博 士 論 文

怒りが Disinformation の共有に及ぼす影響とその対策

Haruka SUZUKI
鈴木　悠

情報セキュリティ大学院大学
情報セキュリティ研究科
情報セキュリティ専攻

2025 年 9 月

目次

1. 序論

2010 年代以降，スマートフォンの普及と共にソーシャルメディアの利用率が上がり[1]，マスメディアからの情報を単方向で受け取る消費者に過ぎなかった個人が，世界規模で情報発信をすることができるようになった．ソーシャルメディアとは，インターネットを利用して誰でも手軽に情報を発信し，相互のやりとりができる双方向のメディアである[2]．マスメディアによるニュース等の公共性の高いマクロ情報から，個人による投稿コンテンツ等の意見や見解といったミクロ情報まで，様々な情報がソーシャルメディアに溢れるようになった．このような多様な情報が瞬時に世界規模で伝播していくことから，ソーシャルメディアは速報性・拡散性に優れるとされている．

しかし，投稿コンテンツはソーシャルメディア上で連鎖的に共有される過程において情報源・発信者と切り離されていくという特徴があり[3]，情報の信憑性が大きな問題となっている．2017 年に欧州評議会は，ソーシャルメディアにおいて世界規模での情報汚染が起こっていると指摘した[4]．歴史的に，戦時における情報戦の一環としてプロパガンダや，意図的に虚偽又は事実が織り交ぜられた情報等を指す Disinformation が拡散されてきたが，ソーシャルメディアの持つ信憑性の曖昧さが平時の情報戦「Cyber Influence Operations」の場として悪用されている．Cyber Influence Operations は，対象者の選択，考え，意見，感情，又は動機に影響を与えるためにサイバー空間で実行される情報操作及び影響工作を指す[5]．このような悪意のある影響から個人や集団を守るためには，認知的レジリエンスを強化する「コグニティブセキュリティ」[6]の観点からの対策が必要である．

本研究は，Disinformation を共有するユーザを対象に，Disinformation の共有を減らす効果的な対策を提案することを目指す．現在問題となっているのはソーシャルメディア上での Disinformation の拡散であるが，これらは個々のユーザによる共有の連鎖によって生じている．Disinformation を共有したユーザが拡散することを意図していたのか，またユーザの共有が将来的に拡散につながるかどうかは分からないため，ユーザ向けの対策として拡散を抑制することは難しいと考えられる．このため，本研究では拡散の連鎖を生み出す個々のユーザを対象とした Disinformation の共有を減らす対策を提案することによって，ソーシャルメディア上での Disinformation の拡散の抑制及び情報環境の健全化に貢献する．

1.1. Disinformation の定義

「Disinformation」という言葉は，1887 年に米カンザス州の新聞[7]で初めて使用されたことが確認されている．ソ連が第一次世界大戦においてドイツ軍が敵対勢力を混乱させるために流布した情報操作技術を採用し[8]，ロシア語の「дезинформация」を英訳した「Dezinformatsiya」を使用し始めたのが語源ともされる[9]．情報を指す「information」に接頭辞の「Dis」を使用することで，否定的な情報という意味合いが付与されている．

政策的な取組みにおいて Disinformation をはじめて定義したのは EU である．欧州理事会は，2015 年に欧州対外行動庁（EEAS）に「East StratCom Task Force」を設置し，ロシアが流布した Disinformation に対抗すべく「EUvsDisinfo」というレビューサイトの運用を開始した[10]．2017 年には欧州評議会が，虚偽（FALSE）と有害（HARMFUL）の観点からの識別による，Mis-information，Dis-information，及び Mal-information という新たな概念的枠組みを提案した（図 1-1）[4]．この中で，Disinformation は「害をもたらすために意図的に共有される虚偽情報」と定義されている．

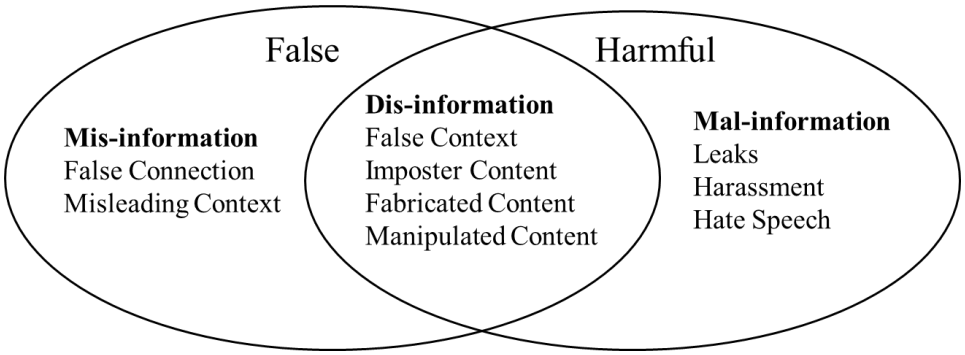


図 1-1 Disinformation の定義（欧州評議会, 2017）

一方，2018 年に欧州委員会が設立した EU ハイレベル専門家グループ（HLEG）は，意図性により Disinformation と Misinformation を分類した．HLEG による Disinformation の定義は，「意図的に公共の害をもたらすため，あるいは利益を得るために設計，表示，及び宣伝された，あらゆる形態の虚偽の，不正確な，又は誤解を招くような情報」[11]である．これを参考に，欧州委員会は政策文書 COM(2018) 236[12]を作成・公表し，Disinformation の定義を「経済的利益を得るため，あるいは意図的に公衆を欺くために作成，表示，及び流布され，公共の害をもたらす可能性のある，検証可能な虚偽又は誤解を招くような情報」とした．同様に，イギリスの DCMS 下院特別委員会も意図性による識別を採用して

いる．2019 年発行の報告書では，Disinformation は「害をもたらす目的で，あるいは政治的，個人的又は金銭的な利益のために，公衆を欺き，誤解させることを意図した，虚偽及び/又は操作された情報を意図的に作成し，共有すること」，Misinformation は「不注意により虚偽情報を共有すること」と定義された[13].

しかし，Disinformation には事実と捏造が混在したものや，どこまでも正しい情報を出すように意図されたものもある[14]．長迫は，Disinformation には不都合な真実も含まれるとし，「害意を以て故意に広められ，真なる情報と偽の情報の双方を含むものの，それが誤った文脈や詐欺的な内容，でっち上げや操作された内容に組み合わされることで，攻撃対象を認知するプロセスを歪ませる情報の集合体」という定義を提案した[15]．日本では 2021 年に Disinformation 対策フォーラムが，EU ハイレベル専門家グループ（HLEG）の定義を基本的には踏襲しながらも，「より広範にインターネットを通じて流通する情報による社会的・経済的・公共的利益が損なわれる事態を議論の対象とすることが適切である」と述べた．経済的利益や公衆を欺く意図の無い誤情報（Misinformation）と事実を基にした情報であるが個人や集団に害を与える意図をもって発出される情報（Malinformation）についても議論の対象とすべく，Disinformation を「あらゆる形態における虚偽の，不正確な，又は誤解を招くような情報で，設計・表示・宣伝される等を通して，公共に危害が与えられた，又は，与える可能性が高いもの」と定義した[16].

このように様々な定義がある中，本論文では日本の Disinformation 対策フォーラムによる定義を採用する．第一に，Disinformation は必ずしも虚偽ではなく，事実又は事実と虚偽が混在している場合がある．虚偽情報はファクトチェックによる事実の証明を以って反論や削除等の対処が可能であるが，必ずしも虚偽ではない場合には当該対処は困難となる．このような問題は，虚偽情報のみを対象とする欧州評議会の定義では解決することができない．第二に，Disinformation は違法ではないが有害な場合がある．違法なコンテンツと認定され得る情報（例えば，ヘイトスピーチ）は法的対処が可能だが，ヘイトスピーチのような攻撃性はなく，社会に不当に扱われている被害者であると不満や怒りを煽ることで社会的な意見の対立・分断から社会を混乱させるような有害な情報には対処することができない． Disinformation の拡散による「社会的・経済的・公共的利益が損なわれる事態」の解決策を検討するにあたっては，虚偽情報だけではなく，違法性がないものの有害な事実又は事実と虚偽の混在についても対象とするのが相応である．

1.2. 背景

ソーシャルメディアが情報拡散の場として注目されたきっかけの 1 つにアラブの春がある．アラブの春は 2010 年から 2011 年にかけてアラブ諸国において発生した民主化運動の総称であり，ソーシャルメディアで呼び掛けられた市民が集まることで大規模な抗議活動へとつながった[17]．Facebook が初期段階における情報拡散のインフラとなり，抗議につながる不満が急速に伝播したとされる[18]．ソーシャルメディアは情報の速報性や拡散性に優れる反面，信憑性や情動伝染を仲介する媒介として課題があることが指摘されている[19].

このソーシャルメディアが持つ影響力に注目したのがロシアである[20]．ロシアは 1920 年代から戦略的に Disinformation を用いており [9]，1950 年代には敵国の政治体制内部に存在する緊張や矛盾を悪化させるために，事実，虚偽，又は両者のどちらか分からない事実と虚偽の混合物を流布していたとされる [14]．ロシアは，事実として存在する人種差別を暴露することは正当な手段として考えており，Disinformation の成功には少なくとも一部は現実と合っていなければならないとした．2014 年にはクリミア併合に関連して，ロシアの民間企業であるインターネット・リサーチ・エージェンシー社（以下，IRA 社とする）が Disinformation キャンペーンを展開していた[21].

2016 年には選挙に関連した Disinformation キャンペーンが行われ，ソーシャルメディアで蔓延する Disinformation は民主主義への脅威として捉えられるようになった．イギリスの欧州連合離脱是非を問う国民投票（通称 Brexit）では，イギリスの選挙コンサルティング会社であるケンブリッジ・アナリティカ社（以下，CA 社とする）による行動マイクロターゲッティングが問題となった．行動マイクロターゲッティングとは，個人の性格タイプに合わせた特定のメッセージや広告を作成して配信する手法である[22]．CA 社は，Facebook アプリで取得された心理測定テストの結果と「いいね」「シェア」「投稿」といった行動履歴データ等から，個人の民族性や政治的所属等を高い精度で予測する[23]サイコグラフィックスと呼ばれる手法を用いた [22]．サイコグラフィックスは，心理学のパーソナリティ分類法であるビッグファイブ理論（OCEAN とも呼ばれる）に基づき，個人の性格を判別する仕組みであった．ビックファイブ理論の 5 つの因子である「開放性」「誠実性」「外向性」「協調性」「神経症傾向」から個人の特性や行動傾向を分析し，「神経症傾向スコアが高いユーザは恐怖にかられて意思決定をする傾向がある」といった分析をしている [22]．例えば，「神経症とダークトライアド特性（ナルシズム，マキャベリズム，

及びサイコパシー傾向の総称）を持つ集団」と「平均的な市民よりも衝動的な怒り又は陰謀論に傾きやすい集団」へは，感情に火をつけて煽ることでエンゲージメントを高めるという行動マイクロターゲッティングを行っていた[24].

2016 年のアメリカ合衆国大統領選挙では, IRA 社が米社会に根付く社会問題を提起し，米国民を分極化させ，選挙を妨害することを目的にソーシャルメディアで Disinformation を流布した [21]. この Disinformation キャンペーンには, CA 社も関与していたとされる [22]. 当時 Facebook は 2015 年 6 月から 2017 年 5 月までの広告購入を調査し，「LGBT 問題，人種問題，移民，及び銃の権利等の話題を用いて社会や政治の分断を煽るメッセージを増幅することに焦点をあてているようにみえる」と報告した[25]. IRA 社はこれら複数の話題について，有権者の属性に応じてメッセージをカスタマイズしている．例えば，共和党支持の有権者には，極度の怒りと疑念を抱かせることで投票意欲を高めようとしていた[26]．投稿内容は，陰謀論，有権者による不正投票，及び選挙への違法参加を暗にほのめかし，ヒラリー・クリントンが選挙を「盗んだ」場合の反乱の必要性を述べていた．CA 社も古い価値観を引きずっている白人男性（特に高齢者）に対して，現在の社会慣行において不当に扱われている被害者であると刺激して怒りの爆発を誘導することが有効と考えていた [24]．民主党支持の有権者には，ネイティブアメリカン，LGBT，及びイスラム教徒といったコミュニティのアイデンティティとプライドに焦点をあて，ヒラリー・クリントン以外の候補者へ投票するよう呼び掛けていた．アフリカ系の有権者には，社会的疎外感又は警察官による暴行をテーマにした投稿を選挙直前まで続け，選挙が近づくにつれて選挙をボイコットする又は誤った投票手続きに従うよう呼び掛けていた．メキシコ系・ヒスパニック系の有権者には, 米国制度に不信感を抱くように誘導していた [21]. これらの Disinformation は，米国民の社会不安と人間の認知バイアスを利用し，現実と虚構の境界線を曖昧にし，民主主義そのものへの信頼を損なわせるために作られていた [26]. 欧州委員会は，「Disinformation は制度，デジタル及び伝統的メディアに対する信頼を損ない，市民が十分な情報を得た上で意思決定を行う能力を阻害することで，民主主義に害を及ぼす」[11] と警鐘を鳴らしている．

さらに，近年の Disinformation キャンペーンは選挙や国民投票に留まらない．中国は選挙に直接干渉するのではなく，アメリカの国際的な地位を弱体化させるために新型コロナウイルスに関連する Disinformation（新型コロナの発生源はアメリカとする説，新型コロナは生物兵器である等）や陰謀論を拡散したと指摘されている[27]．Disinformation

は欺瞞的なコンテンツに最も影響を受けやすい特定のオーディエンスにリーチすることで，分極化やシニシズムを高めることを目的としている[28]．過去の Disinformation キャンペーンで最も拡散したのは人々の感情に訴えかける Disinformation であり [4]，その中でも怒りは拡散しやすいという傾向がある[29]．社会的比較によって引き起こされる怒りは，報復的な情報や解決策の選好を促進し[30]，価値を追求する能力を麻痺させるため人々の最善の行動につながらない可能性がある[31]．怒りを引き起こす Disinformation がソーシャルメディアで蔓延することは，過激な思想や活動を助長し，国民の健康，環境，安全の保護等の公共に害を及ぼす [12] 可能性が高まる．この Disinformation による国民への長期的な影響について，欧州評議会は兼ねてより懸念を示している [4]．

1.3. 問題提起

Disinformation に対処するための包括的なアプローチが各国で推進されている[32]が，多くの場合において対処対象としているのは，虚偽又は違法性がある場合に限定されている．その理由は，Disinformation は一部ないしは全てが真実であるケースが存在するという複雑性に起因している．Disinformation への対処にあたっては大きく 3 つの問題が存在している．

第一に，Disinformation には法的対処における限界がある．Disinformation は必ずしも虚偽ではなく，事実や一個人の見解等，様々なケースが存在するため，概念を法的な枠組みに落とし込むことが難しい[33]．例え有害であっても合法的なコンテンツは表現の自由によって保護され，削除が正当化され得る違法コンテンツとは異なる対処が必要となる [12]．実際の Disinformation キャンペーンでは，政治的価値観，民族性，人種，及び宗教観といった属性が標的にされた [4]．この属性間の対立を先鋭化する方法として，自分と同じ属性を持つグループが不当に扱われていると「被害を訴える Disinformation」が用いられている．つまり，対立グループを攻撃するヘイトスピーチに該当するような違法性が必ずしも含まれているわけではない．日本においても，刑罰，民事法，又は行政処分といった既存の法制度の構成要件に Disinformation により生じた事態が該当し得る場合には対応が可能[34]だが，必ずしも虚偽を含まない Disinformation は具体的なケースを想定しつつ，検討が必要であると述べるに留まっている[35]．

第二に，Disinformation には技術的な検出の困難さがある．Disinformation を拡散する主体には，ソーシャルメディアユーザとソーシャル Bot の 2 種類がある[36]．ソーシャル

Bot の投稿コンテンツは，Bot 同士よりも，人間のユーザアカウントが共有する割合が高い[37]．また，ソーシャル Bot はコミュニティ内でのネガティブな感情の増幅に影響を及ぼし，このネガティブな感情が増幅する過程においてエコーチェンバーが形成されやすい傾向がある[38]．このようにソーシャル Bot も Disinformation の拡散を生み出す要因の 1 つだが，ソーシャル Bot は検出ツール（例えば，Botometer[39]）による高い精度での判別が可能であり，摘発により停止措置がなされることがある[40]．一方，ユーザによる Disinformation の拡散に対しては，情報システム技術又はファクトチェック団体等の人による検出がある．情報システム技術による検出では，過去画像の流用[41]，生成 AI で作られた画像・動画[42]，既存の Misinformation に類似したコンテンツ[43]，ChatGPT で生成されたテキスト[44,45]等については検出するためのツールが存在する．しかし，これらツールの主な検出対象は，有害コンテンツ又は虚偽であることが検証可能なコンテンツに留まる．その理由の 1 つに，プラットフォーム事業者は「真実の裁定者」になるべきではないという事業者側の考え方がある[46,47]．このため，X では認定されたユーザが誤解を招く可能性がある投稿に注釈を書き込む「コミュニティノート」を導入した[48]が，注釈を提供する認定ユーザの党派性（支持する特定の党派への偏り）が評価に影響を及ぼしていることが指摘されている[49]．

第三に，Disinformation は人による精査が難しいものが多い．多くのソーシャルメディアユーザは正確ではないと思うコンテンツを共有したくないと考えており[50]，そのように判断されたコンテンツの共有は減少する[51]．このため，ユーザ向けの Disinformation 対策として，真偽を見分ける能力を高めるメディア情報リテラシー教育が推進されていることが多い．しかし，ユーザが Disinformation を共有する要因は，知識やスキルの不足によるものだけではない[52]．いくつかある共有を促進する要因のうち，感情による影響が大きい可能性が示唆されている[53,54]．過去事例において最も拡散した Disinformation は怒り等の感情に影響を与えることを狙ったものであり [4]，人は情報の真理値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[55]．Disinformation の怒りがその真偽に関わらず共有行動を促進するのであれば，メディア情報リテラシー教育によって真偽を見分ける知識やスキルを向上させても，ユーザは怒りのままに Disinformation を共有している可能性がある．

このように，Disinformation に対する真偽を起点とした対策には限界がある可能性がある．このため，Disinformation の共有を減らすには，その特徴である怒り等の感情に

着目することが有用である可能性がある．Disinformation は法的対処や技術的な検出が困難であるため，ソーシャルメディア上の Disinformation の拡散を抑制するためにはソーシャルメディアユーザの能力に頼らざるを得ない．しかし，メディア情報リテラシー教育によりユーザの知識やスキルを向上させても，Disinformation が悪用する怒りが真偽に関わらず共有を促進している可能性がある．Disinformation には怒りを生み出す要因があるために共有が促進されているのであれば，ユーザに共有を促す「怒り」に着目したユーザ向け対策が Disinformation の感情的な共有を減らすのに有用な可能性がある．これは，従来の真偽に着目した Disinformation 対策とは異なる視点からの取組みであり，現対策と併用することで対処が難しい Disinformation の影響を低減することに役立つことが考えられる．

1.4. 本論文の構成

本論文は 8 章から構成される．2 章では，現在実施又は検討されている Disinformation 対策について調査する．3 章では，関連研究を調査することで怒りが共有を促進するメカニズムを明らかにし，2 章の現対策が怒りによる Disinformation の共有を十分に解決し得るか，またどのようなユーザ介入策が Disinformation の怒りによる共有に有効な可能性があるか考察する．4 章では，2 章における現対策の調査及び 3 章における関連研究の調査結果をもとに現状の課題を示し，それらを解決するための目的を示す．5 章では，Disinformation の怒りを生み出す要因が共有に及ぼす影響を明らかにするための実験を行う．6 章では，3 章での関連研究の調査結果をもとに Disinformation の怒りに着目することでユーザの共有行動を減らす有効策を作成し，その効果について共有行動に介入する既存の対策と比較評価する実験を行う．7 章では，6 章で作成した怒りに着目して共有を減らす有効策が Disinformation の共有を減らす効果について，既存の教育と比較評価する実験を行う．8 章では，これまでに得られた調査及び実験の結果を総括し，本研究の貢献，限界，及び提言と今後の課題について述べる．

2. Disinformation 対策の調査

本章では，これまでに実施又は検討されている Disinformation 対策について広く調査し, その概観を述べる. 特に,「1.3.問題提起」において, ユーザに共有を促す Disinformation の怒りに対しては，従来の真偽を起点とした対策では限界がある可能性が考えられた. このため, Disinformation 対策を広く調査する中で, Disinformation の怒りを含む感情的側面に言及しているものがある場合，その対策の詳細についても述べる.

調査結果は，対策の実施主体ごとに整理して述べる．Disinformation 対策には様々な種類があるため，Kozyreva ら（2020）[56]，Roozenbeek ら[57]，及び Kozyreva ら（2024）[58]において推奨策として列挙された Disinformation/Misinformation 対策を調査した．これらの対策を実施主体ごとに整理した上で，その概観を説明する.

2.1. 政府機関における取組み

政府機関における Disinformation，Misinformation，及び Malinformation 対策の方針を調査した．これは，各政府機関によって Disinformation 対策の指針や法規制による対処の方針が異なることが考えられるからである．例えば，誤りであることが証明可能な Misinformation と有害な Malinformation は，それによって生じた事象に該当する現行法が適用されることが多い．これに対し，Disinformation は有害でも違法性のないコンテンツが存在する．そこで，各政府機関における Disinformation 対策の方針及び法規制の状況について調査し，併せて Disinformation，Misinformation，及び Malinformation をどのように位置づけて対処しているのか確認した.

