moral disengagement among college students. *International journal of bullying prevention*, Vol. 6, No. 1, pp. 28–40, 2024.

[5] 総務省. プロバイダ責任制限法の一部を改正する法律（概要）, 2022. https://www.soumu.go.jp/main_content/000836903.pdf（参照 2026-01-06）.

[6] 西谷千乃与, 安藤一秋. アスリートに対する sns 上の誹謗中傷の分析と検出. Web インテリジェンスとインタラクション研究会予稿集 第 20 回研究会, pp. 25–32. Web インテリジェンスとインタラクション研究会, 2024.

[7] 正木博明, 柴田健吾, 星野秀偉, 石濱嵩博, 齋藤長行, 矢谷浩司. 若年層 sns ユーザに対するプライバシ・安全上の行動に関するナッジの大規模定量調査. 情報処理学会論文誌, Vol. 61, No. 12, pp. 1892–1902, 2020.

[8] 村田拡介, 秋田航汰, 小野祐紀, 河合求真, 森実輝. Sns 上の誹謗中傷投稿への同調に関する分析: ナッジメッセージによる誹謗中傷抑制の試み. 大阪大学経済学, Vol. 73, No. 4, pp. 55–55, 2024.

[9] Susan Athey, Niall Keleher, and Jann Spiess. Machine learning who to nudge: causal vs predictive targeting in a field experiment on student financial aid renewal. *Journal of Econometrics*, Vol. 249, p. 105945, 2025.

[10] Benjamin E Hilbig, Isabel Thielmann, Ingo Zettler, and Morten Moshagen. The dispositional essence of proactive social preferences: The dark core of personality vis-à-vis 58 traits. *Psychological Science*, Vol. 34, No. 2, pp. 201–220, 2023.

[11] Ryan O. Murphy, Kurt A. Ackermann, and Michel Handgraaf. Measuring social value orientation. *Judgment and Decision Making*, Vol. 6, No. 8, pp. 771–781, 2011.

[12] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pp. 745–760. Springer, 2018.

[13] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11, pp. 512–515, 2017.

[14] Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, New Haven, CT, 2008.

[15] Kevin Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, Vol. 39, No. 3, pp. 629–649, 2017.

[16] Jan Luca Pletzer, Daniel Balliet, Jeff Joireman, D Michael Kuhlman, Sven C Voelpel, and Paul AM Van Lange. Social value orientation, expectations, and cooperation in social dilemmas: A meta–analysis. *European Journal of Personality*, Vol. 32, No. 1, pp. 62–83, 2018.

[17] Alex Williams and Li Chen. Rapid diffusion of sports-related abusive posts after events. *Journal of Online Behavior*, Vol. 12, No. 1, pp. 15–32, 2025.

[18] 笹川スポーツ財団. スポーツ観戦率：スタジアムや体育館などでの直接スポーツ観戦率は 26.2%, 12 2025. https://www.ssf.or.jp/thinktank/sports_life/data/sportsspectating.html（参照 2026-01-07）.

[19] 国際大学グローバル・コミュニケーション・センター (GLOCOM). Innovation nippon 2022: ジャーナリストへの誹謗中傷の実態 報告書〈概要版〉. Technical report, 国際大学グローバル・コミュニケーション・センター (GLOCOM), April 2023. 概要版 PDF, 調査報告書.

[20] Joaquín Ungaretti, Talía Gómez Yepes, María Laura Sánchez Pujalte, and Edgardo Etchezahar. Cyberbullying perpetration among spanish adults: The roles of fear of missing out and critical thinking. *Societies*, Vol. 15, No. 9, p. 249, 2025.

表 4: テーマが野球であり，かつ SVO_p を説明変数に用いたときの重回帰分析の結果，推定された回帰係数．表内の値は回帰係数と括弧内に標準誤差を示す．有意な値には，$p < 0.05$ のとき下線，$p < 0.01$ のとき**太字**を付す．