調査対象とする政府機関は,「法規制の類型」を参考に選択した．Disinformation に関する法規制は大別して 3 つの類型, ①プラットフォーム事業者の規制型（EU, イギリス, ドイツ，フランス等），②外国勢力の介入に対する事後制裁型（アメリカ，台湾等），③虚偽情報全般規制型（シンガポール，ロシア，中国等）に分けられる[59,60]．類型ごとに調査対象を選択することとし，類型①と②は「1.2.背景」の Disinformation キャンペーンを受けて早期に取組みを開始した①に該当する EU，イギリス，②に該当するアメリカとした．類型③は本研究が民主主義に寄与することを目指していることから，③の類型の中で民主主義指数[61]が高いシンガポールを調査対象とした．これらに加え，日本の Disinformation 対策の現状についても調査した.

2.1.1. EU

EU において Disinformation 対策の取組みを推進しているのは，欧州委員会内の通信ネットワーク・コンテンツ・技術総局（DG CONNECT）である[62]．これまで同局は，2010 年に EU の成長戦略である「Europe 2020」，2015 年に EU 加盟国間のデジタル市場を統一する「デジタル単一市場戦略」，及び市民のデジタルスキルやメディアリテラシーを向上させる「Media Literacy for All」「Creative Europe」を推進してきた．また，研究・イノベーション総局（DG RTD）と連携し，新しい研究開発を支援するプログラム「Horizon 2020」「Horizon Europe」の枠組みに Disinformation 対策技術の開発を編入する等をして取組みを推進してきた．

(1) Disinformation 対策の方針

2017 年 11 月，欧州委員会はフェイクニュースやオンラインで広がる Disinformation に対抗するための政策イニシアティブについての助言を得るために，ハイレベル専門家グループ（HLEG）を設立した．HLEG は，2018 年 1 月の初回会合において Disinformation に関連する問題とそれに対処するための方法について討議し，3 月に Disinformation 対策の提言を含む報告書を公開した[11]．この報告書内の提言をもとに，2018 年 4 月に欧州委員会は Disinformation 対策の取組みをまとめた政策文書「オンライン Disinformation への取組み：欧州のアプローチ（COM(2018) 236 final）」[12] を公開した．以降，Disinformation 対策に関連した政策文書として，取組み（Approach），行動計画（Action Plan），及び実務規範（Code of Practice）が定期的に公開されている．これらに記載されている主な取組みを総括すると，①プラットフォーム事業者によるユーザ保護の要請（Code of Practice の策定と定期的な報告），②ファクトチェック，集合知，検出能力の強化，③オンラインにおける説明責任の強化，④新技術の活用，⑤安全かつ強靭な選挙プロセスの構築，⑥リテラシー教育による育成，⑦質の高いジャーナリズムへの支援，⑧EU 加盟国間での戦略的コミュニケーションの計 8 つに整理できる．

(2) 規制における Disinformation の位置づけ

2018 年公開の「COM(2018) 236 final」では，EU における Disinformation の用語を定義した上で，報道の誤り，風刺，パロディ，明確に識別可能な党派的なニュースやコメン

ト，又は違法なコンテンツとは区別するとした．また，有害性があっても違法性のないコンテンツについては，「一般的に表現の自由によって保護され，コンテンツ自体の削除が正当化され得る違法なコンテンツとは異なる方法で対処する必要がある」と述べられている．

2024 年 2 月に施行された「Digital Services Act（DSA）」では，プラットフォーム事業者に対して違法コンテンツ（例えば，ヘイトスピーチ，テロリストのコンテンツ，差別的コンテンツ等）の対応が義務化された．違法コンテンツの定義は，第 3 条（h）において欧州連合法又はそれに準拠する加盟国の法律に反する情報とされている[63]．違法ではないが有害なコンテンツに対しては，「違法でない限り，有害なコンテンツは違法コンテンツと同じように扱われるべきではない」とし，「表現の自由を完全に尊重しつつ，違法コンテンツの削除又は削除の促進措置のみを課す」とされている[64]．また，DSA の前文（104）において，「Disinformation，操作的な悪用行為，又は未成年者への悪影響等のシステミックリスクが社会と民主主義に及ぼし得る負の影響も考慮すべき」とし，プラットフォーム事業者に行動規範（Code of Conduct）の遵守と実践を求めている．それまでプラットフォーム事業者に対するユーザ保護の要請は「Code of Practice」で実践されてきたが，2025 年 2 月に DSA の枠組みに統合されることが承認され，2025 年 7 月から施行予定である[65]．

EU 外からの Disinformation への対処については「East StratCom Task Force」による「EUvsDisinfo」に加え，外国干渉に関する特別委員会（ING2）にて推進されている．2020 年 12 月に公開された「欧州民主化行動計画（COM(2020) 790 final）」において，外国からの干渉に対抗するために，Disinformation の「加害者にコストを課すことを可能にする新たな手段の導入を提案する」とした[66]．2022 年 3 月には欧州議会が外国干渉に関する特別委員会（ING2）を設置した．EU のレジリエンスを強化し，外国からの干渉に対抗するための協調戦略を策定することを目的として活動している[67]．

2.1.2. イギリス

イギリスにおいて Disinformation 対策の取組みを推進しているのは，デジタル・文化・メディア・スポーツ省（DCMS）（現在の科学・イノベーション・技術省（DSIT））である[68]．2016 年 6月のイギリスの欧州連合離脱是非を問う国民投票において，EU 離脱を支持する組織等から Disinformation が拡散されて国民投票の結果に影響を与えた可能性が

あるとして[69]，2017 年 1 月に英国議会下院の超党派議員で構成される下院 DCMS 特別委員会が調査を開始した[70]．2018 年 1 月テリーザ・メイ首相により，国家主体等による Disinformation に対抗するための国家安全保障通信専門部門が内閣府に，英国内の Disinformation 対策の主管部門が DCMS のデジタル・技術政策局内に設置された．その後，DCMS から 2023 年に新設された DSIT に移管された．

(1) Disinformation 対策の方針

2018 年 7 月に DCMS が公開した中間報告書では，フェイクニュースという用語ではなく「Disinformation」及び「Misinformation」という言葉を定義して用いるよう提言する等，計 53 個の政府への提言がまとめられた．これらの提言に対する政府側からの暫定的な回答を経て，最終報告書においてイギリスにおける Disinformation の定義が初めて記載された [13]．最終的な提言は 51 個にまとめてられており，大別すると①義務的な倫理規範の策定，②独立した規制機関の確立と監督，③テック企業の責任の強化と罰金，④選挙法の見直し，⑤デジタルリテラシー向上の計 5 つに整理できる．これらの提言は，2019 年 4 月に DCMS と内務省が共同で発行した「Online Harms White Paper」[71]に反映された．

2017 年 10 月に DCMS 及び DCMS 国務長官が公開した「Internet Safety Strategy Green Paper（インターネット安全戦略)」[72]はデジタル憲章を掲げたものであったが[73]，2019 年 4 月に DCMS と内務省が公開した「Online Harms White Paper」には Disinformation 対策が盛り込まれた．Disinformation の対処に関するプラットフォーム事業者の自主的な取組みに対し，首尾一貫した新たな規制の枠組みが必要であるとして，新しい法的な注意義務とその義務を果たすための実務規範（Code of Practice）を設ける予定であることが記載された．その他，2019 年 4 月に政府コミュニケーションサービス（GCS）が「RESIST Counter Disinformation Toolkit」（2021 年に RESIST 2）[74]を公開し，組織が Disinformation に対処するためのフレームワークが提供されている．

(2) 規制における Disinformation の位置づけ

2019 年の「Online Harms White Paper」において，規制の範囲とする有害コンテンツは「有害の定義が明確であるもの」と「有害の定義が明確ではないもの」に大別されている [71]．「有害の定義が明確であるもの」としては，児童の性的搾取と虐待，テロ関連コンテンツ，組織的な移民犯罪，現代的な奴隷制度，過激なポルノ，リベンジポルノ，嫌が

らせとサイバーストーカー行為，ヘイトクライム，自殺の助長又は支援，暴力の煽動，違法な商品/サービスの販売等が挙げられている．一方で，「有害の定義が明確ではないもの」として，Disinformation やトロール行為が挙げられている．これらは，違法なコンテンツと並んで，合法ではあるが潜在的に有害なオンラインコンテンツと位置づけられており，「刑事犯罪には至らない場合でも，有害で破壊的な影響を与え，有害なオンライン環境を作り出し，ユーザがオンラインで自己表現する能力に悪影響を及ぼす可能性がある」としている[75]．このように個人に重大な害を与える可能性のある Disinformation 及び Misinformation は，注意義務の対象とされている．

規制は，電気通信・放送等の規制機関である英国情報通信庁（Ofcom）が担当しており，Ofcom による規制の枠組みを規定した「Online Safety Act（オンライン安全法）」が 2023 年 10 月に制定された[76]．同法は，違法なコンテンツ及び子どもに有害なコンテンツによるリスクの特定・軽減・管理する義務をプラットフォーム事業者に課すものであり，義務に違反している場合には制裁金が課される可能性がある[77]．法案の時点では，合法だが有害なコンテンツとして Disinformation も規制対象として想定されていたが，審議過程において削除され，規制対象とする優先犯罪として外国干渉に基づく犯罪（「National Security Act（国家安全保障法）」第 13 条）を含むことで包含されることとなった[78]．

国外からの Disinformation への対処については，2023 年 7 月に制定された「National Security Act」に含まれている[79]．外国干渉罪の主な目的は，イギリスの利益や権利等を損なう外国からの敵対的な活動を抑止し，国家の安全と利益を保護することである[80]．干渉活動は必ずしも敵対的なものではないとしつつ，一方で透明性を持って行われないものは外交の規範外であるとし，隠蔽的で悪意ある政治的干渉活動として国家支援型の Disinformation の問題が指摘されている．

2.1.3. アメリカ

アメリカでは，合衆国憲法修正第 1 条における言論の自由のもと，Disinformation 対策としての一元的な取組みは推進されていない．また，2025 年 1 月 20 日の第二次トランプ政権発足日にホワイトハウスから「言論の自由を回復し，連邦政府の検閲を終わらせる」という大統領令が出され[81]，Disinformation 対策に関連する政府の取組みが廃止され始めている[82]．

(1) Disinformation 対策の方針

国内向けに行政機関が推進している一元的な Disinformation 対策はなく，プラットフォーム事業者による独自の取組み，ファクトチェック団体による情報提供，州レベルで推進されているメディア情報リテラシー教育等がある．

安全保障に関連する Disinformation 対策の取組みについては，2017 年 2 月に同盟国内の Disinformation キャンペーンを追跡し，戦術を分析し，対抗するための機関として国務省内に「Global Engagement Center（GEC）」が設立されたが，2024 年 12 月に閉鎖されることとなった[83]．2018 年 5 月に国土安全保障省（DHS）が設立した「対外影響対策タスクフォース（CFITF）」も同様の状況にある．CFITF は外国の干渉や情報活動（Disinformation 等）に関するレジリエンスを構築するために設立され，その後サイバーセキュリティ・社会基盤安全保障庁（CISA）に移管されたが[84]，2023 年に憲法修正第 1 条の権利を侵害しているとする訴訟を受けて活動は抑制された[85]．また，2022 年 4 月に発表された国土安全保障省の Disinformation ガバナンス理事会も批判を受けて 3 週間で活動を停止した[86]．

(2) 規制における Disinformation の位置づけ

国内に関しては，「アメリカ人による政治的に分断的なコンテンツ又はあからさまな Disinformation の拡散は，憲法修正一条により憲法で保護された言論の自由として認められている」[87]．憲法修正第 1 条における言論の自由により，「その言論が「虚偽」であることのみを理由として規制立法を展開するということは憲法上，許されない可能性が高い」[88]．このため，「基本的に連邦レベルでの Disinformation 対策の法規制は存在していない」[89]とされる．

2020 年 11 月に当時の大統領であるトランプ氏が，民主党が選挙を盗もうとしている等の投稿を Twitter にしたところ，誤解を招く恐れがあるという理由で警告ラベルが付与された[90]．このような Twitter の措置に対してトランプ氏は，「オンラインの検閲の防止に係る大統領令 13925 号」に署名し，言論の自由を確保するためにプラットフォーム事業者による恣意的なユーザ投稿の削除等を限定するよう求めた[91]．これまでプラットフォーム事業者は，憲法修正第 1 条の言論の自由が重視される背景から，通信品位法 230 条で広範な免責が認められてきた[92]．このため，以降司法省による勧告や議員らによる改正法案がいくつか提出されたが，現時点でいずれも成立していない．

一方で，州レベルでは保守派の州においてプラットフォーム事業者によるコンテンツ削除等のコンテンツモデレーションを規制する法律が制定されている（フロリダ州法 SB7072，テキサス州法 HB20）．また，選挙運動におけるディープフェイク等を規制する法律がテキサス州（SB751）及びカリフォルニア州（AB730）で可決されている．

2.1.4. シンガポール

シンガポールにおいて Disinformation 対策の取組みを推進しているのは，デジタル開発・情報省（MDDI）の傘下機関である情報通信メディア開発庁（IMDA）である．2017 年 4 月，シャンムガム法務・内務大臣が，国会においてフェイクニュース対策の必要性について述べた[93]．2018 年 1 月に情報通信省と法務省が Green Paper を公開し，提言としてオンライン虚偽情報（Falsehoods）の意図的な拡散という課題に対して広範な議論を行うための特別委員会を設置するよう求めた[94]．同月，「意図的なオンライン虚偽情報に関する特別委員会」が設置され，特別委員会は 9 月に立法を含む必要な措置を実施すべき等の提言を含む調査報告書を公開した[95]．2019 年 4 月に「オンライン虚偽情報及び情報操作防止法案」が議会に提出され，5 月に可決，10 月より施行された．本法案の制定以後，IMDA の所掌下にある POFMA Office[96]にて虚偽情報の監視，通報処理，及び事例公開等が行われている．

(1) Disinformation 対策の方針

2018 年 1 月に公開された「意図的なオンライン虚偽情報：課題と影響」と題した Green Paper では，7 か国における意図的なオンライン虚偽情報の拡散事例について調査した結果と共に，外国からの政治への干渉を禁止する既存の規制である政治献金法（Political Donations Act），結社法（Societies Act），及び公共秩序法（Public Order Act）と同様の規則をサイバー空間にも適用すべきとする提言が述べられた [94]．また，虚偽情報の拡散を防止・対抗するための対応原則と立法を含む具体的な措置を検討するため，特別委員会を設置するよう求めた．

特別委員会は，同年 9 月に「オンライン上の意図的な虚偽情報に関する特別委員会報告書」を公開し，意図的な虚偽情報の拡散を阻止するための 22 の勧告を提言した [95]．その内容を大別すると，①国家レベルの戦略と協調的なアプローチの策定，②政府の対応権限と立法措置，③政府及び公共機関による支援・研究・情報提供，④プラットフォー

ム事業者による報告と監査，⑤ファクトチェック団体との連携，⑥教育の取組み，⑦ジャーナリストのスキル向上と専門基準の順守の計 7 つに整理できる．このうち，特に推進されているのは②の立法措置に関する整備であり，2019 年 10 月に「オンライン虚偽情報及び情報操作防止法（Protection from Online Falsehoods and Manipulation Act: POFMA）」[97]，2021 年 10 月に「外国干渉防止法（Foreign Interference (Countermeasures) Act: FICA）」[98]，2022 年 11 月に「オンライン安全法（Online Safety (Miscellaneous Amendments) Act）」[99]が制定された．

(2) 規制における Disinformation の位置づけ

オンライン虚偽情報及び情報操作防止法（POFMA）が制定される以前は「Falsehoods」「Fake」「Disinformation」といった用語が使用されており，その定義や使い分けについては明らかではなかった．POFMA において「虚偽の事実言明（False statements of fact）」に表現が統一され，規制対象として①虚偽情報，②国の治安を害する可能性がある情報，③公衆衛生，公安，治安，財政を害する可能性がある情報，④他国との友好関係を害する情報，⑤選挙や国民投票の結果に影響を与える情報，⑥集団間の敵意，憎悪，又は悪意の感情を煽ること，⑦政府の職務遂行に対する国民の信頼を低下させる情報が該当するとした（第 7 条）．これらの情報が，マルチメディアメッセージングサービス（ソーシャルメディア等）だけでなく，ショートメッセージサービス（iPhone の iMessage 等）といったクローズドなプラットフォームで伝送された場合も対象となる（第 3 条）．「虚偽の事実言明」として規制対象となる情報を投稿又は拡散した場合，訂正指示がなされ，当該指示に従わない場合に罰金又は禁固刑に処される．実際にどのような情報が「虚偽の事実言明」に該当するかの判断は大臣に委ねられている（第 10 条）ことから，定義が不明確であり恣意的な判断が可能であるといった批判が生じている．

オンライン安全法では，ユーザの安全なオンライン利用を保護するために，悪質なコンテンツ（Egregious content）を規制している．悪質なコンテンツに該当するものとして，自殺，暴力，性的暴力，児童ポルノ，国の公衆衛生措置を妨害・リスクをもたらす行為，人種又は宗教グループに対する敵意等を引き起こすコンテンツ，テロリズムを擁護・指導するコンテンツ等が明示されている（第 45 条 D 項）．

国外からの Disinformation への対処については，外国干渉防止法（FICA）に規制対象と対抗措置が記載されている．FICA は，外国による国内政治への干渉を防止・検知・妨

害する能力を強化し，公共の利益を保護することを目的としている．FICA では明言されていないものの，具体的には①オンライン上の敵対的情報キャンペーン（Hostile Information Campaigns：HICs）と②国内の代理人を活用するオフラインの影響工作に対処することを指す[100]．外国の主権者又はその代理人として行動する者が，政治的議論を操作して社会を混乱させることを目的に POFMA 第 7 条で規制対象とされている情報を用いた活動を行うことに対し，内務大臣の指示により調査と抑制措置が講じられる．措置には当該情報の削除，アクセス無効化等が含まれ，調査の結果虚偽又は誤解を招く情報であった場合罰金又は懲役が科される．なお，外国主体の代理人である場合を除き，シンガポール人が政治問題に関する自身の見解を表明する場合には適用されない[100]．

2.1.5. 日本

日本において Disinformation 対策の取組みを推進しているのは，総務省である[101]．2019 年 4 月に総務省が主催する「プラットフォームサービスに関する研究会」が中間報告書を公開し，その中で Disinformation に関して具体的な施策の方向性の検討に向けた整理が必要である旨が述べられた[102]．2020 年 2 月の最終報告書で Disinformation 対策に関するフォーラム設置の必要性が述べられ[103]，2020 年 6 月に Disinformation 対策フォーラムが設置された[104]．その後，2023 年 11 月に総務省が主催する「デジタル空間における情報流通の健全性確保の在り方に関する検討会」が開催され，Disinformation に焦点をあてた包括的な対策の議論がなされた．

国家安全保障の観点においては，2022 年 12 月に国家安全保障戦略が発表され，能動的サイバー防御と共に，偽情報の拡散は安全保障上の脅威であるとして対応の強化が述べられた[105]．これまで日本におけるサイバーセキュリティ対策の推進は，2000 年に内閣官房が設置した内閣官房情報セキュリティ対策推進室を改組した，内閣官房情報セキュリティセンター（NISC）が担ってきた．NISC では，サイバーセキュリティに関連する企画立案，情報収集，関係機関との調整，行政機関のセキュリティ監視等を行ってきたが，更に安全保障分野の政策をも一元的に総合調整するための新たな組織として改組されることとなった[106]．2025 年 5 月に「能動的サイバー防御」を導入するための法律が成立したことに伴い，7 月に NISC を改組した国家サイバー統括室（NCO）が発足した[107]．NCO は能動的サイバー防御（サイバー攻撃に該当するアクセス無害化等）についての対処を検討・決定し，その実施主体として警察と自衛隊が連携して対応（アクセス無害化措置

の執行等）する[108]．2025 年 1 月には警察と自衛隊の合同拠点が新設される等[109]，海外からのサイバー空間を経由した安全保障上の脅威に対抗する仕組みが整備されてきている．2025 年 7 月に行われた第 27 回参議院議員通常選挙では，他国による選挙介入が行われたとの指摘や報告があったとして，NCO を中心に海外事例の情報収集・分析を行い，偽情報に関する規制又は新法の制定等の対策を検討するとしている[110]．

(1) Disinformation 対策の方針

Disinformation 対策フォーラムの中間報告書では，「公職者の発言や公的機関による発表，メディアによる報道については直接の対象としない」とした上で，対策の取組みとして「SNS での個人の投稿を主たる対象としたファクトチェックの実施（検証結果の効果的な伝達を含む）」と「SNS の利用に焦点を当てた情報リテラシー教育」の 2 つが挙げられた [16]．最終報告書に取組みとして実施されたことが記載されたが，プラットフォーム事業者が自主的に実施した取組み以外の施策については，具体化に向けた留意点や実施の方向性に関する記載に留まっていた[111]．

2024 年 9 月に総務省が公開した「デジタル空間における情報流通の健全性確保の在り方に関する検討会 とりまとめ」では，有識者からの情報提供，プラットフォーム事業者へのヒアリング，及び諸外国における動向の調査をもとに，制度的な対応 5 点と制度的な対応以外の 5 点からなる計 10 点の総合的な対策が提言された [35]．制度的な対応として，①情報伝送プラットフォーム事業者による偽・誤情報への対応，②情報伝送プラットフォームサービスが与える情報流通の健全性への影響の軽減，③マルチステークホルダーによる連携・協力の枠組みの整備，④広告の質の確保を通じた情報流通の健全性確保，及び⑤質の高いメディアへの広告配信に資する取組みを通じた健全性確保が挙げられた．また，制度的な対応以外として，⑥普及啓発・リテラシー向上，⑦人材の確保・育成，⑧社会全体へのファクトチェックの普及，⑨技術の研究開発・実証，及び⑩国際連携・協力が挙げられた．

(2) 規制における Disinformation の位置づけ

Disinformation/Misinformation を規制する法令はなく，それによって引き起こされた事態によって関連する法律が適用される [34]．これに加え，「デジタル空間における情報流通の健全性確保の在り方に関する検討会 とりまとめ」では，原則として対応を検討す

べき偽・誤情報が示され，①検証可能な誤りが含まれていること，かつ②各要素（違法性や客観的な有害性，拡散することによる社会的影響の重大性，検証の容易性）の有無・軽重に照らし，具体的な方策との関係で比例性が認められること，の両方の要件を満たすものとした[35]．このうち，「客観的な有害性」と「社会的影響の重大性」が認められ得るか，また「必ずしも誤りは含まれていないが文脈上誤解を招く（ミスリーディングな）情報」や「事実ではあるが人を害する意図を持って発信された悪意ある情報」への対応については，具体的なケースを想定しつつ更なる検討が必要であるとした．

国外からの Disinformation への対処については，日本では対応方針が定められていないが，外務省において外国からの情報操作に関して海外諸国と連携する取組みが進められている[112]．日本としてどのように対処すべきかという点に関しては，有識者から「Disinformation を用いた外国勢力の干渉に関する情報収集センターを設置し，事後制裁及び国際法上許容される対抗措置を行うことを可能にする法律の制定を検討する」[60]「国家及びその手足となるアクターには人権としての表現の自由はない」[113]といった意見もあり，表現の自由の保護領域からの議論があった．

2.2. プラットフォーム事業者における取組み

プラットフォーム事業者が，自社がサービス提供するソーシャルメディアに実装又は実装を検討している Disinformation 対策について調査した．プラットフォーム事業者は，ユーザがソーシャルメディアを利用している最中に直接介入する対策を実装することができる．しかし，プラットフォーム事業者による投稿コンテンツへの介入，すなわちコンテンツモデレーションには表現の自由等の基本的価値と公衆衛生の保護との間にトレードオフの関係がある[114]．このため，Disinformation 対策の方針は各事業者の理念やスタンスによって様々である[115]．本研究では，調査対象とするソーシャルメディアを絞り込んだ上で，調査対象の公式ブログ，ブログ執筆者の論文，及び関連するメディアニュースから，実施又は検討している Disinformation 対策を調査した．

ソーシャルメディアは Disinformation が大きな問題となる以前より，エコーチェンバー[116]，匿名性によるオンライン脱抑制効果[117]，評判や推薦等を指標に情報源の信頼性を簡便に評価する傾向[118]，及びフィルターバブル[119]等の懸念が指摘されていた．これらの問題が，ソーシャルメディアが提供するレコメンデーション機能やターゲッティング機能によって助長され，Bot や生成 AI が用いられることで増幅されるようになった．ソ

ーシャルメディア上の感情についても，Facebook 上で他者の感情にさらされる機会を少なくするとユーザの投稿数が減るという撤退効果がある[120]とし，ソーシャルメディアはユーザが他者の感情に接触する頻度とそれに伴う情動伝染を増幅する仲介者としての役割を果たしていると指摘されている [19].

プラットフォーム事業者が Disinformation 対策に取組むきっかけとなったのは，2016 年 11 月 8 日のアメリカ合衆国大統領選挙である．選挙後，落選したヒラリー・クリントン氏を支持する政治団体やマスコミが，「Facebook 等の大手ソーシャルメディアがフェイクニュースを規制しなかったから不正に負けた」という主張をはじめた[121]．これに対し，13 日に Facebook 創始者は「Facebook 上のコンテンツは 99%信頼できるもの」であり，「虚偽のニュースやでっちあげはごく一部である」と見解を示した[122]．しかし，18 日に主張を一転させ，対策に取組むことを表明した[123]．2017 年 9 月，Facebook は 2015 年 6 月～2017 年 5 月までの広告購入を調査した結果を公表し，特定の広告とアカウントが LGBT 問題，人種問題，移民，及び銃の権利等の話題によって分断的な政治メッセージを増幅していたことを報告した [25]．これを受けて，米上院下院の情報特別委員会は Facebook，Twitter，及び Google のプラットフォーム事業者に対して公聴会で証言するよう求め[124]，そこで対策が不十分であったことが指摘された[125].

また，EU においても 2016 年 6 月イギリスの欧州連合離脱是非を問う国民投票における Disinformation キャンペーンを受けて設置された EU ハイレベル専門家グループ（HLEG）は，自己規制的なアプローチとして「Code of Practice」を策定し，プラットフォーム事業者の目標とさせるよう提案した [11]．2018 年 9 月に欧州委員会は，迅速かつ効果的にユーザを Disinformation から保護するよう，Facebook，Google，Twitter，及び Mozilla の各プラットフォーム事業者に対して Code of Practice を策定してその実施状況を年次で報告するよう求めた[126].

これらの背景を踏まえ，調査対象とするソーシャルメディアは，本研究の検証において想定する X（旧 Twitter）の他，米議会及び欧州委員会の要請対象となった Facebook と Google とした.

2.2.1. X（旧 Twitter）

X の目的は，「公共の場における会話に寄与すること」であり，暴力，嫌がらせ，脅し，又は恐怖を与えて他ユーザが発言できないようにする行為を禁止している[127]．X は 2006

年 3 月に「twttr」という名称でサービスが開始され，2010 年に他ユーザの投稿コンテンツを共有するリツイート機能（以下，現在の機能名称であるリポストに統一する）が追加されてから投稿コンテンツが拡散するようになった．ルールを設けて違反コンテンツを明示してきたが，2014 年にアカウントの不正使用やハラスメントが増加して問題視されるようになった[128]ため，攻撃的な投稿コンテンツを報告するハラスメント防止ツール[129]や専門家らによる評議会「Twitter Trust & Safety Council」を設立する[130]等の対応をしてきた．X の主流な対策は，ポリシーを策定・強化し，そのポリシーに基づき違反コンテンツを検出してラベルを付与するというものである．

(1) Disinformation 関連対策の概要

Disinformation 対策としてはポリシーを強化すると共に，透明性を高めることに重点が置かれた[131]．X は透明性を高める一環として，データセットの公開[132]（2021 年 12 月公開終了）やレコメンデーションアルゴリズム等のソースコードを公開している[133]．また，透明性を確保しつつユーザが正確な情報を入手できるようにする機能として「コミュニティノート」が提供された．コミュニティノートは，2021 年 1 月にパイロット版として開始された「Birdwatch」[48] の改称である．「Birdwatch」は，誤解を招く可能性がある投稿コンテンツに協力ユーザたちが注釈（note）を書き込むことで有益な文脈を提供することを意図している．書き込まれた注釈を読んだユーザは，読んでいないユーザよりも誤解を招く可能性がある投稿コンテンツに同意する確率が 20〜40%低くなったと報告されている[134]．一方で，文脈を提供する Birdwatch メンバーは自分の支持政党と対立する党派からの投稿コンテンツに対してネガティブな評価を書く傾向が強いことが指摘されている[49]．このような透明性を高める取組みに加え，2024 年 6 月にはユーザのプラバシー保護を強化するために各投稿コンテンツの「いいね」の数が非公開となり[135]，続けてリポスト数も非公開となった．

(2) ユーザの行動に介入する取組み

X ではユーザの行動に介入する機能がいくつか実装されている．Code of Practice 署名後の 2019 年 6 月，ルール違反の投稿コンテンツに対して，閲覧者に内容が見えないよう違反通知をオーバーレイ表示する機能を導入した[136]．閲覧者は通知メッセージの「表示する」を選択することで投稿コンテンツを読むことができる．これは，ルールに違反す

る投稿コンテンツであっても，公共の利益に照らして閲覧を認めるというXのスタンスによるものである．

2020年6月，投稿コンテンツ内に記載されたニュース記事のURLをクリックして開かずにリポストしようとしたユーザに対して，当該ニュース記事の内容を確認するよう促す警告をポップアップ表示する機能を試験的に導入した[137]．その結果，警告表示を見たユーザがニュース記事を開く回数が40%増加し，記事を開くユーザ数も33%増加したと報告されている．この機能は有効であると評価されたことから，正式に実装された．

同年10月，誤解を招く情報ラベルが付与された投稿コンテンツに警告画面を表示し，リポストしようとすると引用画面がポップアップ表示される仕組みを導入した[138]．誤解を招く情報とは，公衆衛生当局等の各分野の専門家によって，誤りであること又は誤解を生じる可能性があることが確認されている発言や主張のことを指している[139]．リポストはリポストボタンを1回クリックするだけで共有されるのに対し，引用はコメントを追加してからリポストボタンをクリックすることで共有される．ユーザに必ずコメントを追加させるという摩擦（フリクション）を生じさせることによって，誤解を招く情報のリポスト数が減ると予測されていた．結果として，誤解を招く情報ラベルが付与された投稿コンテンツの引用は約29%減少したと報告されている[140]. しかし, その2ヶ月後, 引用を奨励しても，引用に追加されたコメントの45%は1つの単語しか含まれず，投稿コンテンツに対する思慮深いコメントが増加するようには見えなかったとして当該機能は廃止された．

2.2.2. Facebook

Facebookのサービスを提供するMeta Platform社は，「コミュニティづくりを応援し，人と人がより身近になる世界を実現する」ことを企業理念とし，ユーザの表現の場を提供しつつも，安全性，尊厳，信頼性，及びプライバシーとのバランスを重視している[141]．Facebookは2004年2月にサービス提供が開始されてから，様々なプライバシー懸念が指摘されてきた．例えば，2009年12月に変更された新しいプライバシー設定が適用された際，全体公開がデフォルト設定であった等の指摘がある[142]．しかし，2013年のスノーデン事件を受けて米テック企業[143]においてユーザ情報保護への取組みが推進されるようになった[144]．翌年の2014年には「プライバシーチェック」サービスを提供し[145]，2015年には利用にあたってのコミュニティ基準を定め[146]，2016年には虚偽・誇大広告の表示

優先度を下げる[147]等の対策を行ってきた.

(1) Disinformation 関連対策の概要

2017 年 4 月に公開された対策の取組みは大別すると，①虚偽ニュースの経済的なインセンティブを阻害し拡散を抑制するための製品開発をすること，及び②ユーザの意思決定を支援することだった[148]. ①拡散の抑制に関する 2024 年までの主流の対策は，外部のファクトチェック団体と連携してファクトチェックを強化し，低品質コンテンツを下位表示することであった[149]. 2019 年 4 月に Google の PageRank アルゴリズムをもとにした機能「Click-Gap」を開発し，Facebook から外部 Web サイトへ不均衡な量のトラフィックが発生しているドメインを探してニュースフィードのランキングを下げるということをしている[150]. 2020 年 11 月には AI を活用した予測検出に取組んでおり，既知の Misinformation の複製に近いものを大規模に照合する画像照合モデル「SimSearchNet++」や，複数言語の意味的な類似性をより正確に評価する「LASER cross-language sentence-level embedding」等を開発した [43]. 予測検出された投稿コンテンツはファクトチェック団体の記事と比較され，既存の Misinformation に合致しない場合はファクトチェックを依頼するといったサイクルを構築していた（図 2-1）. しかし，Facebook は 2025 年 1 月にファクトチェック団体との連携を終了し，コミュニティノートへ移行することを発表した[151]. これにより，ファクトチェックされた投稿コンテンツの下位表示を停止し，警告をオーバーレイ表示するラベルを使用するとしている. ②ユーザの意思決定を支援するため，2017 年 10 月に「コンテキストボタン」が試験的に導入された[152]. 当該機能は，ニュースフィードの記事タイトルの右上にある「i」マークのアイコンをクリックすると，ニュース記事の情報源と記事に関する背景情報が提供されるというものである. 当該機能は 2018 年 4 月に正式に導入され[153]，2020 年には新型コロナウイルスに関連する投稿コンテンツで 90 日以上経過した記事をシェアしようとすると警告画面がポップアップ表示される仕組みが追加された[154].

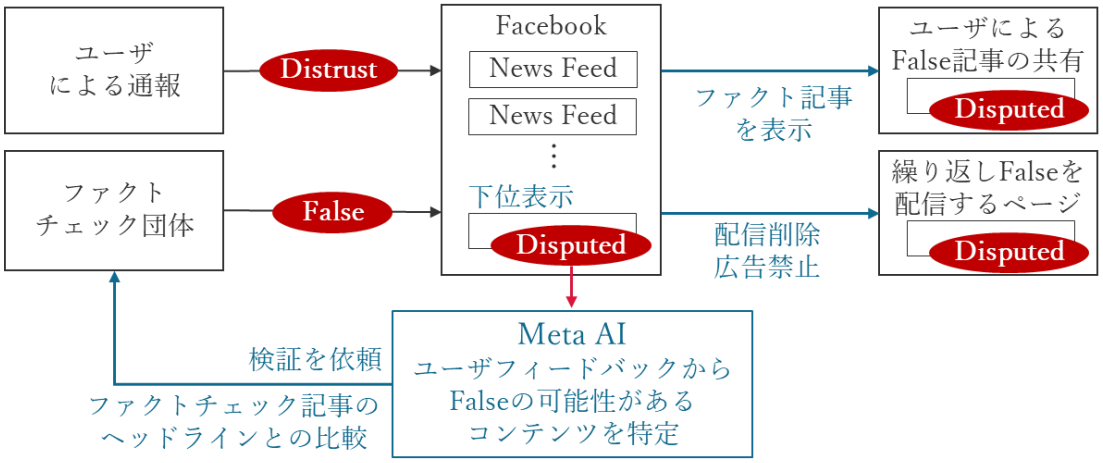


図 2-1 Facebook のファクトチェックサイクル

(2) ユーザの行動に介入する取組み

アメリカ合衆国大統領選挙の翌月である 2016 年 12 月より外部のファクトチェック団体と連携しており，ファクトチェック団体から提供された情報に基づいてユーザに対して警告ラベルをポップアップ表示している[155]．警告ラベルが表示されるタイミングは，虚偽の疑いがある投稿コンテンツに対してユーザがシェアボタンを押して共有しようとした時，ユーザが虚偽又は加工された写真を表示しようとした時，及び虚偽情報を繰り返し投稿しているアカウントをユーザがフォローしようとした時等がある．

2.2.3. Google

Google が掲げる使命は「世界中の情報を整理し，普遍的にアクセス可能で有用なものにすること」であり[156]，検索サービスでは最も関連性が高い信頼できる情報を提供することを目的としている[157]．Google は 1998 年 9 月にサービス提供が開始されてから，2010 年に社内シンクタンク「Google Ideas」（現 Jigsaw）[158]の設立，2015 年に報道支援チーム「Google News Lab」の設立[159]，非営利団体のファクトチェック団体「First Draft」の設立[160]を支援してきた．2016 年 10 月，翌月のアメリカ合衆国大統領選挙に向けて，Google ニュースの記事に Fact Check タグとファクトチェック記事の URL をラベル表示する仕組みを導入した[161]．また，2017 年 2 月にコメント機能があるアプリケーションを公開している運営者向けに，コメントの有害判定をするツール「Perspective API」の提供を開始した[162]．これは，ユーザからのコメントを機械学習モデルにより 6 つの属性（毒性等）

ごとに 0-1 にスコア化して評価し，設定した閾値（例えば 0.8）を超えた場合にアラートをあげるといった有害コメントの検出に利用することができる[163].

(1) Disinformation 関連対策の概要

2017 年 10 月，2016 年アメリカ合衆国大統領選挙における Disinformation キャンペーンの調査結果と今後の対策方針を公開した[164]. さらに, 2019 年 2 月に「Google White Paper」を公開し，①ランキングシステムの表示において品質を重視する，②悪意のあるアクターに対抗する，③ユーザにコンテキストを提供するという 3 つの戦略的枠組みに基づいて Disinformation 対策の取組みを実行するとした[165]．①ランキングシステムの表示に関する取組みとして，2022 年 8 月に信頼できる複数の情報源（専門知識，権威，信頼性を示す情報源）から関連情報を表示するように設計された「Multitask Unified Model (MUM)」を開発した[166]．これにより，高品質な情報がランキングで優先表示されるようになったと報告している．②悪意のあるアクターについては脅威分析グループが定期的に脅威情報を公表しており[167], 2022 年 9 月には脅威インテリジェンス企業大手である Mandiant 社を買収して影響工作への対処を強化した[168]．③コンテキストの提供としては，検索結果において追加のコンテキストを提供する機能が導入された．2020 年 6 月に Google で画像検索した際に，ファクトチェックラベルが表示されるようにした[169]．2021 年 2 月には「この結果について（About this result）」[170]，2023 年 3 月には「この著者について（About this author）」[171]，2023 年 5 月には「この画像について（About this image）」[172]というコンテキストを提供する機能が追加されており，Google の検索結果画面の右上に表示される縦三点リーダー（⋮）をクリックすることでコンテキストを見ることができる.

その他，AI に関する取組みが推進されており，2023 年 8 月に AI が生成した画像に電子透かしを埋め込む「SynthID」[173]及び 2025 年 5 月には画像に SynthID が含まれているかを検証する「SynthID Detector」[174]が公開された.

(2) ユーザの行動に介入する取組み

Google が提供している主なサービスは検索サービスであることから，ユーザの行動に対して介入する取組みはなかった.

2.3. ファクトチェック団体における取組み

ファクトチェック団体が，どのような情報を検証対象としているか調査した．歴史的な背景として，新聞等の伝統的なメディアにおいて正確な情報を提供するための真偽の検証，すなわちファクトチェックが行われてきた．しかし，インターネットの普及に伴い Web メディアやソーシャルメディアが増加し，伝統的なメディア以外においてもファクトチェックの重要性が求められるようになった．本研究では，主にオンラインメディアの情報を対象に真偽の検証をしているファクトチェック団体を調査し，どのような情報がファクトチェック対象となり得るのかを確認した．

現在ファクトチェックの事実上の世界標準となっているのは IFCN（International Fact-Checking Network）である．IFCN は 2015 年にアメリカのジャーナリズム研究機関である Poynter 社が設立した．世界最大のファクトチェック団体の連合組織として機能しており，IFCN の認証を得ることで世界的なファクトチェック基準を満たしていることを示すことができる．IFCN の基準は 5 つの原則（Code of Principles）としてまとめられており，①非党派性と公正性，②情報源の基準と透明性，③資金源と組織の透明性，④検証方法の基準と透明性，⑤オープンで誠実な訂正方針が求められる[175]．現在，442 のファクトチェック団体が活動しており[176]，そのうち 189 団体が IFCN の認証団体である[177]．日本のファクトチェック団体で認証を取得しているのは「InFact」「日本ファクトチェックセンター（JFC）」「リトマス」の 3 団体である．

IFCN は原則に基づく認証とトレーニングを提供しているだけであり，ファクトチェックのための詳細な手順やルールを定めているわけではない．このため，IFCN 認証を取得しているファクトチェック団体において，どのような基準で真偽の検証を行っているのかを調査した．調査対象は，IFCN 認証団体のうち，1994 年と最も古いオンラインメディアのファクトチェック団体である「Snopes」，IFCN 認証を提供している Poynter 社のファクトチェック団体である「PolitiFact」，及び日本の Disinformation 対策フォーラムの報告書を受けて設立された「日本ファクトチェックセンター」とした．

(1) Snopes

1994 年にアメリカ在住のマイケルソン夫妻が，都市伝説，デマ，民間伝承に関する真偽について検証し，その結果を公開するサイトとして「Snopes.com」を開設した[178,179]．広告収入から得た資金を元手に会社を設立し，現在に至るまで独立系のファクトチェッ

ク団体として活動している．真偽の検証結果には，20 種類のラベル（調査中，真実，ほとんど真実，混合，ほとんどが虚偽，虚偽，未確認，根拠なし，詐欺，風刺，都市伝説等）が付与される[180]．これは，複雑なトピックや出来事の真実性を 1 つの単語で評価することは困難であるという考えに基づく．

(2) PolitiFact

2007 年，フロリダ州の新聞「Tampa Bay Times」において，選挙における政治家の発言の正確性を検証するために設立された．2018 年に所有権が親会社の Poynter 社に移管され，非営利団体である Poynter Institute for Media Studies の所属となった[181]．広告や助成金，寄付金等により運営されているが，匿名の情報源，特定の政党，政治関係者，又は利益相反とみなされる場合は寄付を受け付けていない．真偽の検証結果には，6 種類の評価（真実，ほとんど真実，半分真実，ほとんどが虚偽，虚偽，嘘）が付与される．これらの判定は，当該情報の検証者と編集者で確認した後，編集者 2 名が追加され，基本的には編集者 3 名中 2 名の評価が採用されている．

(3) 日本ファクトチェックセンター

日本ファクトチェックセンターは，2022 年 10 月に一般社団法人セーファーインターネット協会が，Google の慈善事業部門である Google.org と LINE ヤフー株式会社の支援を受けて設立した[182]．これは，2020 年 6 月に同協会が設立した Disinformation 対策フォーラムの報告書を受けての取組みである．運営団体であるセーファーインターネット協会から独立した予算体系で管理されており, Google.org, LINE ヤフー株式会社, 及び Meta 社からの助成金等で運営されている[183]．真偽の検証結果は，5 種類の評価（正確，ほぼ正確，根拠不明，不正確，誤り）に分けられる．あくまでも事実を検証するものであり，意見を検証（オピニオンチェック）するわけではないことを重要な点として説明している[184]．日本ファクトチェックセンターで検証した結果については，「Yahoo！ニュース」で配信されている．

ファクトチェックにおける課題についても議論されており，検証対象とする情報の範囲，偽情報の判定の困難性（ファクトチェッカー同士でも判定が分かれることがある），及び表現の自由を過度に制約する危険性等が挙げられている[185]．また，運営実施体制や資金面においてリソースを確保することが難しく，検証可能な記事数の量的課題がある

ことが指摘されている[186].