説明変数	いいね		リポスト		引用ポスト	
	誹謗中傷あり	誹謗中傷なし	誹謗中傷あり	誹謗中傷なし	誹謗中傷あり	誹謗中傷なし
切片	**0.1104 (0.021)**	**0.0862 (0.024)**	**0.0927 (0.021)**	0.0350 (0.021)	0.0289 (0.019)	0.0332 (0.020)
利己損失強調	0.0458 (0.047)	-0.0371 (0.053)	0.0176 (0.046)	0.0220 (0.045)	0.0286 (0.041)	0.0086 (0.043)
利他損失強調	0.0480 (0.044)	0.0863 (0.050)	0.0293 (0.043)	-0.0252 (0.043)	0.0174 (0.039)	-0.0100 (0.041)
多数派帰属強調	0.0115 (0.048)	0.0281 (0.055)	0.0267 (0.048)	0.0317 (0.047)	-0.0130 (0.043)	0.0227 (0.045)
少数派脱却強調	0.0050 (0.048)	0.0089 (0.055)	0.0191 (0.047)	0.0064 (0.047)	-0.0042 (0.043)	0.0119 (0.045)
PROSELF_p	0.0081 (0.034)	0.0337 (0.039)	0.0480 (0.034)	0.0482 (0.033)	-0.0171 (0.030)	0.0049 (0.032)
PROSOCIAL_p	**0.1023 (0.034)**	0.0525 (0.038)	0.0447 (0.033)	-0.0132 (0.033)	0.0460 (0.030)	0.0283 (0.031)
利己損失強調× PROSELF_p	0.0629 (0.076)	0.0990 (0.086)	0.0691 (0.075)	0.0312 (0.074)	-0.0404 (0.067)	0.0442 (0.070)
利己損失強調× PROSOCIAL_p	-0.0171 (0.071)	-0.1361 (0.081)	-0.0515 (0.070)	-0.0092 (0.069)	0.0690 (0.063)	-0.0356 (0.066)
利他損失強調× PROSELF_p	0.0687 (0.067)	-0.0886 (0.076)	0.0359 (0.066)	-0.0286 (0.066)	0.0296 (0.059)	0.0014 (0.062)
利他損失強調× PROSOCIAL_p	-0.0207 (0.072)	0.1749 (0.082)	-0.0066 (0.071)	0.0034 (0.070)	-0.0122 (0.064)	-0.0114 (0.067)
多数派帰属強調× PROSELF_p	-0.0500 (0.075)	0.0520 (0.085)	-0.0473 (0.074)	0.0851 (0.073)	0.0012 (0.066)	0.0592 (0.069)
多数派帰属強調× PROSOCIAL_p	0.0615 (0.078)	-0.0239 (0.088)	0.0740 (0.077)	-0.0534 (0.076)	-0.0142 (0.069)	-0.0365 (0.072)
少数派脱却強調× PROSELF_p	-0.0735 (0.079)	-0.0288 (0.089)	-0.0097 (0.078)	-0.0396 (0.077)	-0.0076 (0.070)	-0.1000 (0.073)
少数派脱却強調× PROSOCIAL_p	0.0786 (0.073)	0.0376 (0.083)	0.0288 (0.072)	0.0460 (0.071)	0.0034 (0.065)	0.1118 (0.068)

表 5: テーマがサッカーであり，かつ SVO_p を説明変数に用いたときの重回帰分析の結果．表内の値は回帰係数と括弧内に標準誤差を示す．有意な値には，$p < 0.05$ のとき下線，$p < 0.01$ のとき**太字**を付す．