2.4. 学術・研究機関における取組み

学術・研究機関が，研究開発及び効果検証をしている Disinformation 対策について調査した．Disinformation に関する研究は，2016 年の Disinformation キャンペーン以降，急激に増加している[187]．Disinformation 対策に関する研究は多岐にわたるため，Kozyreva をはじめとする Disinformation/Misinformation 研究において著名な研究者 30 名が推奨策として挙げている Disinformation 対策 [58] を参考に，対策としての効果が検証されている研究を中心に調査した．

2.4.1. 警告とファクトチェックラベル

警告ラベルは「特定の情報やその情報源に惑わされる可能性を明示的に警告するもの」であり，ファクトチェックラベルは「プロのファクトチェッカーが付与した信頼性評価」を指す [58]．これらのラベルが付与されたニュース記事や投稿コンテンツは，正確さが低いとユーザに認知されて共有意思を低下させる[188]．ラベルの形態は二項ラベル（真/偽）だけでなく，「誤解を招く」「論争中」といったテキストメッセージによるラベルや，真偽確率を 0（偽）から 1（真）の間で数値化してカラーバーで示すもの[189]等がある．Misinformation が各ラベル（デマ/疑わしい/事実）に分類される割合をドーナツチャートで示すことは，共有に反対する人を増加させた[190]．また，ニュース記事の論争スコア（同意する/同意しない）と共に，感情スコアを-1（ネガティブ感情）から 1（ポジティブ感情）の間で数値化して表示するインタフェースを提案する研究[191]もあったが，その効果については未検証である．

単純な二項ラベルによる警告はユーザの信念には影響を与えないとする結果や[192]，未検証であることからラベルが付与されていないものを正確であると判断して共有意思が高まる（暗黙の真実効果）[188] という限界が指摘されている．また，スコアラベルの効果を検証する研究は少ない．

2.4.2. 情報源の信頼性ラベル

情報源の信頼性ラベルは，ニュース記事の提供元がプロのファクトチェック団体によってどのように評価されたかを示すものである [58]．警告ラベルは特定の投稿コンテンツ

を対象に付与されるものであるのに対し，情報源の信頼性ラベルはニュースサイトごとに付与される．信頼できない情報源のニュースをユーザが判別できるよう支援することが目的であり，信頼度のレーティングが表示されることにより虚偽ニュースの見出しが共有される可能性が低くなることが示されている[193]．この効果は，評価者がファクトチェッカーであっても一般ユーザであっても同様にみられたが，ファクトチェッカーの評価を使用する方が効果は高かった．最も大規模な情報源評価データベースを持つ米 NewsGuard 社は，9 つのジャーナリズム基準に基づいて 0 から 100 の間で情報源の信頼性を数値化して評価・提供している[194]．

情報源の信頼性ラベルを提供するファクトチェック団体はその評価の安定性や有用性を高める等の取組みをしている[195]が，情報源の評価結果がファクトチェック団体ごとに異なることが指摘されている[196]．

2.4.3. デバンキングと反論

デバンキングとは，特定の誤解が広まってしまった後に，その誤解に対処するための訂正情報を提供することである [58]．目的はユーザの誤った信念を減らすことであり，誤りであることの説明と共に事実情報を提供したり（トピック反論），誤解を招くために使用された修辞的な戦術を暴露したりする（テクニック反論）．単純に訂正情報だけを提供した場合, Misinformation を信じ続けるという誤情報持続効果が生じることが指摘されている[197]．特に，怒り等のネガティブな感情を喚起させる Misinformation は，誤情報持続効果を誘発する可能性が高いと言われている [52]．このため，思考に影響を与え続ける Misinformation の影響を減らすためには, 訂正情報だけを提供するよりもデバンキングすることが有効であると考えられている[198]．

デバンキングと反論において, Misinformation に対する反論メッセージが怒りを和らげる効果が報告されていた. 例えば, 実験で新型コロナワクチンに関する Misinformation に対して反論メッセージを提示したところ，Misinformation に起因する怒りが和らぎ，好意的なワクチン接種態度が促進された[199]．一方で，実世界におけるデバンキングの成功には，実施者の信頼性が大きく影響を与える [198]．正式な情報源であっても，自分の信念とは一致しない情報には注意を払わない場合があるため，対象グループに信頼されている発信者を使うといった工夫が必要となる．

2.4.4. プレバンキング（予防接種）

プレバンキングとは，Misinformation の影響を中和するために，遭遇前にユーザに対して認知的な予防接種を行うことである[200]．これは医療におけるワクチン（弱毒化したウイルス）の予防接種と同様の理論であり，①信念や態度に対する攻撃の警告と②先制的な反論をすることで事前に態度的な抵抗力を構築する[57]．プレバンキングの効果はデバンキングよりも高いという主張もある[200]．その理由として，プレバンキングの焦点が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であることから，メッセージに批判等が含まれず共感を得やすいと説明されている[201]．

プレバンキングにおいて，Misinformation が悪用する操作テクニックとして「感情操作的な言葉」を説明するものがあった．Misinformation における 5 つの操作テクニック（感情操作的な言葉，支離滅裂，誤った二分法，スケープゴート，人格攻撃）を学ぶ動画を見ることは，Misinformation に対する共有判断の質を向上させた[202]．また，2024 年 5 月に Jigsaw が公開した「Prebunking with Google」[203]では，Misinformation が使用する 10 個の操作テクニックを解説しており，その 1 つが「感情的な言葉（Emotional Language）」だった．「感情的な言葉，特に恐怖や怒りといったネガティブな感情を喚起する言葉を使用するとデジタルコンテンツの拡散可能性が高まる」「人は感情が高ぶった状態に追い込まれると理性的で意図的な意思決定に集中できなくなり，影響を受けやすくなる」と解説されている．この操作テクニックへの対処法として，共有や返信をする前に時間を掛けること，コンテンツに対する自身の感情的な反応を振り返ることを推奨している．

Disinformation に対するプレバンキングも効果があることが示されている．Jigsaw は 2023 年 2 月[204]と 10 月[205]に実施したプレバンキングキャンペーンにより，動画の視聴者の識別能力が改善したと報告した．ウクライナ難民に対する態度を操作しようとする Disinformation の 2 つのナラティブ「生活費の高騰をウクライナ難民のせいにする」「難民の暴力的な性質を煽る」を特定し，それを事前に否定するための 6 本の動画が作成された．この動画をキャンペーン期間中，YouTube，Facebook，Instagram，X，及び TikTok 等で広告として配信した．動画には Disinformation の操作テクニックについての教育が含まれており，特に恐怖を煽る操作テクニックを学ぶ動画を見た視聴者は，動画を見たことがない視聴者よりも恐怖を煽る事例に対する検出能力が 4.5〜8.0%向上した[205]．

プレバンキングの形式は，テキストや動画等による受動的なプレバンキング[206]と，ゲ

ームによる能動的なプレバンキング（例えば，「Bad News」[207]「Breaking Harmony Square」[208]「Go Viral!」[209]）がある．いずれもユーザが Misinformation を共有する意欲を低下させる効果があり，能動的なプレバンキングの方が受動的プレバンキングよりも効果が高く[210]，少なくともその効果は 3 ヶ月持続した[211]．一方で，能動的なプレバンキングは自主的な参加に依存しているため効果は一部のユーザに限られる．また，プレバンキングがその効果を発揮するためには，対象とする Disinformation/Misinformation の内容をある程度予測する必要があり，プレバンキングの内容と実際の Disinformation/Misinformation の内容に大きな乖離がある場合は効果が低い可能性がある [57]．

2.4.5. ナッジ

ナッジの定義は「選択を禁じることも，経済的なインセンティブを大きく変えることもなく，人々の行動を予測可能な形で変える選択アーキテクチャのあらゆる要素」[212]である．Disinformation/Misinformation 対策におけるナッジの主な目的は，ユーザの共有行動に介入することで共有の再考を促すことである．Disinformation の共有を減らす効果が検証された推奨策には，「フリクション」「正確さナッジ」「社会的規範ナッジ」がある．

フリクションは，意図的に関連するプロセスを遅くしたり，手間をかけさせたりするデザインのことを指す [58]．フリクションはユーザの行動を一時停止させて，その行動を冷静になって考え直すよう促すことを目的としていることから，熟慮を促す効果があると言われている[213]．主に Disinformation 対策としては，ユーザが共有しようとした時にポップアップウィンドウを表示して共有行動を再考させたり，確認行動をさせたりする仕組みとして既にソーシャルメディアに導入されている．フリクションは情報提供や危険性を伝達する警告よりもユーザの行動変容を促す効果が高いとされており[214]，共有しようとした時に内省の時間が与えられることに好意的なユーザもいる[215]．一方で，行動が中断されることを迷惑だと感じたり，不要又は押しつけがましいと感じたりするユーザもいる[216]．

正確さナッジは，ユーザがニュースを共有する前に，ニュースの正確さについて考えるよう促すものである[217]．これは，ユーザは基本的に Misinformation の共有を回避したいと願っているにも関わらず，党派性等が正確さ以外の要素に注意を向けさせてしまう[218]ために共有してしまうという考えに基づく．正確さナッジは，ユーザの注意を正確さ

に向けさせる．X で見出しの正確さを評価するよう促されたユーザは，その後の共有の判断がより慎重になった [50]．一方で，正確さナッジは，正確ではないと判断しても故意に共有すること [51]，強い党派性を持つユーザには効果がないこと[219]，真実と虚偽の区別が難しいコンテンツの場合や真実だと信じている場合[220]は効果がないことが指摘されている．

社会的規範とは「ある集団のメンバーによって理解される規則や基準のことで，法律の力を借りずに社会的行動を導く又は制約するもの」と定義されている[221]．社会的規範ナッジは，ニュース記事の共有における意思決定の質を向上させることを目的とし [57]，Misinformation を信じたり共有したりしないように社会的な情報（仲間の影響力）に注意を向けさせるものである [58]．社会的規範には，「大多数が承認する/承認しないこと」を知らせる命令的規範と，「一般的に行われていること」について知らせる記述的規範の 2 種類がある[222]．命令的規範はユーザによる Misinformation の報告率を高め[223]，記述的規範は虚偽のニュース記事を共有しようとしたユーザの割合を減らした[224]．ニュース記事の URL リンクの上に社会的規範ナッジのテキストメッセージを配置するだけで実装が可能だが，反社会的な行動が蔓延している状況では機能しない可能性があることが指摘されている [224]．

2.4.6. メディアリテラシーのヒント

メディアリテラシーのヒントとは，ソーシャルメディアをユーザが利用している時にニュースの真偽を見抜く方法のヒントを提供することである[225]．ニュース見出しの正確さを評価する際に，自動生成される記事のプレビュー，写真，情報元の Web ドメイン等のヒントをユーザに提供することによって，ニュースの真偽を見分けるユーザの能力が向上した [225]．2022 年 10 月に OECD（経済協力開発機構）がメディアリテラシーの 5 つのヒント「情報源を調べる」「証拠を確認する」「他の報告も見てみる」「見出しを疑う」「通常とは異なる書式に注意する」を表示した後の共有行動を測定する実験を行い，共有前に正確さに注意を向けさせる「正確さナッジ」よりも 3.5 倍誤った見出しの共有を減らしたと報告した[226]．疑わしい投稿コンテンツに関するコンテキストをユーザへ提供する仕組みの研究も行われており，ユーザが投稿コンテンツを共有しようとした時にソーシャル Bot 検出や情報源の信頼性等をフィードバックする [33]，投稿コンテンツ内で使用されているプロパガンダのテクニックをハイライト表示するといったようなものがあ

った[227]．一方で，効果には個人差があり [225]，またその効果は時間と共に減衰するという限界が指摘されている [58]．

2.4.7. AI 又は自動プログラムによるユーザ支援

ユーザが Disinformation に対抗できるよう支援するために，AI エージェントや自動プラグラムを活用する研究がいくつかある．EU の Horizon2020 に採択された「Co-Inform: Co-Creating Misinformation-Resilient Societies」[228]では，Misinformation の検出を支援するインテリジェントプラットフォームを開発し，それと連携する「MisinfoMe Bot」を提供した[229]．このソーシャル Bot は，開発したインテリジェントプラットフォームにて既に Misinformation と判定されたコンテンツの URL をユーザが X で共有した時に，当該ユーザに対して警告をダイレクトメッセージで送信する．ユーザとのダイレクトメッセージでのやり取りの合間に，MisinfoMe Bot は「Misinformation は，恐怖や怒りといった強い感情を引き起こす傾向があります」等の追加情報を提供する．

大規模言語モデル（LLM）や AI エージェントによって Misinformation の拡散を抑制しようという試みも行われている[230]．例えば，LLM を用いて Misinformation に対する事実に基づいた応答を生成することでデバンキング又はプレバンキングをすることが提案されている．また，オンラインチャットでのカスタマーサポートに AI エージェントを活用する事例のように，他者に相談しにくい公衆衛生（性感染症等）に関する Misinformation に対して AI エージェントが共感的コミュニケーションをすることは, 人間よりも効果的なユーザ支援をすることができる可能性が示唆されている[231]．特定の性格特性を与えた AI エージェント同士で Misinformation の拡散における説得を試みた研究では，性格特性の組み合わせによって説得の成功率が異なること，また対立的なアプローチよりも感情的につながる相互理解と信頼の構築が高い効果を発揮することを示唆した[232]．ただし，AI エージェントによる介入が逆効果を引き起こす可能性が指摘されており [230]，また AI エージェントによるユーザ支援が Disinformation 対策としても有用かどうかは，今後更なる検証が必要である，

2.5. 教育団体等における取組み

Disinformation 対策には，ユーザの能力を対象に認知的抵抗力を高める教育とブーストの試みがある. ブーストとは,「既存能力の育成又は新たな能力を習得させることで, 人々

が自らの意思決定をより容易に行えるようにする介入」と定義されている[233]．教育とブーストの目的は，Disinformation/Misinformation 対策に関連する能力を向上させ，ソーシャルメディアユーザの認知的抵抗力を高めることである．その推奨策には，「メディア情報リテラシー」と「横読みと検証戦略」がある．

2.5.1. メディア情報リテラシー

メディア情報リテラシーとは 2008 年に UNESCO が提唱した用語であり，発展経緯の異なる情報リテラシー，メディアリテラシー，デジタルリテラシー，及びニュースリテラシー等が包括されている[234]．メディア情報リテラシーの定義は「市民が個人的，職業的，及び社会的活動に参加し関与するために，批判的，倫理的，及び効果的な方法で様々なツールを用いることで，あらゆる形式の情報やメディアコンテンツにアクセスし，検索し，理解し，評価して利用し，創造し，共有するための一連の能力」とされている[235]．この定義では，批判的思考（critical thinking）もメディア情報リテラシーの 1 つとして位置づけられている．メディア情報リテラシーは個人の意思決定に及ぼす反社会的なメッセージの影響を低減する戦略の 1 つであり，フェイクニュースの識別能力を高め[236]，共有を減らす効果があることが先行研究より示唆されている[237]．一方で，教育対象（学生以外への効果的な介入），教育内容と効果測定指標のバラつき，及び地域・文化による効果の違い等の限界があることが指摘されている [57]．

国内外で提供されている代表的なメディア情報リテラシー教育プログラム 22 件（付録 1）を調査した．なお，教育団体がメディア情報リテラシー教育をするための教材として，政府機関が公開しているものも本調査に含めた．その結果，Disinformation の感情的側面に言及するものが 5 件あったため，その詳細について述べる．

(1) Mind Over Media: Analyzing Contemporary Propaganda[238]

2016 年に EU の資金提供プログラム Media Literacy for All で採択されており，現代的なプロパガンダを分析するために必要な知識とスキルを身に付けることを目的としている．プロパガンダが用いる 4 つの技法「強い感情を呼び起こす」「公衆の要望に訴えかける」「情報や考えを単純化する」「対立相手を攻撃する」を認識することが，批判的思考を身に付ける重要な第一歩であると説明している[239]．

(2) GET YOUR FACTS STRAIGHT![240]

2018 年に EU の資金提供プログラム Media Literacy for All で採択されており，若者，両親，及び祖父母を対象に Disinformation は何かを理解し対処方法の基本を学ぶことでメディアリテラシースキルを習得することを目的としている．感情を刺激するものはより早く広まるという「感情と Disinformation の関連性」を説明し，情報が真実か虚偽かを確認するためにできることとして「文章に感情的な表現がないか確認」するよう推奨している[241].

(3) START2THINK[242]

2019 年に EU の資金提供プログラム Media Literacy for All で採択されており，Disinformation の可能性を警告し，認知的な予防接種を行うことで批判的思考スキルを強化することを目的としている．Disinformation が用いる手法として，「閲覧者の感情に働きかけるために極端な意見，見解，声を増幅させること」「特定のナラティブを広めるために感情的な Disinformation を提示すること」が紹介されている[243].

(4) Spot and fight disinformation[244]

2021 年 1 月に欧州委員会が教育者向けのツールキットとして公開したものであり，15〜18 歳の学生に対して Disinformation がもたらす脅威や身を守る方法を説明して考える機会を提供することを目的としている．教育の中で，Disinformation による感情操作について触れ，「権威ある情報源に従った客観的な意見ではなく，強い感情を呼び起こすために，特定の状況における被害者として誇張されたアクターを使う」と説明している．対処法では，「強い感情を呼び起こすようにデザインされているかもしれない」として「共有する前に考える」ことを推奨している．2024 年 5 月に改訂版が公開された[245].

(5) インターネットとの向き合い方〜ニセ・誤情報にだまされないために〜[246]

2022 年 6 月に日本の総務省が教育教材として作成・公開したものであり，受講生がニセ・誤情報に関する理解を深め，情報を適切に扱う力を養うことを目的としている．欧州の「Spot and fight disinformation」と「GET YOUR FACTS STRAIGHT!」を参考に，日本人向けの事例への変更や有識者による意見が盛り込まれた．ニセ・誤情報は感情を悪用して拡散させようとすることを学び，「転送や拡散の前にひと呼吸」するよう推奨してい

る．この教材は事前の効果検証テストにおいて，講座前よりも平均点が有意に増加したことが示されている[247]．2025 年 2 月に公開された第 2 版では，ニセ・誤情報の感情的側面に関する説明が追記された[248].

その他，Disinformation 対策ではなくヘイトスピーチ対策として「社会性と情動の学習（SEL：Social and Emotional Learning）」の教育コンテンツが 1 件あった．「SELMA（Social and Emotional Learning for Mutual Awareness）」[249]は，EU の資金提供プログラム Rights, Equality and Citizenship Programme 2014-2020 で採択されており，相互認識，寛容，及び尊重を促進することによってオンラインのヘイトスピーチの問題に取組むことを目的としている．ヘイトスピーチに関わる状況下では感情のコントロールができなくなり，状況を悪化させ，他者を危険にさらすような反応や行動につながる可能性があるため，このような感情を認識し，調整し，適切な行動をとるためのスキルが重要であるとした[250].

2.5.2. 横読みと検証戦略

横読みとは，ファクトチェッカーが情報の信頼性を評価するために使用する戦略である [58]．ファクトチェッカーは，見知らぬコンテンツに深く入り込む前に計画を描き（舵取り），その情報はざっと見るだけで他のインターネットリソースを調べることで情報について詳しく知り（横読み），乱雑に検索結果をクリックするのではなく注意深く吟味した（クリック抑制）[251]．このファクトチェッカーが使用した戦略を学ぶことで，情報源を調査し，証拠を批評し，信頼できる情報源を見つける能力が向上した[252]．メディア情報リテラシーの教育プログラムにおいて，この横読みの手法（情報源，情報の確認，情報の追跡）を学ぶことを目的としたものがあった[253]．一方で，教育に時間がかかること，ユーザが学んだ戦略を実行する意欲を持つ必要があるといった限界が指摘されている [58].

2.6. 小括

本章では，これまでに実施又は検討されている Disinformation 対策について広く調査し，その中で Disinformation の怒りを含む感情的側面に言及しているものについても確認した.

第一に，民主主義の価値観を重視する国及び政府機関では，表現の自由を尊重し，Disinformation によくみられる「違法ではないが有害な投稿コンテンツ」の対処は，注意

義務又は行動規範の遵守を求めるに留まっていることが分かった.

第二に，ファクトチェック団体においても Disinformation の判定は難しく，プラットフォーム事業者はファクトチェックに基づく真偽よりも，投稿コンテンツに関するコンテキストを提供することでユーザの判断を支援する対策に移行しつつあることが分かった.

第三に，学術・研究機関及び教育団体等の取組みにおいて，Disinformation の怒りを含む感情的側面に言及する対策は 3 つ（デバンキングと反論，プレバンキング，及びメディア情報リテラシー教育）があることが分かった.

これらの調査結果から，民主主義を重視する国においては表現の自由の観点から Disinformation に対する法的対処，ファクトチェック判定，及びプラットフォーム事業者による対処が難しく，ユーザにその判断が委ねられていることが分かった．そのため，ユーザ向けの取組みは Disinformation の真偽を見分けるものが多く，Disinformation の怒り等の感情に着目するものは少なかった.

3. 関連研究

本章では，人が情報を共有するメカニズムについて関連研究を調査し，現対策が怒りによる Disinformation の共有を十分に解決し得るか，またどのようなユーザ介入策が Disinformation の怒りによる共有に有効な可能性があるか考察する．「2.Disinformation 対策の調査」では，これまでに実施又は検討されている Disinformation 対策について広く調査した．いくつかの現対策において，怒りを含む感情的側面に言及するものがあった．しかし，いずれも Disinformation が悪用する怒りに対処することを主目的とした対策ではなかったため，人が情報を共有するメカニズムの観点からは不十分な可能性がある．これらの対策が Disinformation の怒りによる共有に対して十分な効果があるか明らかにするために，人が情報を共有するメカニズムについて心理学等における関連研究を調査する．明らかになった怒りの共有メカニズムに対する有効策を検討するにあたり，Disinformation 対策以外で怒りを含む感情に着目した介入策とその実装方法を調査し，Disinformation の共有を減らす対策として応用可能かどうか考察する．

3.1. 怒りが共有に及ぼす影響

Disinformation の怒りを生み出す要因とその影響を明らかにするために，人が情報を共有するメカニズムの観点から感情及び怒りがユーザに対してどのような影響を与えているのか関連研究を調査した．現対策が Disinformation の共有を減らす上で有効かどうかを明らかにするためには，Disinformation の怒りがどのような要因から生じ，共有を促進しているのかを理解することが重要である．そこで，心理学等の関連研究を調査し，感情及び怒りによる影響を明らかにした．

3.1.1. 感情による影響

ソーシャルメディアにおいて共有を促進する要因は複数報告されているが，その中でも感情の影響が大きいことが示唆されている．例えば，感情の他には，投稿コンテンツのエンゲージメント率（いいねの数等）[254,255]，社会的地位や影響力を高めようとするオピニオン・リーダーシップ行動[256]，情報そのものの価値の高さ[257]，情報が正確であるという真偽性判断 [51]，個人の価値観や党派性からなる先有態度[258,259]がある．これらの要因の中でも感情の影響が大きいことがいくつかの研究で示唆されており，共有する動機は

情報的な動機よりも感情的な動機が強く[53]，他者を助ける，相互性を生み出す，自分の評判を高めるための共有よりも感情的な動機による共有は多い[54]．また，フェイクニュースの正確性を判断する際に理性ではなく感情に頼ると誤って正確だと認識しやすく[260]，人は情報の真理値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ことが予測されている[55]．したがって，感情による影響がなければ，ユーザは情報的な動機，他者や自己のため，そして理性的な真偽性判断に基づいた共有をする可能性がある．

感情は人の認知，意思決定，行動プロセスに強い影響を与える．1884 年に James が「What is an emotion?」という論文[261]を発表して以来，心理学領域において感情に関する多くの研究が行われてきた．感情は厳密に定義することが難しく，いまだに標準的な定義は存在していない[262,263]．最も広義な定義として，Ortony らの「感情とは，人が心的過程の中で行う様々な情報処理のうちで，人，物，出来事，環境についてする評価的な反応である」というものがある[264]．感情の種類についても一意な定義はなく，Ekman and Friesen は 6 種類の基本感情（後に 7 種類に変更），Izard は 10 種類の基本感情（後に 6 種類に削減），Plutchik は 8 種類の基本感情があると定義している[265]．これら感情は，自由連想，想像力，社会的認知等の認知プロセス[266]，意思決定プロセス[267]，及び行動プロセス[268]に強く影響を与えるとされている．

人は，感情的な出来事を体験した後に，その出来事やそれに対する自分の反応について他者に話す「情動の社会的共有」という対人プロセスを開始する[269]．社会的共有をされた受け手がそれを第三者へ話すプロセスは「二次的な社会的共有」と呼ばれる[270]．これは，ソーシャルメディアユーザによるコンテンツ投稿（ポスト）が社会的共有であり，他者の投稿コンテンツを共有すること（リポスト/シェア）が二次的な社会的共有にあたる[271]．マーケティングの観点では，情報をバイラル化するにあたり，受け手との感情的なつながりを構築する重要性が語られてきた．バイラルニュースの定義は，「他のニュース記事よりもはるかに迅速かつ広範に，主にソーシャルメディアを通じてオンライン上で拡散するネットワーク化されたニュース記事」とされている[272]．Berger はバイラル性を促進する 6 つの要素（STEPPS）として，社会的価値，トリガー，感情，公共性，実用的価値，及びストーリーを挙げた[273]．このうち，成功したバイラルマーケティングキャンペーンは受け手に感情的な反応を引き起こしており[274]，感情とバイラル性には強い関係があることが示唆されている[275]．また，バイラル性の高さは感情の種類によって異なり，ポジティブ（畏怖）又はネガティブ（怒り/不安）な感情を喚起するコンテンツはバ

イラル性が高く，悲しみを喚起するコンテンツはバイラル性が低い[54].

このような感情の種類によるバイラル性の違いが，真実のニュースと虚偽のニュースの拡散の違いを生み出している可能性がある．Twitter に投稿されたニュースへの返信に含まれる感情を Plutchik による 8 つの基本感情に分類して評価した研究では，真実のニュースは期待・喜び・信頼・悲しみの返信が多く，虚偽のニュースは驚き・嫌悪の返信が多かった[276]．同様に，Weibo に投稿されたフェイクニュースとリアルニュースとそのコメント欄に含まれる感情を分析した研究では，リアルニュースは幸せが多く，フェイクニュースは怒りが多かった[277].

3.1.2. 怒りによる影響

怒りは，「特定の認知的・知覚的歪みや欠陥，主観的ラベリング，生理的変化，社会的に構築・強化された組織的行動規範に従う行動傾向等に関連する，ネガティブで現象的（又は内的）な感情」と定義されている[278]．一般的に怒りが生じる 2 つの状況は，(a)自分たちが被害を受けたと思う場合，(b)自分たちが不当に被害を受けたと思う場合である[279]．このような状況には，自分自身だけでなく自分が所属するグループも関係する[280]．怒りには権力に対して真実を語るという役割もある[281]．しかし，怒りは利益よりも有害な結果をもたらすことが多い[282]．例えば，怒りは社会的対立の解決を妨げる意思決定を誘発する．Disinformation は怒りを悪用して，非効果的な行動又は逆効果な行動（例えば選挙のボイコットの呼びかけ[26]）等の有害な目標を追求するように集団を挑発する．怒りは一旦活性化されると，怒りの原因と関係あるかどうかに関わらず，人々の認識を変え，行動を導く[283]．多くのユーザにとって社会的比較によるネガティブな感情は，自分たちの価値観の追求に役立つどころか無力化するだけであり，自分やグループ全体のウェルビーイングにつながる最善の行動を促進することができない[31].

怒りは，対面でもソーシャルメディアにおいても共有行動を促進する．対面においては，ネガティブな感情が強いほど他者への共有行動が促進される[284]．ソーシャルメディアにおいても，畏敬，怒り，又は不安を喚起するコンテンツは共有されやすく，悲しみを喚起するコンテンツは共有されにくい[54]．Weibo の投稿コンテンツを 4 種の感情カテゴリー（喜び/怒り/嫌悪/悲しみ）に分類して拡散傾向を分析した研究では，怒りと喜びは伝染力が高いことが示された[29]．ただし，その拡散傾向は異なり，喜びがコミュニティ内（友人又はフォロワー同士）で共有されるのに対し，怒りは見知らぬ人が頻繁に

共有するため異なるコミュニティにまで広く拡散していた．また，真実よりも虚偽のニュースが拡散しやすいのは[276]，怒りが虚偽を信じさせることによって共有が促進される傾向があるからである[285]．怒りは，単純な手がかり（ステレオタイプ，情報源の専門性，及び情報源の信頼性）に頼る傾向が強く[286]，真偽性判断の正答率が低い[287]．Bago らは真偽を見分ける際に熟慮を促すことによって直感的な誤りが修正され，虚偽のニュースを信じる可能性が低くなったことを報告した[288]．これは，人の判断や意思決定の認知処理プロセスは直感又は熟慮から構成されるという「二重過程理論」[289]に基づく．怒りは熟慮的認知プロセスを妨げる可能性が指摘されている[290]．多くのユーザは正確ではないと判断したコンテンツを共有したくないと考えているが[50]，怒りによって誤って正確だと判断し，それが共有につながっている可能性がある．

3.1.3. 怒りに対する現対策の考察

「2.Disinformation 対策の調査」において確認された，感情に着目した Disinformation 対策が，感情と怒りによる共有を十分に解決し得るか考察した．「3.1.1.感情による影響」において，感情が真偽に関わらず共有を促進している可能性があることが先行研究において予測されていることが分かった．「3.1.2.怒りによる影響」においては，怒りが直感的認知プロセスを促進し，熟慮的認知プロセスを抑制することによって誤った真偽性判断から Disinformation の共有が促進されている可能性が示唆された．そこで，これらの感情と怒りによる共有メカニズムの影響を，現在実施されている Disinformation 対策が減少させることができるか考察した．

現対策のうち，怒りに対する効果が見込まれるのは，能力を対象とした「メディア情報リテラシー」と信念を対象とした「デバンキングと反論」「プレバンキング」の 3 つだった．「デバンキングと反論」は事後対策であることから，怒りが Disinformation の共有を促進した時に介入するには間に合わない．一方，「メディア情報リテラシー」と「プレバンキング」は事前対策であるため，怒りが Disinformation の共有を促進した時に有効に機能する可能性がある．しかし，これらの介入策は能力又は信念を対象としたものであり，怒りによる共有という行動に直接介入するものではない．怒りの共有メカニズムにおいて，Disinformation の怒りがユーザの熟慮を妨げた場合，事前対策の効果が十分に発揮されないまま，ユーザは Disinformation を共有してしまう可能性がある．このため，Disinformation の怒りが共有行動を促進した時，すなわち怒りの共有メカニ

ズムに介入する怒りに着目した有効策がより効果が高いと考えられる．

3.2. 怒りに対する有効策と実装方法

怒りの共有メカニズムに対して効果が見込める有効策と，その有効策を実装する方法について調査した．「3.1.怒りが共有に及ぼす影響」において，感情が真偽に関わらず共有を促進している可能性や，怒りが熟慮的認知プロセスを抑制することで Disinformation の共有が促進されている可能性が示唆された．この怒りの共有メカニズムに対し，有効と考えられる介入策を検討する必要がある．そこで，Disinformation 対策以外において，怒りを含む感情に着目した行動的介入策があるかどうか調査し，それを Disinformation の介入策として実装する方法を確認した．

3.2.1. 怒りに対する有効策

怒りに限定されていないものの，感情に着目してユーザの行動に介入する研究があった．Kiskola らは，ユーザがコメントを投稿する際，自己内省と情動調節を支援することを目的としたユーザインタフェースデザインを提案した[291,292]．このユーザインタフェースは，ユーザがニュースサイトのコメント欄に無礼なコメントを書いて投稿しようとした時に，コメントテキスト内に感情的な表現が含まれていることをユーザに認識させるメッセージ等を表示した．Syrjämäki らも同じく，ニュース記事のコメント欄にユーザが無礼なコメントを書き込んだ時に，「あなたは強い情動を表しているようだ」とコメントのトーンを説明するユーザインタフェースを提案した[293]．この介入は，無礼なコメントによって何が起こるかユーザに考えるよう促すメッセージよりも，ユーザの無礼なコメント投稿を緩和する効果があった．このように，メッセージ内の感情的な要素を認識しやすくするだけで情動調節が実現する可能性があるとし，Kiskola らはこのような仕組みは「選択の自由を保ちながらユーザを優しく誘導するため，情動調節に向けたナッジのアプローチとして活用できる」[291] と述べている．

3.2.2. 情動調節

Kiskola ら及び Syrjämäki らの研究をもとに，怒りに対する対処戦略として，自身の感情に気付いてちょうど良い状態に調節する「情動調節」に着目した．情動調節とは，「どのような感情をいつ持ち，どのようにその感情を経験し，表現するかについて，個人が

影響を及ぼすプロセス」と定義されている[294]．人は感情が強過ぎると衝動的・反応的になりやすく，反対に感情が弱過ぎると無感覚になり何も考えられない状態になる[295,296]．このため，人が思慮深い決断を下すためには，感情の強さは強過ぎず弱過ぎず，バランスがとれている必要がある．そこで，Disinformation の怒りに対して情動調節を促すことができる可能性がある情動調節の戦略を調査した．Gross が提唱した情動調節過程モデルでは，感情生成の 4 つのモード（状況，注意，評価，反応）と各モードに関連する 5 つの情動調節戦略（状況選択，状況修正，注意の方向付け，認知的変化，反応調整）が挙げられている[297]．このうち，Disinformation を共有しようとしている状況において実行可能な注意の方向付け，認知的変化，及び反応調整に怒りに対する効果が見込める戦略があるか調査してまとめた．

(1) 注意の方向付け（気晴らし，集中，反芻）

注意の方向付けとは，特定の状況において，自分の感情を調節するために注意をどのように向けるかを指す．注意の方向付けの戦略には，気晴らし，集中，反芻がある．

気晴らしとは，「状況の異なる側面に注意を向けるさせるか，あるいは状況から注意を完全に逸らすもの」[297] である．例えば，近所の郵便局の配置や通りを走る 2 階建てバス等，外部の非感情的な内容に注意を向けるといったものがある[298]．特に，感情を喚起する刺激や感情とは無関係なポジティブな事柄について考えることを促す気晴らしは効果が高いとされている[299]．ポジティブな気晴らしには，楽しい記憶を心に浮かべてそれについて詳しく考える[300]，愛と思いやりの感情を他者に向けて相手の健康や幸福を祈るといったものがある[301]．

集中とは，「状況の感情的な特徴に注意を向ける」ことであり，「注意が感情やそれに伴う結果に繰り返し向けられる」ことを反芻と呼ぶ[297]．集中と反芻は，感情を誘発した怒りの記憶を再体験するため，怒りを増幅してしまうことがいくつかの研究で示されている[298,302,303]．

(2) 認知的変化（再評価，視点取得）

認知的変化とは，自分が置かれている状況に対する評価や視点を変えることで，その感情の意味合いを変えることを指す．認知的変化の戦略には，感情刺激の再評価と視点取得がある．再評価とは，「本格的な感情的な反応が引き起こされる前に，その状況の評

価方法を精神的に修正すること」[304]を指す．再評価には潜在的に否定的なシナリオにおいて好ましい結果を想像することによって，肯定的な考え方を採用することが含まれる[299,305]．例えば，描かれた出来事が良くなることを想像する[306]，状況は時間の経過と共に改善すると考える[307,308]，教会の前で泣いている女性の涙は死に関連した悲しみではなく結婚式に関連した喜びを意味するものと再解釈するといったものがある[309]．

視点取得とは，「自発的に他者の視点を取り入れ，他者の視点から物事を見ようとすること」[310]を指す．この能力は，他者の行動や反応を予測することで，対人関係をより円滑で有益なものにする[311]．視点取得において重要なのは，自己の視点ではなく，他者の視点から想像することである．例えば，公平な第三者の視点から出来事を見る [303]，その人になったつもりで想像する[312]といったものがある．特定の第三者の視点が指定されることもあり，科学者になったつもりで客観的・分析的にみるよう促すものもあった[313,314]．

(3) 反応調整（抑制）

反応調整とは，感情によって生じた生理的，経験的，又は行動的な反応に対して，可能な限り直接的に影響を与える試みを指す．反応調整で最も一般的な戦略は抑制であり[299]，感情表出の抑制（感情を表に出さない），感情体験の抑制（感情を追い出す），感情思考の抑制（考えないようにする），感情に伴う行動の抑制がある．このうち，感情体験の抑制と感情思考の抑制は効果的ではないと言われている [299]．感情表出の抑制はポジティブな感情を抑えるのには有効であるものの，嫌悪感等のネガティブな感情を減少させる効果はなかった [294]．怒りの表出を抑制することが攻撃的行動を減らすとは限らないため，怒りを感じた時の行動を抑制する能力が重要であるという主張がある[315]．

(4) 外発的情動調節

情動調節は他者の感情に影響を与えることがあり，この過程は外発的情動調節と呼ばれる [297]．例えば，友人が落ち込んでいる時に励ますことによって，友人の感情が緩和したりポジティブになったりするという影響が及ぶ．外発的情動調節においては，悲しみや苦痛といったネガティブな感情に対して，視点取得と共感的対応が有効であることが示されている[316]．共感的対応とは，他者の感情的経験に対する理解，承認，及び思いやりを伝えるという他者への働きかけに該当する．この共感的対応は，プログラミングされた仮想エージェントが表示するメッセージでも効果がみられた[317]．

3.2.3. 介入策としての実装方法

「3.2.2.情動調節」で確認された怒りに対して効果が見込める戦略を，ユーザが Disinformation を共有しようとした時の介入策として応用する場合の実装方法について関連研究を調査した．ユーザによる Disinformation の怒りによる共有を減らすためには，怒りが共有を促進した時に，ユーザに対して効果的な情動調節戦略を提供する必要がある．Kiskola らは，感情的な要素を認識しやすくするだけで情動調節が生じる可能性があるとした．そこで，Disinformation を共有しようとした時に，どのような方法でユーザに情動調節を促すことができるか，その実装方法について確認した．

(1) 情動調節の実装要件

ユーザが自ら情動調節を試みるためには，自身の感情への気付きと情動調節の必要性を認識する必要がある．情動調節を成功させるための重要な要素は，感情区分と，柔軟に感情を調節する方法を知ることである[318]．感情区別とは，人が自分の感情を識別し区別できる精度のことを指す[319]．個人が自分の感情反応に気づいていない場合，それを効果的に調節することはまず不可能であると言われている [318]．Webb らは，感情を調節する最初の課題は，調節の必要性を識別することであると示唆した[320]．この必要性は，人の現在の感情と，その人の感情基準によって定義される望ましい状態との間に不一致があるときに生じる．この不一致は，感情が強過ぎたり，頻度が高過ぎたり，持続時間が長過ぎたり, 状況に対して不適切な感情の種類であったりする場合に生じる [292]. 例えば，悲しい雰囲気の状況において笑いが込み上げてきた時に，その笑いを堪えるよう調節を試みるといったことがある．この不一致を検出する比較機能は，現在の自分の感情に注意が向けられた時に発動する[321]．

(2) 実装デザインの事例

情動調節を用いた介入策の実装方法として, 行動を対象に介入するナッジに着目した．これは，怒りの共有メカニズムに介入するためには，ユーザが Disinformation の怒りによって共有行動をしようとした時に介入する必要があるからである．情動調節は現在の自分の感情に注意が向けられた時に発動するため，ユーザが怒りから Disinformation を共有しようと行動した時にナッジにより介入をし，同時に感情に注意を向けさせる仕組み

を提供することで情動調節につながる可能性がある．

ナッジは，2008 年に提唱されてから主に公共政策で導入されてきたが，デジタル環境においてもユーザのプライバシーとセキュリティの意思決定を支援するために活用されている．ナッジは，ユーザが集中できない状態において意思決定をしなければならない時，又は判断することが難しい時に用いられる[212]．このため，デジタル環境におけるナッジは，ユーザインタフェースの設計を通じて，ユーザの意思決定の複雑さを克服することを目的としている[322]．Disinformation/Misinformation 対策の文脈においては，ユーザの共有行動を一時停止させたり[137]，真偽性判断を支援するためのコンテキストを提供したりする際にナッジが使用されている[153]．

ナッジは人々に望ましい行動を促すものであるが，その操作性に関しては倫理的な観点から議論がある．ナッジは，①意思決定とその実行プロセスを簡単にするものと，②特定の選択をするようゆるやかに促すものの 2 種類がある[212]．このうち，②特定の方向にナッジしようとする操作において倫理的な問題が生じるとして議論がなされてきた[323]．ナッジは透明性があれば倫理的であると主張する者もいれば[324]，問題は不透明であることよりもナッジが浅い認知プロセス（自動的で直感的なプロセス）を対象としていることだと主張する者もいる[325]．Hansen らは，ナッジアプローチの責任ある許容可能な使用のための枠組みとして，認知的な思考様式（内省的/自動的）×認識的な透明性（透明/非透明）の 4 カテゴリーからなるフレームワークを提案した[326]．このうち，内省的かつ透明なナッジは，操作性が低く，その意図と効果がユーザに認識されやすいことから，倫理的な問題を引き起こす可能性が低いと評価されている[327]．Caraban らは，これまで用いられてきたナッジの種類を 23 種に分類し，フレームワークの 4 カテゴリーにマッピングした．このうち，もっとも内省的かつ透明なナッジは視覚化だった．

視覚化は，インタフェースとして使用される介入デザインの一形態であり，情報を説明しながら提供する効果的で透明性の高い方法である[328]．人間の行動修正を直接的又は間接的に促すことを目的に説得力のある可視化（persuasive visualization）の研究が 2010 年代から行われてきた．データをグラフで視覚化することは，その話題について強い初期態度を持つ場合（懐疑的又は特定の信念に固執している等）を除き，テキスト又は表形式の情報よりも説得力があるとみなされる傾向がある[329]．Web サイトのパスワード作成画面に導入されているパスワードメーターは，ユーザが作成したパスワードの強度をカラーバーで示すことによって強度の高いパスワードを設定するよう促す効果がある[330]．

「2.Disinformation 対策の調査」でも前述した通り，Misinformation 対策の研究では真偽確率をドーナツチャートで示すことでユーザの共有を減らす試みがある[190]．この研究成果をもとに，Amin らは「ソーシャルメディアでの情報共有におけるユーザ行動に関連する技術を構築する際には，視覚的手法を優先する」ことを推奨している．

感情情報を視覚化する試みも行われている．2002 年にメールに入力したテキストから感情情報を抽出してチャーノフの顔グラフで視覚化する電子メールブラウザが開発されている[331].「2.Disinformation 対策の調査」でも前述したニュース記事の論争スコアと感情スコアを数値化してカラーバーで表示する研究[191]の他，オンラインチャットの会話テキストに含まれる感情情報から文章に応じた顔文字を提案する仕組み[332]，ソーシャルメディアにユーザが投稿した日記のテキストから顕著なトピックと感情を検出して画像モチーフを生成するという試みも行われている[333]．感情情報を視覚化するにあたり，感情の種類に合った色を研究するものもあり，英語話者を対象に評価した結果では怒りを表現する色は「赤」が最も多く選ばれ，次いで黒，灰色の順であった[334].