説明変数	いいね		リポスト		引用ポスト	
	誹謗中傷あり	誹謗中傷なし	誹謗中傷あり	誹謗中傷なし	誹謗中傷あり	誹謗中傷なし
切片	**0.0789 (0.022)**	0.0259 (0.022)	0.0444 (0.020)	0.0217 (0.019)	0.0126 (0.017)	-0.0055 (0.017)
利己損失強調	0.0809 (0.047)	-0.0374 (0.048)	0.0319 (0.043)	-0.0102 (0.041)	-0.0393 (0.036)	-0.0254 (0.037)
利他損失強調	0.0136 (0.045)	0.0435 (0.046)	-0.0311 (0.041)	0.0548 (0.038)	0.0729 (0.034)	**0.1015 (0.035)**
多数派帰属強調	0.0220 (0.049)	0.0458 (0.050)	0.0591 (0.045)	0.0075 (0.042)	0.0083 (0.038)	-0.0371 (0.039)
少数派脱却強調	-0.0376 (0.049)	-0.0259 (0.050)	-0.0154 (0.045)	-0.0303 (0.042)	-0.0293 (0.038)	-0.0445 (0.039)
PROSELF_p	0.0411 (0.035)	0.0157 (0.035)	0.0078 (0.032)	-0.0278 (0.030)	-0.0124 (0.027)	-0.0138 (0.027)
PROSOCIAL_p	0.0378 (0.034)	0.0102 (0.035)	0.0367 (0.031)	0.0495 (0.029)	0.0250 (0.026)	0.0083 (0.027)
利己損失強調× PROSELF_p	0.0717 (0.077)	-0.0042 (0.079)	0.0069 (0.071)	0.0163 (0.066)	-0.0064 (0.059)	0.0901 (0.061)
利己損失強調× PROSOCIAL_p	0.0092 (0.072)	-0.0332 (0.074)	0.0250 (0.066)	-0.0265 (0.062)	-0.0329 (0.056)	-0.1155 (0.057)
利他損失強調× PROSELF_p	-0.0160 (0.069)	0.0031 (0.070)	-0.0211 (0.063)	-0.0192 (0.059)	0.1034 (0.053)	0.0060 (0.054)
利他損失強調× PROSOCIAL_p	0.0296 (0.073)	0.0404 (0.074)	-0.0100 (0.067)	0.0740 (0.063)	-0.0305 (0.056)	0.0956 (0.058)
多数派帰属強調× PROSELF_p	0.0179 (0.076)	0.0325 (0.077)	0.0087 (0.070)	0.0386 (0.065)	-0.0885 (0.058)	-0.0237 (0.060)
多数派帰属強調× PROSOCIAL_p	0.0041 (0.079)	0.0133 (0.081)	0.0503 (0.072)	-0.0311 (0.068)	0.0969 (0.061)	-0.0134 (0.063)
少数派脱却強調× PROSELF_p	-0.0325 (0.080)	-0.0157 (0.082)	0.0132 (0.073)	-0.0635 (0.069)	-0.0209 (0.062)	-0.0862 (0.064)
少数派脱却強調× PROSOCIAL_p	-0.0051 (0.075)	-0.0102 (0.076)	-0.0286 (0.068)	0.0332 (0.064)	-0.0084 (0.057)	0.0417 (0.059)

表 6: テーマが野球であり，かつ SVO_n を説明変数に用いたときの重回帰分析の結果．表内の値は回帰係数と括弧内に標準誤差を示す． 有意な値には，$p < 0.05$ のとき下線，$p < 0.01$ のとき**太字**を付す．

説明変数	いいね		リポスト		引用ポスト	
	誹謗中傷あり	誹謗中傷なし	誹謗中傷あり	誹謗中傷なし	誹謗中傷あり	誹謗中傷なし
切片	**0.1149 (0.021)**	**0.0790 (0.024)**	**0.0913 (0.021)**	0.0340 (0.021)	0.0299 (0.018)	0.0356 (0.020)
利己損失強調	0.0461 (0.046)	-0.0355 (0.052)	0.0195 (0.045)	0.0136 (0.045)	0.0270 (0.040)	0.0001 (0.043)
利他損失強調	0.0387 (0.043)	0.0707 (0.049)	0.0252 (0.043)	-0.0197 (0.043)	0.0059 (0.038)	-0.0109 (0.041)
多数派帰属強調	0.0151 (0.048)	0.0250 (0.055)	0.0231 (0.047)	0.0329 (0.047)	-0.0020 (0.042)	0.0241 (0.045)
少数派脱却強調	0.0150 (0.047)	0.0188 (0.054)	0.0235 (0.046)	0.0071 (0.046)	-0.0010 (0.041)	0.0223 (0.044)
PROSELF$_n$	-0.0012 (0.033)	0.0638 (0.038)	0.0220 (0.033)	0.0438 (0.033)	-0.0757 (0.029)	-0.0060 (0.031)
PROSOCIAL$_n$	**0.1161 (0.034)**	0.0152 (0.038)	0.0693 (0.033)	-0.0098 (0.033)	**0.1056 (0.029)**	0.0416 (0.031)
利己損失強調× PROSELF$_n$	0.1446 (0.074)	0.0232 (0.085)	0.1281 (0.073)	-0.0914 (0.073)	-0.0247 (0.064)	-0.0297 (0.069)
利己損失強調× PROSOCIAL$_n$	-0.0985 (0.071)	-0.0587 (0.081)	-0.1086 (0.070)	0.1051 (0.070)	0.0517 (0.061)	0.0298 (0.066)
利他損失強調× PROSELF$_n$	-0.0042 (0.070)	-0.1394 (0.080)	-0.1015 (0.069)	0.0160 (0.068)	-0.0713 (0.061)	0.0554 (0.065)
利他損失強調× PROSOCIAL$_n$	0.0429 (0.067)	**0.2101 (0.077)**	0.1267 (0.066)	-0.0357 (0.066)	0.0772 (0.058)	-0.0663 (0.063)
多数派帰属強調× PROSELF$_n$	-0.0547 (0.073)	0.0915 (0.083)	0.0488 (0.072)	0.0374 (0.072)	-0.0263 (0.063)	0.0203 (0.068)
多数派帰属強調× PROSOCIAL$_n$	0.0698 (0.080)	-0.0666 (0.091)	-0.0257 (0.078)	-0.0045 (0.078)	0.0243 (0.069)	0.0039 (0.074)
少数派脱却強調× PROSELF$_n$	-0.0870 (0.074)	0.0884 (0.085)	-0.0534 (0.073)	0.0818 (0.073)	0.0467 (0.064)	-0.0520 (0.069)
少数派脱却強調× PROSOCIAL$_n$	0.1019 (0.075)	-0.0696 (0.086)	0.0769 (0.074)	-0.0747 (0.074)	-0.0477 (0.065)	0.0743 (0.070)