円グラフ等のグラフィカルな視覚化は，付随するテキストメッセージを提供することでその効果を強化することができる．例えば，パスワードメーターの研究において，カラーバーとテキストメッセージを組み合わせて表示する方が，いずれかのみを表示する場合よりも強度の高いパスワードの作成を促した[330,335]．また，画像とテキストを組み合わせて情報を視覚的に表現する教育ツールは，ユーザの認知的負荷を軽減し，虚偽のニュースを識別する能力を高めたとする研究があった[336].

3.2.4.　ナッジに関する留意事項

関連研究を参考に，情動調節を用いた介入策をナッジにより実装するにあたり，留意すべき事項を調査した．Kiskola らと Syrjämäki らは，情動調節をナッジにより実装していたが，ナッジにはその操作性の観点から倫理的な議論がある．既に「3.2.3.介入策としての実装方法」にて挙げた認知的な思考様式（内省的/自動的）×認識的な透明性（透明/非透明）のフレームワーク以外にも，ナッジに関して留意すべき事項があるか調査した．

ナッジには倫理的な問題に関する指摘がある．例えば，あらかじめ設定された初期設定を選択させる「デフォルト設定型ナッジ」は[337]，初期設定をユーザに選択することを強制しているとみなされることがある．このような強制的なデフォルト設定は，人を騙す欺瞞的なデザインを意味するディセプティブパターン（ダークパターン）において「事

前選択（Preselection）」と呼ばれている[338]．ナッジの操作によって生じる倫理的な問題として，自律性（自由を保障するか？），福祉（ウェルビーイングを促進できるか？），長期的弊害（長期的に人を非合理にするか？），及び民主主義と熟議（民主主義を弱体化するか？）への懸念がある[323]．自律性については，選択の自由としてナッジの影響に抵抗する真の機会を与えること，主体性を尊重して内省的かつ透明であることが対策として述べられている．この主体性に関しては，理性的な能力が回避されることが問題であることから，熟慮能力へ訴える「理性へのナッジ」が重要であるとする主張がある[339]．ナッジの提唱者である Thaler & Sunstein は，「ナッジが人々を特定の方向にナッジするのであれば，ナッジされる当人が自分の状況はよくなっていると感じる結果になる可能性が高いと確信していなければならない」と述べている[212]．つまり，ナッジにおける「望ましい方向」とは，「社会厚生（社会全体の幸福度の総和）を高める方向」であり，「パレート改善する（誰かを現状より不幸にすることなく社会厚生を上昇させる）方向」である[340]．ナッジを作成するにあたっては，倫理的な問題がないか確認するチェックリストが公開されている[341]．

ナッジの限界として，教育的な効果がないこと[342]，効果が長期間持続しないこと[327]，反発反応を引き起こす可能性があること[343]，慣れによって効果が減少すること[344]，及び実験室と比較して実環境では効果が減少すること[345]等がある．このうち，教育的な効果がないことと反発反応については，デフォルト設定型ナッジを利用する際のリスクとして挙げられている．解決策として，このようなナッジの潜在的な問題を認識し，教育との併用を検討したり，長期的な効果を測定したりする研究が必要である[342]．

3.2.5. 怒りに対する有効策の考察

関連研究の調査から得られた知見をもとに，Disinformation の怒りに対する効果が見込めるユーザ介入策を考察した．Disinformation の怒りが共有を促進するという問題に対し，怒りに有効な介入策は Disinformation の感情的な共有を減らす可能性がある．しかし，「2.Disinformation 対策の調査」で確認された怒りに効果が見込める介入策は，ユーザの共有行動に介入するものではなかった．このため，怒りの共有メカニズムに介入することが可能な行動を対象としたナッジに着目し，ユーザによる共有行動を減らす上で，どのような怒りに対する介入策が効果的かつ実装可能か「3.1.怒りが共有に及ぼす影響」及び「3.2.怒りに対する有効策と実装方法」で得られた知見をもとに考察した．

(1) 情動調節を用いた介入

共有を促す怒りに対して情動調節を用いた介入をすることで，Disinformation の共有を減らすことができる可能性がある．怒りは共有されやすく[54]，また Disinformation の感情が真偽に関わらず共有を促進している可能性がある[55]．この Disinformation の怒りが共有を促進するという問題を解決するためには，ユーザが怒りによって共有しようとした時に，怒りに着目した介入が効果的である可能性がある．ただし，怒りを過度に抑制することは，個人が望むより良い未来への変革を促す原動力を阻害する可能性がある．怒りという感情を大切な心の表現として扱っていくことが重要であることから，ユーザが自発的に自身の感情をちょうど良い強さに調節する情動調節を採用することが望ましいと考えた．いくつかある情動調節戦略のうち，怒りを増幅させる集中・反芻以外の，気晴らし，再評価，視点取得，行動の抑制，及び共感的対応が，Disinformation の怒りに対して有効な可能性がある．

(2) ナッジによる情動調節の実装

Disinformation の怒りによる感情的な共有に介入する方法として，従来のナッジに着目した．ただし，ナッジを採用するにあたっては懸念される倫理的な問題に対処する必要がある．ナッジの操作性に対しては，操作性がもっとも低いと考えられている内省的かつ透明な視覚化[327]を採用することとする．ユーザが Disinformation を感情的に共有しようとした時に内省的な熟慮を促すことは，人々を自律した責任ある主体として扱う理性へのナッジに該当する[339]．また，選択の自由の観点から，ナッジは自由に無視できるものである必要がある[323]．あくまでもユーザにとって望ましい意思決定を導くことが重要であり，ユーザの自律性や表現の自由は尊重されるべきである．

情動調節をナッジに組み込むにあたっては，視覚化のうち，説得力が高いとされる円グラフ[329]を用いるのが有用な可能性がある．また，その効果を強化するために付随するテキストメッセージを提供することが考えられる[330,335]．これにより，ユーザの認知的負荷が軽減され，ユーザの共有判断能力が高まる可能性がある[336]．

3.3. 小括

本章では，関連研究を調査することで怒りの共有メカニズムを明らかにし，怒りを含

む感情に着目した介入策が Disinformation の共有を減らす対策として応用可能かどうか考察した.

第一に，怒りの共有メカニズムとして，感情が情報の真理値に関わらず共有を促進する可能性があること，特に怒りと喜びが拡散しやすく，怒りは虚偽を信じさせることによって共有が促進される傾向にあることが分かった.

第二に,「2.Disinformation 対策の調査」において怒り等の感情に言及する対策が 3 つあったが，いずれも怒りの共有メカニズムによって対策の効果が十分に発揮されない可能性が考えられた.

第三に，怒りに効果が見込める情動調節を共有行動に介入可能なナッジで実装することが，Disinformation の怒りによる共有に対して効果がある可能性があることが考えられた.

現対策は，Disinformation の怒りによる共有に対して効果が不十分な可能性がある．このため，怒りの共有メカニズムに対する効果が見込める怒りに着目した介入策（情動調節を用いたナッジ）は，現対策の補完策となり得る.

4. 本研究で解決を目指す課題

本章では，調査及び関連研究より得られた知見に基づいて現状の Disinformation 対策における課題を挙げ，その課題を解決するための本研究の目的を示す．「1.序論」において，Disinformation は怒りを悪用することで共有を促していることが分かった．このため，「2.Disinformation 対策の調査」において実施又は検討されている Disinformation 対策を広く調査し，それらの対策が Disinformation の怒りに対しても効果的かどうかを「3.関連研究」において考察した．これらの結果をもとに現状の課題を述べ，その課題を解決するための本研究の目的及び対象範囲を明確にする．

4.1. 現状の課題

「2.Disinformation 対策の調査」の結果より，民主主義を重視する国及び政府機関においては表現の自由の観点から Disinformation に対する法的対処，ファクトチェック判定，及びプラットフォーム事業者による対処が難しく，ユーザにその判断が委ねられていることが分かった．しかし，ユーザ向けの取組みの多くは Disinformation の真偽を見分ける能力の向上又はその支援に重点が置かれており，Disinformation の怒りに着目するものは少なかった．Disinformation の怒りを含む感情的側面に言及する対策は 3 つ（デバンキングと反論，プレバンキング，及びメディア情報リテラシー教育）があるが，いずれも Disinformation が悪用する怒りに対処することを主目的とした対策ではなかったため，人が情報を共有するメカニズムの観点からは不十分な可能性がある．

「3.関連研究」からは，Disinformation の怒りが共有を促進するという問題を解決するためには，ユーザが怒りによって共有しようとした時に介入する「怒りに着目した介入策」が有効である可能性が示唆された．関連研究より，怒りの共有メカニズムがあることが分かった．デバンキングと反論は事後対策であり，メディア情報リテラシー教育とプレバンキングは事前対策だが怒りの共有メカニズムによって効果が十分に発揮されない可能性がある．このため，怒りの共有メカニズムに介入することが可能な，行動を対象としたナッジにより，怒りに着目した情動調節を用いた介入策を実装することが，感情的な共有を減らす可能性がある．

これらの調査結果より，Disinformation の怒りによる共有を減らすためには怒りの共有メカニズムに対処する必要があるにも関わらず，現対策では怒りの共有メカニズムに

対する効果は十分ではない可能性が考えられる．このように，現対策の限界について明確に示した研究はこれまでにない．Disinformation の怒りによる共有という現対策の限界を補完する怒りに着目した新たな介入策が必要だが，効果的な介入策を考案するにあたっては検討すべき重要な課題が 3 つある．第一に，Disinformation の怒りによる共有メカニズムが実証されていない．先行研究では，怒りがユーザの共有行動を促進することが実験により示されており [54]，実際のソーシャルメディア環境のデータ分析からは怒りはユーザ同士のコミュニティを超えて拡散する傾向があった [29]．このため，共有を促す怒りに対する介入策が，Disinformation の怒りによる共有の影響を減らす上で効果的である可能性がある．しかし，「Disinformation の怒り」が共有に及ぼす影響が明らかになっていないため，介入策による効果を検証することができない．

第二に，既存の Disinformation 対策において，ユーザが Disinformation を共有しようとした時に，怒りに着目した介入策によって Disinformation の共有を減らすものはない．人は感情が強過ぎると衝動的・反応的になりやすいため [295,296]，感情のバランスを整えるのを支援することがユーザの思慮深い判断につながる．怒りに対する介入策は，Disinformation の怒りによる影響を調節し，怒りの共有メカニズムによる感情的な共有を減らす可能性がある．しかし，「2.Disinformation 対策の調査」をした結果，Disinformation の共有を減らすために，ユーザが Disinformation を共有しようとした時に怒りに着目して介入する対策はなかった．

第三に，Disinformation の怒りに着目した介入策が，現対策と比較しても有用かどうか明らかではない．「2.Disinformation 対策の調査」において，怒りに関する Disinformation 対策として現在最も使用されているのは，メディア情報リテラシー教育だった．メディア情報リテラシー教育は Disinformation を識別する能力を高めるが [57,236]，その効果は怒りによって十分に発揮されない可能性がある．これは，「3.関連研究」に基づく予測であり，真偽性判断において怒りは直感的思考を促進し [287]，真偽性判断に必要な熟慮を妨げる [290] という怒りの共有メカニズムがあるからである．怒りに対する介入策は，この怒りの共有メカニズムに介入し，ユーザに思慮深い判断を促すものである．このため，メディア情報リテラシー教育の効果を十分に発揮させるためには，その効果を限定し得る怒りに対する介入策を補完的に実施することが重要である．しかし，Disinformation の怒りに着目した介入策が，メディア情報リテラシー教育の補完策としても有用かどうかは検証されていない．

4.2. 本研究の目的

本研究の目的は，Disinformation の怒りを生み出す要因が共有に及ぼす影響を明らかにし，Disinformation の共有を減らすために怒りに着目した有効策を提案することである．この目的を達成するために，本研究では以下の実験を行う．

第一に，Disinformation の怒りを生み出す要因による共有メカニズムを明らかにする（図 4-1 の実験①）．Disinformation の怒りと信憑性判断のそれぞれが共有に及ぼす影響の有無と，その影響の大きさを比較検証する．

第二に，Disinformation の怒りに着目した情動調節ナッジを作成し，従来のナッジよりも Disinformation の共有を減らす効果が高いことを比較評価により明らかにする（図 4-1 の実験②）．「3.関連研究」での考察に基づき，Disinformation の怒りに対して効果が見込める情動調節ナッジを作成する．情動調節ナッジが Disinformation の共有を減らす効果については，既存のナッジの効果と比較評価することで示す．

第三に，情動調節ナッジが既存のメディア情報リテラシー教育の補完策として有効であることを明らかにする（図 4-1 の実験③）．怒りに着目した介入策である情動調節ナッジが Disinformation の共有を減らす効果について，怒りに関する Disinformation 対策として現在最も使用されているメディア情報リテラシー教育の効果と比較評価する．これにより，Disinformation の怒りに対する現対策の限界を示すと共に情動調節ナッジが補完策として有用であることを示す．

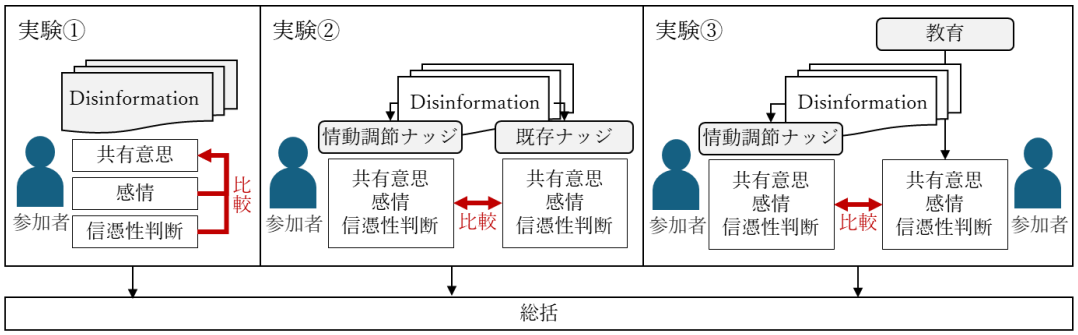


図 4-1 各実験の概要図

4.3. 対象範囲と想定する状況

本研究では，Disinformation がソーシャルメディアユーザ同士での社会的な意見の対

立・分断を狙うために用いられた状況を想定する．何故なら，本研究では Disinformation の共有を減らすために，共有を促す怒りに着目しているからである．Disinformation には，ワクチン接種に関連するもの，気候変動等の環境に関するもの，災害に関連するもの，株価操作を狙ったもの，対立・分断を狙うことで選挙干渉を試みるもの等がある．このうち，Disinformation により対立・分断を狙った過去事例では，怒りを悪用することでユーザの共有行動を促していたことが分かっている．このような社会的比較によって引き起こされる怒りは，報復的な情報や解決策の選好を促進し [30]，ユーザの最善の行動につながらない可能性がある [31]．このため，怒りによる共有を促す Disinformation の影響とその対策を検証することは，ユーザの最善の行動を導くことにつながると考える．

また，Disinformation の形態は，テキストのみの投稿文を対象とする．Disinformation はテキストと画像の両方を含む「マルチモーダル」であることが多いが，マルチモーダルの Disinformation で悪用される画像・動画は，95〜98%の高い精度で検出可能なツールが開発されており，技術的な対処が比較的容易と考えられる [42]．これに対し，テキストのみの場合は，表現の自由の観点から違法又は有害と認定され得る用語（例えば，ヘイトスピーチ）の検出 [43] に留まっている．生成 AI によるテキスト型の Disinformation の拡散についても危惧されているが[346]，開発されている検出ツールの精度は 26%であり，精度を改善するために利用できない状態が続いている [45]．このため，本研究の検証対象をテキストのみの Disinformation とし，有効策を検討することは有意義と考える．

検証は，実際のソーシャルメディア環境ではなく実験環境とし，X（旧 Twitter）に Disinformation が投稿された状況を想定する．Disinformation の共有は対面コミュニケーションにもおいても行われるが，ソーシャルメディアでは一度に不特定多数へ大量に拡散されることから，拡散数が対面の約 464 倍[347]と社会的な影響が大きい．X は，欧州委員会の調査において Disinformation と認定されたコンテンツの割合が最も高く[348]，日本においても総務省の調査で X は Misinformation 又はミスリーディング情報に最もさらされた媒体だった[349]．しかし，ソーシャルメディアの実環境における実験的操作 [120] には倫理上の問題があるとして，過去の研究で批判された経緯がある [19]．このため，本研究ではユーザが Disinformation に遭遇する確率が高く，その Disinformation をユーザが共有した場合に不特定多数への拡散につながる可能性がある X を想定した実験室実験とする．

4.4. 小括

本章では，現状の Disinformation 対策における課題を挙げ，その課題を解決するための本研究の目的を示した．現対策では怒りの共有メカニズムに対する効果は十分ではない可能性があるため，補完策として怒りに着目した新たな介入策が必要である．しかし，効果的な介入策を考案するにあたっては検討すべき重要な課題が 3 つあった．

第一に，Disinformation の怒りによる共有メカニズムを明らかにする必要がある．

第二に，ユーザが Disinformation を共有しようとした時に，「怒りに着目した介入」によって Disinformation の共有を減らす対策がない．

第三に，Disinformation の怒りに着目した介入策が，現対策と比較しても有用かどうか明らかではない．

これらの課題に対し，本研究では Disinformation の怒りを生み出す要因が共有に及ぼす影響を明らかにし，Disinformation の共有を減らすために怒りに着目した有効策を提案することを目的とした．この目的を達成するために，各課題を解決する実験を行う．

5. 怒りが Disinformation の共有に及ぼす影響

本章では，Disinformation の怒りを生み出す要因による共有メカニズムを明らかにする．「3.関連研究」において感情や怒りによる影響と有効なアプローチ手法を調査したところ，怒りがユーザの共有行動を促進していることが分かった．このため，共有を促す怒りに対する介入策が，Disinformation の怒りによる共有の影響を減らす上で効果的である可能性がある．しかし，Disinformation の怒りが共有に及ぼす影響が明らかになっていないため，介入策による効果を検証することができない．そこで，Disinformation の怒りを生み出す要因が共有を促進しているかを確認する本実験を実施する．また，本実験で使用する Disinformation 等の刺激を選定するための予備実験を事前に行う．

5.1. 予備実験

予備実験の目的は，本実験で使用するテキスト投稿刺激を選定することであった．テキスト投稿刺激とは，X の投稿文を模して実験者が作成した文章のことを指す．過去事例において拡散した Disinformation には，ユーザに怒りを認識させるという特徴があった．このような特徴を組み込んだテキスト投稿刺激を作成し，意図した特徴を備えているか予備実験で確認した．各テキスト投稿刺激に対する参加者の感情的な反応を測定し，選定基準に基づき評価することで，本実験で用いるテキスト投稿刺激を選定した．

5.1.1. 方法

(1) 参加者

実験者が所属する大学院と企業で募集された 26 人（男性 15 人，女性 11 人；年齢 20 代～60 代）より回答を得た（2021 年 8～9 月）．実験はデータ収集前に研究倫理委員会の承認を得ており，学術目的の調査であることを説明した上で参加者から同意を得た．

(2) テキスト投稿刺激

テキスト投稿刺激は，5 刺激条件×2 テーマの計 10 個に絞り込むために同じ刺激条件に該当するものを 2 種類ずつ（A/B），計 20 個が作成された（表 5-1）．日本語圏における X の投稿文字数制限に準拠した約 140 文字の文章で構成された．文章に含まれる情報の質は，取り上げる話題は 1 つとし，話題に関する具体的な情報源は記載しない等，情

報粒度の均一化が図られた.