表 7: テーマがサッカーであり，かつ SVO_n を説明変数に用いたときの重回帰分析の結果，推定された回帰係数．表内の値は回帰係数と括弧内に標準誤差を示す． 有意な値には，$p < 0.05$ のとき下線，$p < 0.01$ のとき**太字**を付す．

説明変数	いいね		リポスト		引用ポスト	
	誹謗中傷あり	誹謗中傷なし	誹謗中傷あり	誹謗中傷なし	誹謗中傷あり	誹謗中傷なし
切片	**0.0776 (0.022)**	0.0232 (0.022)	0.0425 (0.020)	0.0184 (0.018)	0.0154 (0.017)	-0.0087 (0.017)
利己損失強調	0.0834 (0.047)	-0.0377 (0.048)	0.0290 (0.043)	-0.0116 (0.040)	-0.0418 (0.036)	-0.0270 (0.037)
利他損失強調	0.0038 (0.044)	0.0393 (0.045)	-0.0378 (0.040)	0.0403 (0.037)	0.0738 (0.034)	**0.0940 (0.035)**
多数派帰属強調	0.0244 (0.049)	0.0437 (0.050)	0.0648 (0.045)	0.0063 (0.042)	0.0126 (0.038)	-0.0439 (0.039)
少数派脱却強調	-0.0341 (0.048)	-0.0220 (0.049)	-0.0135 (0.044)	-0.0166 (0.041)	-0.0292 (0.037)	-0.0318 (0.038)
PROSELF$_n$	0.0055 (0.034)	-0.0190 (0.035)	-0.0425 (0.031)	-0.0608 (0.029)	-0.0250 (0.026)	-0.0256 (0.027)
PROSOCIAL$_n$	0.0721 (0.034)	0.0422 (0.035)	**0.0849 (0.031)**	**0.0791 (0.029)**	0.0403 (0.027)	0.0168 (0.027)
利己損失強調× PROSELF$_n$	0.1379 (0.075)	-0.0100 (0.077)	-0.0290 (0.069)	-0.0329 (0.064)	0.0079 (0.059)	0.0613 (0.060)
利己損失強調× PROSOCIAL$_n$	-0.0545 (0.072)	-0.0277 (0.074)	0.0580 (0.066)	0.0213 (0.061)	0.0268 (0.053)	-0.0883 (0.057)
利他損失強調× PROSELF$_n$	-0.1240 (0.071)	-0.0435 (0.072)	-0.0733 (0.065)	-0.0720 (0.060)	0.0470 (0.055)	-0.0227 (0.056)
利他損失強調× PROSOCIAL$_n$	0.1278 (0.068)	0.0828 (0.070)	0.0355 (0.062)	0.1122 (0.058)	0.0268 (0.053)	0.1166 (0.054)
多数派帰属強調× PROSELF$_n$	0.0406 (0.074)	0.1002 (0.076)	0.0463 (0.068)	0.1101 (0.063)	-0.0770 (0.058)	0.0782 (0.059)
多数派帰属強調× PROSOCIAL$_n$	-0.0162 (0.081)	-0.0565 (0.082)	0.0184 (0.074)	-0.1038 (0.068)	0.0896 (0.063)	-0.1221 (0.064)
少数派脱却強調× PROSELF$_n$	-0.0490 (0.075)	-0.0656 (0.077)	0.0135 (0.069)	-0.0660 (0.064)	-0.0028 (0.059)	-0.1423 (0.060)
少数派脱却強調× PROSOCIAL$_n$	0.0149 (0.076)	0.0436 (0.078)	-0.0269 (0.070)	0.0495 (0.065)	-0.0264 (0.059)	0.1106 (0.060)