表 5-1 テキスト投稿刺激の刺激条件

刺激条件	不当性	バイラル性	内容
Disinformation-viral (men/older)	あり	高い	実験者の創作
Disinformation-viral (women/younger)	あり	高い	実験者の創作
Disinformation-control	あり	低い	実験者の創作
True information-viral	なし	高い	既存ニュース
True information-control	なし	低い	既存ニュース

刺激条件は，不当性（あり/なし）とバイラル性（高い/低い）という 2 つの操作要因ごとに，2 個の Disinformation-viral 刺激，Disinformation-control 刺激，True information-viral 刺激，及び True information-control 刺激の計 5 個から構成された．不当性の有無は，Disinformation の特徴である怒りを認識させる操作要因として設定された．不当性がある Disinformation 刺激は怒りを認識し [279]，不当性がない True information 刺激は怒り以外の感情を認識することが予測された．バイラル性はマーケティングの観点から記事の共有を広げるために公共性や感情を高める要素を持つ [273] ことから，公共性から感情の強さを高める操作要因として設定された [274,275]．不当性がある Disinformation であっても，viral 刺激はバイラル性が高いことから強い感情を認識することで共有され，control 刺激はバイラル性が低いことから感情があまり認識されず共有されない可能性がある．

Disinformation-viral 刺激は，不当性がありバイラル性が高い，すなわち怒りが強く認識される刺激条件として実験者により創作された．前述（「4.3.対象範囲と想定する状況」）の通り，本研究では Disinformation が対立・分断を狙うために用いられた状況を想定することから，アメリカの下院情報特別委員会が公開した 2016 年米大統領選挙における Disinformation 事例（3,266 件）[350]の文章と構成が参考にされた．当該事例では，自分が属するアイデンティティが対立グループから不当に扱われていることを表現することで怒りを引き起こそうとしていた．これを日本向けに置き換えるにあたり，日本の社会における格差問題として関心が高い[351]，男女間対立（男性/女性）と世代間対立（高年層/若年層）の 2 つがテーマとして採用された．日本人の関心が高い格差には，他にも所得格差，地域格差，容姿による格差，又は学歴格差等が挙げられるが，自分が対立層のどちらに所属しているかという認識には個人差が大きい．これに対し，男女間対立と世代間対立は，参加者自身が認識している属性（性別/年齢）であり認識の個人差が小さい．

Disinformation-viral 刺激を見て自分が該当する属性[280]が不当に傷つけられたと思った場合[279]に参加者は怒りを認識する可能性がある．Disinformation は各対立層向けに流布されるため，男女間対立のテーマでは男性を狙った Disinformation-viral (men) 刺激と女性を狙った Disinformation-viral (women) 刺激，世代間対立のテーマでは高年層を狙った Disinformation-viral (older) 刺激と若年層を狙った Disinformation-viral (younger) 刺激が設けられた．

Disinformation-viral (older) 刺激の例： 若者のワクチン接種率が低いと聞いてるのに、外ではマスクなしで騒いでる若者がビックリするほど多い！感染者数がようやっと減ってきたものの、またいつ感染者が増えるか分からないからと多くの人ができるだけ活動を自粛して経済も停滞しているというのに。身勝手な若者が感染の脅威をまき散らしていると思うと非常に腹立たしい。

Disinformation-control 刺激は，不当性がありバイラル性が低い，すなわち怒りが弱く認識される刺激条件として実験者により創作された．Disinformation-viral 刺激と同様に対立グループに対する不当性が含まれるものの，社会問題ではなく個人的な関心又は個人的な問題と認識されやすいものとした．このため，不当性から怒りが認識されても，Disinformation-viral 刺激とは異なり感情の強さは弱いことが考えられた．

Disinformation-control 刺激の例： 定年が 65 歳まで延長されたが、60 歳以降は給与も上がらないし現状維持ができてればいいと考えている。年下の上司も仕事を頼みにくそうだし、後進育成といっても今の仕事はパソコンが中心でデジタルに疎い自分が教えられるようなこともない。高齢者の雇用拡大が若者採用を抑制すると批判もあるが、年金も 65 歳からだし生活のためだ。

True information-viral 刺激は，不当性がなくバイラル性が高い，すなわち怒り以外の感情が強く認識される刺激条件として設定された．True information-control 刺激は，不当性がなくバイラル性も低い，すなわち感情そのものが認識されにくい刺激条件として設定された．True information-viral 刺激と True information-control 刺激は，日本の伝統メディア又は Web メディアで発信された既存のニュース記事の要約あるいは一部を抜粋する形で引用された．そのうち，感情的な表現が多いものが True information-viral 刺激，客観的

な論調のものが True information-control 刺激に割り振られた.

True information-viral 刺激の例： 「若者の一票を高齢者の一票よりも重くすべき。」このコロナ禍が変化のチャンスと市長が語っている。若者は自分たちが弱者であることにすら気づいていない。カギを握るのは年金制度を支えている現役世代の行動である。若者は自分たちが相当まずい状況に追い込まれていることに一刻も早く気づくべきだ。

(3) 手続き

実験は，Google フォームを使用したアンケート形式で行われた．参加者には，実験中に提示された投稿は全て実験者によって作成された架空のものであること，感情がおさまってから実験を開始することが教示された．これは，実験以外に起因する感情（例えば，実験前にイライラする出来事を体験した等）による回答への影響を可能な限り減らすことを意図していた.

実験を開始すると，参加者はよく利用しているソーシャルメディアのタイムラインにテキスト投稿刺激が表示されている状況をイメージするよう求められた．テキスト投稿刺激 20 個がランダムな順番で提示され, 提示されるたびに参加者は以下 2 つの質問に回答するよう求められた.

1) 感情の種類：テキスト投稿刺激に対して認識した感情の種類について，Plutchik の「感情の輪」[352]から最も近い感情を 1 つ回答するよう求められた（回答選択肢：期待，喜び，信頼，恐怖，驚き，悲しみ，嫌悪，怒り，又は感情なし）.
2) 感情の強さ：何らかの感情を認識した場合，その強さを回答するよう求められた（回答選択肢：1.弱い～4.強い）．なお，1) 感情の種類において「感情なし」と回答した参加者には当該質問項目は表示されなかった.

実験の最後に，参加者は性別，年代，普段利用しているソーシャルメディア，及び関心のある日本の社会問題に関する質問に回答するよう求められた.

5.1.2. 結果

刺激条件の不当性による違いを確認するために感情の種類の割合について算出して比較した結果, 怒りを認識した参加者の割合は Disinformation 刺激（7.0%）が True information 刺激（1.3%）よりも多かった（図 5-1）．True information 刺激で認識された感情の種類は，

期待（48.1%）が最も多かった．感情の種類によって認識された感情の強さは異なり，信頼，恐れ，嫌悪，及び怒り（*Md* = 3, *QD* = 2-3）が強く，期待と喜び（*Md* = 2, *QD* = 1-3）が弱かった．

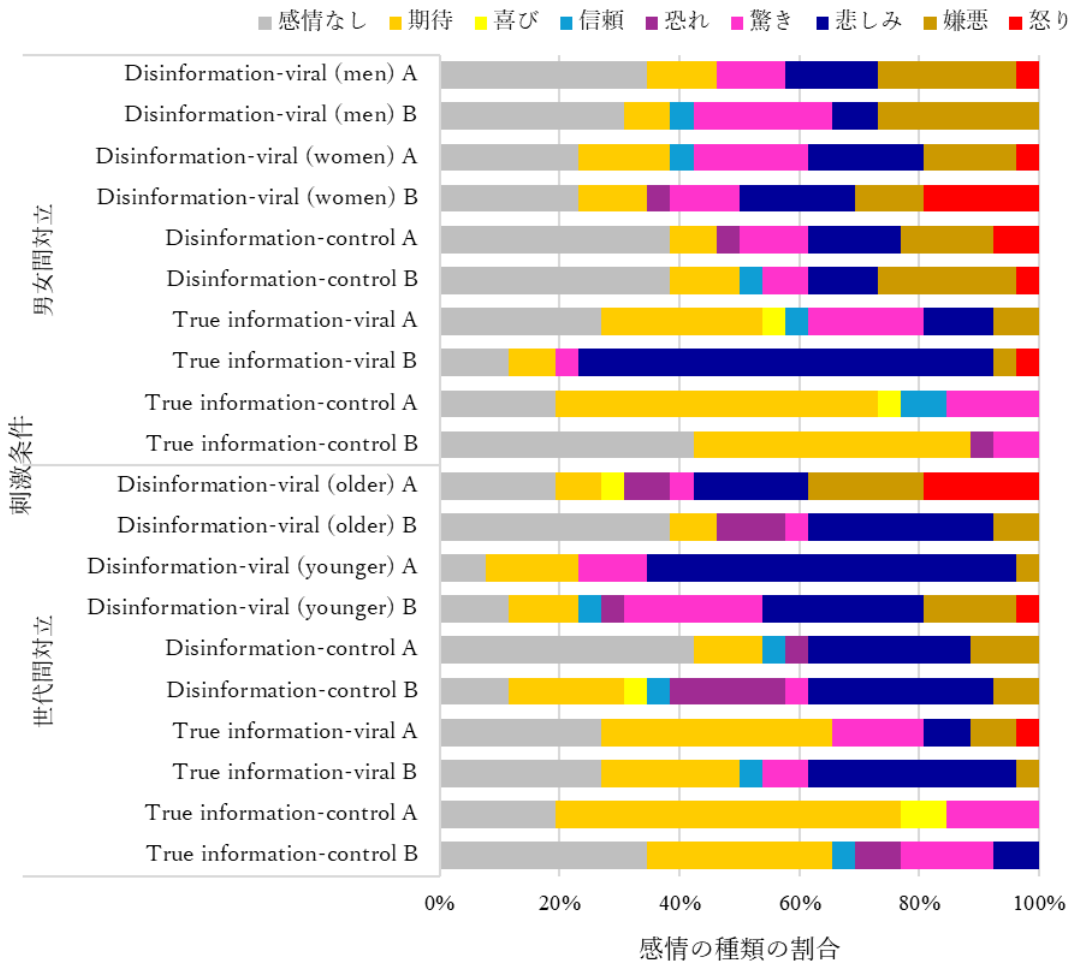


図 5-1 各テキスト投稿刺激で認識された感情の種類の割合

刺激条件のバイラル性による違いを確認するために感情の強さ（0-4）の中央値と四分位偏差を算出して比較した結果，viral 刺激（Disinformation-viral/True information-viral）は control 刺激（Disinformation-control/True information-control）よりも強い感情の回答が多い傾向がみられた．viral 刺激（*Md* = 2, *QD* = 1-3）と control 刺激（*Md* = 2, *QD* = 0-3）の中央値は同じだったが，第 1 四分位数は viral 刺激の方が強かった．テキスト投稿刺激 10 個×2 種類（A/B）ごとに感情の強さの中央値と四分位偏差を算出したところ，最も強い感情（*Md* = 3）が認識されたのはいずれも Disinformation-viral 刺激だった（表 5-2）．

表 5-2 テキスト投稿刺激 20 個の感情の強さ（中央値と四分位偏差）

テーマ	テキスト投稿刺激	A	B
男女間対立	Disinformation-viral (men)	2.0 (0.00-2.75)	2.0 (0.00-3.00)
	Disinformation-viral (women)	2.0 (1.00-3.00)	2.0 (1.00-3.00)
	Disinformation-control	1.5 (0.00-2.00)	1.5 (0.00-3.00)
	True information-viral	2.0 (0.25-2.00)	2.5 (2.00-3.00)
	True information-control	2.0 (1.25-3.00)	1.0 (0.00-2.00)
世代間対立	Disinformation-viral (older)	3.0 (2.00-3.00)	1.0 (0.00-2.00)
	Disinformation-viral (younger)	2.0 (2.00-3.00)	3.0 (2.00-3.00)
	Disinformation-control	1.0 (0.00-2.75)	2.0 (2.00-3.00)
	True information-viral	2.0 (0.25-3.00)	2.0 (0.25-3.00)
	True information-control	2.0 (1.00-3.00)	1.5 (0.00-2.00)

テキスト投稿刺激ごとに感情を認識した（感情の強さ 1-4）参加者と感情なし（感情の強さ 0）と回答した参加者の人数を比較しところ，viral 刺激では 12 個中 10 個の刺激において感情を認識した参加者の人数が有意に多かった．20 個全てのテキスト投稿刺激において，感情を認識した参加者の人数が感情なしと回答した参加者の人数よりも多かったため，その人数に有意な差があるかどうかを検証することとした．各テキスト投稿刺激に対する感情の強さ（1-4/0）を独立変数とし，それぞれの回答人数を従属変数とするカイ二乗検定を行った．その結果，感情を認識した参加者の人数が，感情なしと回答した参加者の人数よりも有意に多かったテキスト投稿刺激は viral 刺激では 12 個中 10 個と多く，control 刺激では 8 個中 3 個と少なかった（表 5-3）．

表 5-3 テキスト投稿刺激 20 個のカイ二乗値

テーマ	テキスト投稿刺激	A	B
男女間対立	Disinformation-viral (men)	2.46	3.85*
	Disinformation-viral (women)	7.45**	7.45**
	Disinformation-control	1.38	1.38
	True information-viral	5.54*	15.38**
	True information-control	9.85**	0.62
世代間対立	Disinformation-viral (older)	9.85**	1.38
	Disinformation-viral (younger)	18.62**	15.38**
	Disinformation-control	0.62	15.38**
	True information-viral	5.54*	5.54*
	True information-control	9.85**	2.46

* $p < .05$, ** $p < .01$, *** $p < .001$

5.1.3. 考察

参加者が Disinformation 刺激に対して認識した感情の種類は True information 刺激と比較して嫌悪又は怒りが多く，不当性があると怒りを認識するという操作要因と一致していることが確認できた．得られた結果は，Vosoughi らによる Twitter データの分析結果でみられた「true ニュースは期待，喜び，信頼，又は悲しみの返信を促進し，false ニュースは驚き又は嫌悪を強める返信を促進した」[276] という傾向と類似する結果だった．また，バイラル性については，テキスト投稿刺激によって傾向はやや異なるものの，概ね viral 刺激の方が control 刺激よりも強い感情を認識した参加者が多いことが確認された．

20 個のテキスト投稿刺激に対して得られた回答に基づき，本実験で使用するテキスト投稿刺激 10 個（2 テーマ×5 刺激条件）を 2 種類（A/B）のいずれかから採用した．バイラル性は感情の強さを高める操作要因であり，この定義を充たすよう，Disinformation-viral 刺激と True information-viral 刺激は，感情を認識した人数が有意に多かった方のテキスト投稿刺激を採用した．一方，Disinformation-control 刺激と True information-control 刺激は，感情を認識した人数に有意差が認められなかった方のテキスト投稿刺激を採用した．2 種類（A/B）とも有意差あり又はなしの場合は，Disinformation-viral 刺激では刺激条件が対象とする属性において感情の強さの中央値等が高かった方，Disinformation-control 刺激は感情の強さの中央値等が低かった方，True information-viral 刺激は感情の強さの中央値等が高かった方のテキスト投稿刺激を採用することとした．

5.2. 本実験

本実験の目的は，Disinformation の怒りを生み出す要因による共有メカニズムについて，要因間で比較検証することにより明らかにすることだった．これは，先行研究 [55] の予測に基づくものであり，これまでに検証されていない．そこで，予備実験で選定したテキスト投稿刺激を用いて，Disinformation の怒りを生み出す要因が共有に及ぼす影響を本実験で確認した．怒りによる影響が大きいかどうかは，感情又は信憑性判断が共有へ及ぼす影響を比較することで示された．

5.2.1. 仮説

要因（感情/信憑性判断）による共有への影響を明らかにするために，二つの仮説からなる検証モデルを設けた（図 5-2）．この仮説検証モデルは，過去の Disinformation キャ

ンペーン事例の特徴及び先行研究の結果に基づき構築された.「5.1.予備実験」において選定されたテキスト投稿刺激を用いて，比較要因である感情と信憑性判断を独立変数，共有意思を従属変数とすることで仮説が検証された.

第一の仮説は，感情は信憑性判断よりも共有意思への影響が大きいというものである（図 5-2 の仮説 1）．Disinformation の怒りを生み出す要因であるバイラル性が共有を促進するメカニズムを明らかにするために，2 つの要因（感情の強さ/信憑性判断）が共有意思へ及ぼす影響を測定する.「3.関連研究」で明らかになった怒りの共有メカニズムに基づくと，感情の強さによっては信憑性判断を経由せずに情報を共有する可能性がある．信憑性判断の方が，感情よりも共有意思へ大きな影響を与える場合，その情報を信じることによって共有につながる．これに対し，感情の方が，信憑性判断よりも共有意思へ大きな影響を与える場合，その情報に対して認識した感情によって共有につながる．その際，信憑性判断の影響は小さい又は影響を与えていないことを意味する．X において虚偽は真実の噂よりも共有されるという先行研究 [276]，及び過去事例において最も拡散した Disinformation は怒り等の感情に影響を与えることを狙ったものであった [4] ことを踏まえると，Disinformation の怒りを生み出す要因が信憑性判断よりも強く働き，信憑性に関わらず共有する傾向を強めていると考えられる.

第二の仮説は，怒りは信憑性判断を経て共有意思を促進するというものである（図 5-2 の仮説 2）．ソーシャルメディア環境で取得された実データの分析では，怒りと喜びは伝染性が高いことが分かっている [29]．怒りは虚偽かどうかに関わらず信憑性を高く評価させることによって共有を促進する傾向がある [305]．Disinformation の怒りを生み出す要因である不当性によって怒りを認識した場合，その怒りが強い場合は仮説 1 の信憑性判断を経由せずに共有する可能性があり，怒りが弱い場合は「信じる」という信憑性判断によって共有されている可能性がある.

なお，本実験の仮説では対象外とするものの，Disinformation の比較対象として True information への反応についても測定する．バイラル性が共有を促進する場合，強い感情が認識される True information は強い感情が情報を信じさせることによって共有が促進される可能性がある．反対に，弱い感情が認識される True information は情報に対する個人の受け止め方によって信憑性判断（信じる/信じない）及び共有意思（共有する/共有しない）は異なることが考えられる.

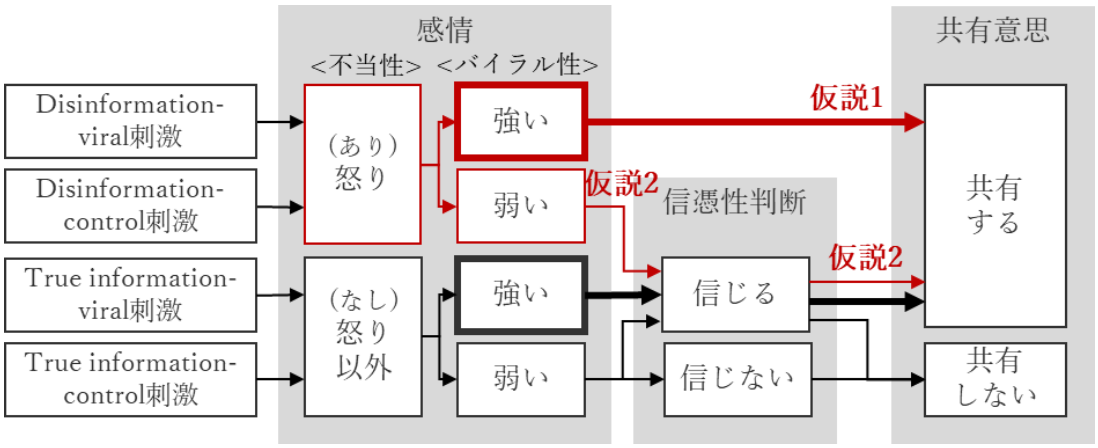


図 5-2 Disinformation の共有メカニズムに関する仮説モデル

5.2.2. 方法

(1) 参加者

参加者は日本在住の 300 人（男性 150 人，女性 150 人；年齢 $M = 44.5, SD = 13.5$）であり，2021 年 11 月に Web 調査会社を通じてオンライン調査モニターとして募集された．参加者には報酬としてリワードプログラムのポイントが付与された．スクリーニングにより，参加者は X を週 3 回以上利用するという条件を満たしていた．この条件は，総務省の調査でインターネットサービス利用者（$N = 10{,}000$）の 95%が少なくとも週 2〜3 回利用しているという利用頻度を参考に設定された [349]．実験はデータ収集前に研究倫理委員会の承認を得ており，学術目的の調査であることを説明した上で参加者から同意を得た．

(2) テキスト投稿刺激

「5.1.予備実験」で選定されたテキスト投稿刺激 10 個が使用された（付録 2 参照）．なお，世代間対立テーマの Disinformation-viral (older) 刺激については，予備実験から本実験までの実施期間の間に話題として取り上げていた新型コロナウイルス感染者数に変化があったため，時勢に合わせて微修正をした．文字数の平均は 141.9 文字（$SD = 9.15$）だった．

(3) 手続き

実験は，Web ベースのリサーチ会社によるアンケートを用いて実施された．参加者に

は，実験中に提示された投稿は全て実験者によって作成された架空のものであること，感情がおさまってから実験を開始することが教示された．これは，実験以外に起因する感情（例えば，実験前にイライラする出来事を体験した等）による回答への影響を可能な限り減らすことを意図していた．

実験を開始すると，参加者はよく利用しているソーシャルメディアのタイムラインにテキスト投稿刺激が表示されている状況をイメージするよう求められた．テキスト投稿刺激 10 個がランダムな順番で提示され，提示されるたびに参加者は以下 4 つの質問に回答するよう求められた．

1) 感情の種類：テキスト投稿刺激に対して認識した感情の種類について，Plutchik の「感情の輪」[352] から最も近い感情を 1 つ回答するよう求められた（回答選択肢：期待，喜び，信頼，恐怖，驚き，悲しみ，嫌悪，怒り，又は感情なし）．
2) 感情の強さ：何らかの感情を認識した場合，その強さを回答するよう求められた（回答選択肢：1.弱い〜4.強い）．なお，1) 感情の種類において「感情なし」と回答した参加者には当該質問項目は表示されなかった．
3) 共有意思：参加者は「あなたのタイムラインにこの投稿が表示されていたら共有すると思いますか？」という質問に対して回答するよう求められた（回答選択肢：共有する/共有しない）．
4) 信憑性判断：参加者は「もしこの投稿があなたのタイムラインに表示されたら，『実際の出来事』だと信じますか」という質問に対して回答するよう求められた（回答選択肢：信じる/信じない）．これは，人は真実か虚偽かという情報の正確さを問われると情報の精査が促され，共有選択時の真実の識別力が高まる [217] ためであった．

実験の最後に，参加者は関心のある日本の社会問題，及び実験後の感情（1＝ ポジティブ〜5 ＝ ネガティブ）を回答するよう求められた．参加者の属性情報（性別，年齢等）については，Web 調査会社より提供された．

(4) 分析

① テキスト投稿刺激の操作チェック

刺激の妥当性を確認するために，参加者が各テキスト投稿刺激に対して認識した感情の種類と強さを確認した．刺激条件の不当性は，Disinformation 刺激に対して怒りを認識

した参加者が True information 刺激よりも多いことで確認することができる．そのため，Disinformation 刺激又は True information 刺激に対して参加者が回答した感情の種類の割合を算出した．

刺激条件のバイラル性は，viral 刺激が control 刺激よりも強い感情が認識されたことにより確認することができる．そのため，感情の強さについてテキスト投稿刺激ごとに中央値と四分位偏差を算出し，差があるか検定した．感情の強さの差を比較するにあたり，Disinformation-viral 刺激と True information-viral 刺激を viral 刺激としてグループ化し，Disinformation-control 刺激と True information-control 刺激を control 刺激としてグループ化した．分析にはウィルコクソンの符号順位検定を行った．独立変数は刺激グループ（viral/control）であり，従属変数は感情の強さ（0.感情なし，1.弱い～4.強い）だった．

② 仮説検証のための分析

仮説 1 の解は，感情の強さと信憑性判断のどちらの要因が共有意思へ及ぼす影響が大きいかをテキスト投稿刺激ごとに確認することで得られる．バイラル性により感情を高める viral 刺激では，control 刺激よりも感情が強いため，感情の強さが信憑性判断よりも共有意思へ及ぼす影響が大きい可能性がある．さらに，虚偽は真実よりも拡散したという過去事例を踏まえると，Disinformation-viral 刺激は True information-viral 刺激よりも，この傾向が強くみられる可能性がある．仮説検証のため，ロジスティック回帰分析を実施した．ロジスティック回帰分析は，従属変数が 2 値（共有する/共有しない）である場合に，複数の独立変数から従属変数である 2 値の結果に及ぼす影響を予測することができる．従属変数は，共有意思（共有する/共有しない）だった．独立変数は，感情の強さ（0.感情なし，1.弱い～4.強い），信憑性判断（信じる/信じない）だった．

仮説 2 の解は，感情の種類ごとに参加者の信憑性判断と共有意思の回答割合を算出することで確認した．怒りは誤って信じることによって共有につながっている可能性があることから，怒りがそれ以外の感情よりも信じる/共有する割合が高いかどうかを確認した．不当性がある Disinformation 刺激は怒りが多く認識され，不当性がない True information 刺激は怒り以外の感情が認識される可能性がある．怒りが他の感情よりも信じる/共有する割合が高ければ，Disinformation 刺激は強い怒りを生み出す要因があるため True information 刺激よりも信じて共有されやすいことが明らかになる．

5.2.3.　結果

(1)　操作チェック

テキスト投稿刺激の不当性による違いを確認した結果，Disinformation 刺激は True information 刺激よりも怒りと嫌悪が多かった（図 5-3）．テキスト投稿刺激に対して認識した感情の種類は，Disinformation 刺激では怒り（16.4%），悲しみ（15.0%），又は嫌悪（12.9%）の感情が多く認識され，True information 刺激では期待（19.4%）又は悲しみ（13.0%）の感情が多く認識された．感情の種類によって認識された感情の強さは異なり，怒り（*Md* = 3, *QD* = 3-4）が最も強く，驚き（*Md* = 2, *QD* = 2-3）が最も弱かった．

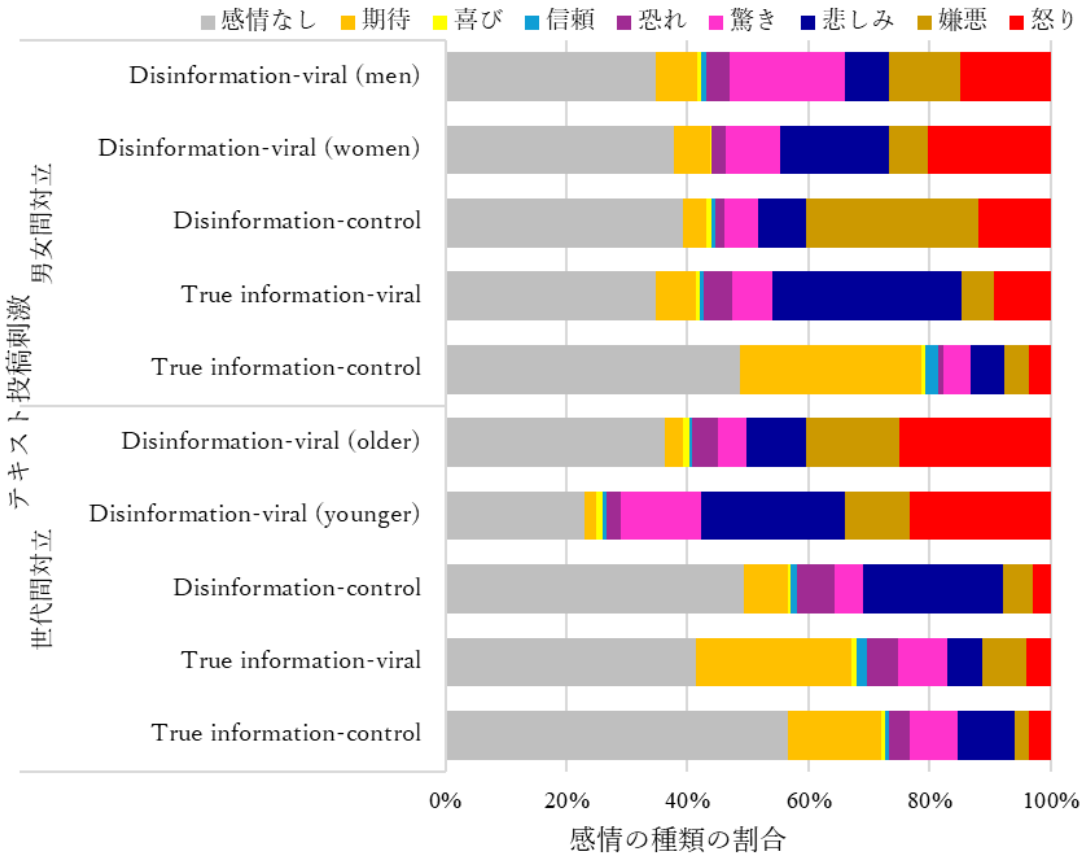


図 5-3 各テキスト投稿刺激で認識された感情の種類の割合

テキスト投稿刺激のバイラル性による違いを確認した結果，全体的に viral 刺激は control 刺激よりも感情が強かった．テキスト投稿刺激ごとに感情の強さの中央値と四分位偏差を算出したところ，viral 刺激（*Md* = 2-3）は control 刺激（*Md* = 0-2）よりも強い

感情の回答が有意に多かった（$p < .001$）（表 5-4）．しかし，男女間対立テーマのDisinformation-control 刺激は, バイラル性が低い刺激条件であるにも関わらず感情の強さの中央値が高く，感情を認識した参加者が感情を認識しなかった参加者よりも有意に多かった（$\chi^2(1) = 13.65, p < .001$）．このため，刺激条件に合致しなかった男女間対立テーマのテキスト投稿刺激を対象外とし，世代間対立テーマのテキスト投稿刺激のみを分析対象とした．

表 5-4 各テキスト投稿刺激の感情の強さ（中央値と四分位偏差）

テーマ	テキスト投稿刺激	*Md* (*QD*)
男女間対立	Disinformation-viral (men)	2.0 (0.00-3.00)
	Disinformation-viral (women)	2.0 (0.00-3.00)
	Disinformation-control	2.0 (0.00-3.00)
	True information-viral	2.0 (0.00-3.00)
	True information-control	1.0 (0.00-3.00)
世代間対立	Disinformation-viral (older)	2.0 (0.00-3.00)
	Disinformation-viral (younger)	3.0 (1.00-3.00)
	Disinformation-control	1.0 (0.00-3.00)
	True information-viral	2.0 (0.00-3.00)
	True information-control	0.0 (0.00-2.00)

実験後の感情（1= ポジティブ～5= ネガティブ）について中央値と四分位偏差を算出したところ，中央値は 3（*QD* = 3-4）であり極端な負への偏りはなかった．

(2) 共有を促進する感情の強さ

Disinformation-viral 刺激が共有されやすいか確認するために, テキスト投稿刺激ごとに信じる/共有すると回答した参加者の割合を算出したところ，Disinformation-viral (younger) 刺激（23.3%）が最も共有すると回答した人数の割合が多かったことが分かった（表 5-5）．反対に，共有すると回答した人数の割合が最も少なかったのは，Disinformation-control 刺激（10.0%）だった．Disinformation-viral (younger) 刺激は，参加者が認識した感情の強さが最も強く（表 5-4），参加者は強い感情を認識したテキスト投稿刺激を共有すると回答する傾向があった（表 5-6）．

表 5-5 テキスト投稿刺激ごとの信じる/共有する割合

テキスト投稿刺激	信じる(%)	共有する(%)
Disinformation-viral (older)	49.3	14.7
Disinformation-viral (younger)	70.3	23.3
Disinformation-control	62.7	10.0
True information-viral	50.3	15.3
True information-control	63.0	11.3
M (SD)	59.1 (8.1)	14.9 (4.6)

表 5-6 感情の強さごとの信じる/共有する割合

	感情なし	弱い	やや弱い	やや強い	強い
信じる(%)	42.7	66.7	67.4	69.7	81.1
共有する(%)	3.9	4.2	13.2	24.9	42.6

仮説 1 を検証した結果，3 個のテキスト投稿刺激において感情の強さが信憑性判断よりも共有意思により大きな影響を及ぼしていたことが分かった（表 5-7）．感情の強さと信憑性判断は，全てのテキスト投稿刺激において共有意思に有意に影響を及ぼす独立変数だった（いずれも $p<.05$）．感情の強さと信憑性判断のどちらがより共有意思に及ぼす影響が大きいか独立変数間で比較するために，標準化偏回帰係数（β）を算出した．共有意思に有意に影響を及ぼす独立変数のうち，標準化偏回帰係数の数値が大きい独立変数の方が，共有意思に及ぼす影響が大きいことを意味している．テキスト投稿刺激のうち，感情の強さが信憑性判断よりも標準化偏回帰係数が大きかったのは，Disinformation-viral (older) 刺激（0.983），Disinformation-control 刺激（0.821），及び True information-control 刺激（1.038）だった．一方，信憑性判断の方が感情の強さよりも標準化偏回帰係数が大きかったのは，Disinformation-viral (younger) 刺激（0.953），及び True information-viral 刺激（1.550）だった．

表 5-7 共有意思に影響を与える独立変数

テキスト投稿刺激	独立変数	β	OR (95%*CI*)	Hosmer-Lemeshow
Disinformation-viral (older)	感情の強さ	0.983***	1.9 (1.4-2.6)	0.824
	信憑性判断	0.960***	6.8 (2.5-18.3)	
Disinformation-viral (younger)	感情の強さ	0.780***	1.7 (1.3-2.3)	0.104
	信憑性判断	0.953***	8.0 (2.4-27.0)	
Disinformation-control	感情の強さ	0.821***	1.8 (1.3-2.4)	0.366
	信憑性判断	0.765*	4.9 (1.4-16.7)	
True information-viral	感情の強さ	0.935***	1.9 (1.4-2.6)	0.425
	信憑性判断	1.550***	22.1 (5.2-94.5)	
True information-control	感情の強さ	1.038***	2.2 (1.6-3.0)	0.669
	信憑性判断	0.825**	5.5 (1.6-19.1)	

* $p < .05$, ** $p < .01$, *** $p < .001$

(3) 共有を促進する感情の種類

仮説 2 を検証した結果，感情の種類のうち参加者が共有すると回答した割合が高いのは，喜び（58.3%）と怒り（30.5%）だった（$M = 22.58\%$, $SD = 14.99$）（図 5-4）．参加者が信じると回答した割合が高い感情の種類は，悲しみ（77.2%），怒り（75.1%），喜び（75.0%），及び期待（73.8%）だった（$M = 64.57\%$, $SD = 12.31$）．Disinformation-viral (older/younger) 刺激で認識された感情の回答割合は，怒りが 23〜25%で喜びが 1%だった．

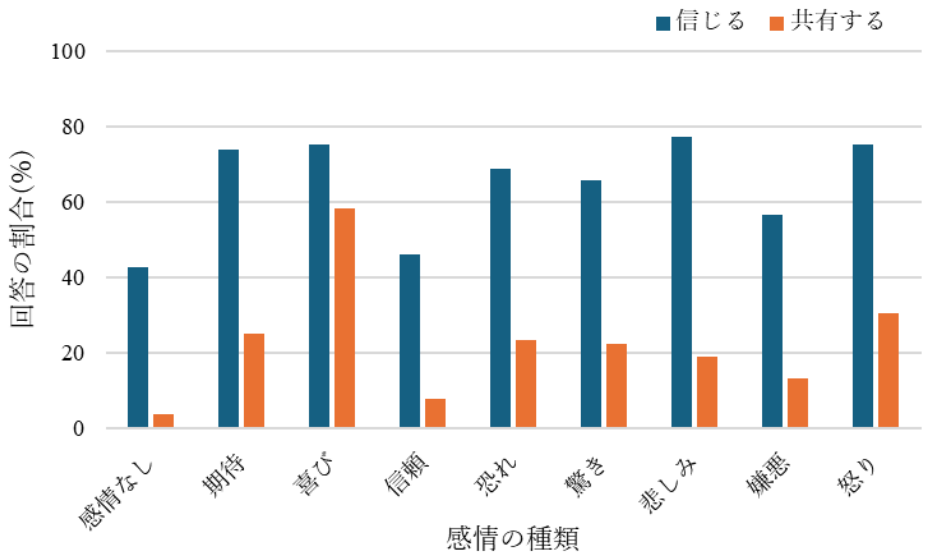


図 5-4 感情の種類ごとの信じる/共有する割合

5.2.4. 考察

Disinformationの怒りを生み出す要因が共有へ及ぼす影響について, 仮説モデル(図 5-2)を部分的に支持する結果が得られた(図 5-5). 仮説1は部分的に支持され, Disinformation-viral (older) 刺激を含む3つのテキスト投稿刺激において，感情の強さが信憑性判断よりも共有意思への影響が大きかった．仮説2は支持され，参加者が信じる/共有すると回答した割合が多かった感情の種類は，喜びと怒りだった．Disinformation-viral刺激は不当性から怒りを認識されることが多く，喜びを認識した参加者は少なかった．

(1) 共有を促進する感情の強さ

バイラル性により感情を高めるviral刺激は，信憑性判断よりも共有意思へ大きな影響を及ぼすことが確認されたが，viral刺激であれば必ず同じ傾向がみられるわけではなかった．Disinformation-viral (older) 刺激では，感情の強さが信憑性判断よりも共有意思への影響が大きかった．この結果は，Lewandowskyが予測した「人は情報の真理値に関わらず感情を刺激する情報を伝える可能性」[55]を支持している可能性が高い．しかし，同じくviral刺激であるDisinformation-viral (younger) 刺激は参加者が認識した感情の強さが最も強かったにも関わらず（表 5-4），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（表 5-7）．この違いは，参加者がviral刺激に対して認識した感情の種類に起因しているものと考えられる．Disinformation-viral (older) 刺激は怒りを認識した参加者が多かったのに対し，Disinformation-viral (younger) 刺激は悲しみを認識した参加者が多かった（図 5-3）．つまり，感情が信憑性に関わらず共有を促進するのは，Disinformationに対して強い怒りが認識された場合であることが示唆された(図 5-5①).

(2) 共有を促進する感情の種類

参加者が信じる/共有する割合が高い感情の種類は喜びと怒りであることから，怒りを生み出す要因を含むDisinformationは信じて共有されやすいことが示唆された（図 5-5②）．この結果は，Fanらによる「喜びと怒りは伝染性が高い」[29]というソーシャルメディアのデータ分析結果と一致している．Disinformation-viral (older) 刺激に対して参加者が認識した感情の種類は怒りが多く（図 5-3），怒りは信じる/共有する割合を高めた（図 5-4）.

これに対し，True information 刺激において多く認識された悲しみと期待は，信じる割合を高めるものの，共有する割合は低かった（図 5-4）（図 5-5④）．これは，Berger and Milkman による「悲しみを喚起した場合には共有される可能性が低くなった」[54] という実験結果と一致している．しかし，悲しみを認識した参加者が多い Disinformation-viral (younger) 刺激と，期待を認識した参加者が多い True information-viral 刺激は，共有すると回答した参加者の割合が高かった（表 5-5）．これは，悲しみ又は期待が，怒りと同じくらい強く認識された場合には，共有が促進される可能性を示唆している（図 5-5③）．つまり，共有するか否かは認識した感情の強さと種類によって異なり，テキスト投稿刺激に対して認識した感情が弱い場合（図 5-5②④），又は強くても悲しみ又は期待の場合（図 5-5③）は信憑性判断の方が感情よりも共有意思への影響が大きくなることが考えられる．この傾向は，朴と大坊による「悲しみが真偽性判断の正答率を高めた」[287] という実験結果でも示唆されている．

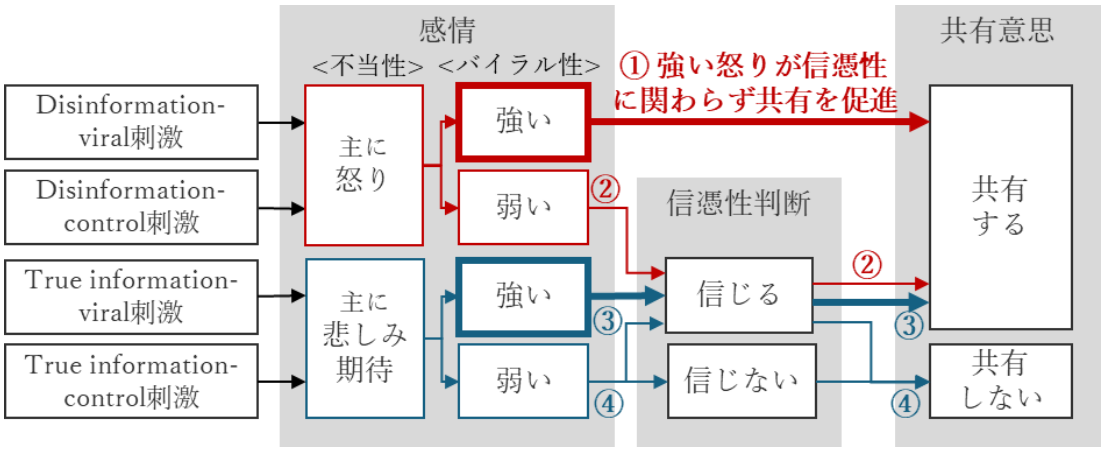


図 5-5 仮説検証結果に基づく Disinformation の共有メカニズム

5.3. 小括

本章では，Disinformation の怒りを生み出す要因による共有メカニズムを明らかにした．Disinformation の特徴である怒りを認識させる「不当性」と，公共性から感情の強さを高める「バイラル性」の 2 つの操作要因を設け，不当性による怒りとバイラル性による強い感情を認識させるテキスト投稿刺激を作成した．Disinformation の怒りを生み出す要因（不当性/バイラル性）による共有メカニズムを明らかにするために，共有意思に及ぼす影響を要因間（感情の強さ/信憑性判断）で比較検証した．

第一に，Disinformation に対して強い怒り（バイラル性＋不当性）を認識すると，信憑性に関わらず強い怒りのままに共有する傾向が見られた．

第二に，喜びと怒りを認識する情報は信じられて共有されやすかった．Disinformation は，不当性から怒りを認識するため，信じられて共有されやすい可能性が示された．

以上の結果から，Disinformation は怒りから「信じて共有するルート」と強い怒りから「信憑性に関わらず共有するルート」の 2 つの経路で共有が促進されることが考えられる．

6. 怒りに着目した情動調節ナッジの提案

本章では，Disinformation の共有を促進している強い怒りに着目し，ユーザによる Disinformation の感情的な共有を減らす有効な介入策を明らかにする．「5.2.本実験」の結果から，Disinformation の怒りを生み出す要因が共有を促進している可能性があることが分かった．このため，共有を促す怒りに対する介入策が，Disinformation の怒りによる共有の影響を減らす上で効果的である可能性がある．しかし，「2.Disinformation 対策の調査」では，ユーザが Disinformation を共有しようとした時に，Disinformation の共有を促進する強い怒りにアプローチすることで共有の再考を促すユーザ介入策はなかった．そこで，Disinformation の強い怒りによる感情的な共有を減らすために，Disinformation の強い怒りに対するユーザ介入策のデザインを考案し，その効果を確認する本実験を実施する．また，本実験で使用する介入デザインを選定するために，先行研究をもとに作成したいくつかの介入デザインを評価するための予備実験を行う．

6.1. 予備実験

予備実験の目的は，本実験で使用する介入デザインを選定することであった．Disinformation の強い怒りが共有を促進するにも関わらず，これまで Disinformation の共有を減らすために怒りに着目したユーザ介入策はなかった．このため，「3.関連研究」で得られた知見をもとに，強い怒りに対して有効性が見込める新たな介入デザインを考案し，Disinformation の共有を減らす効果があるか予備実験で確認した．介入デザインが参加者の共有意思や感情的な反応にもたらす変化を測定し，強い怒りへの有効性を評価することで，本実験で用いる介入デザインを選定した．

6.1.1. 方法

(1) 参加者

実験者が所属する大学院と企業で募集された 47 人（男性 30 人，女性 17 人；年齢 10 代～60 代）より回答を得た（2023 年 12 月）．一部の参加者は予備実験の目的を知っていたため，他者になりきって第三者的な視点で回答するロールプレイング形式とした．これは，実験者が得たい結果を想定して参加者が回答する可能性を抑えるためであった．実験はデータ収集前に研究倫理委員会の承認を得ており，学術目的の調査であることを

説明した上で参加者から同意を得た.

(2) テキスト投稿刺激

「5.2.本実験」で使用された 10 個のテキスト投稿刺激のうち，男女間対立テーマの Disinformation-viral (men) 刺激と Disinformation-viral (women) 刺激の 2 個が使用された（付録 2）．この 2 個を用いた理由は，男女の参加者が実験でなりきる同性の第三者の強い怒りをイメージしやすくするためであった.

また，「5.2.本実験」でのテキスト投稿刺激の操作チェック結果を受けて，より刺激条件に合致したテキスト投稿刺激を再作成して選定することとした．再作成されたのは男女間対立テーマの Disinformation-control 刺激と世代間対立テーマの Disinformation-viral (older) 刺激の 2 刺激である．それぞれの刺激条件（不当性×バイラル性）に，より合致するテキスト投稿刺激に絞り込んだものを本実験で使用するため，2 刺激とも 2 種類ずつ（A/B）の計 4 個を作成した．男女間対立テーマの Disinformation-control 刺激は，バイラル性が低い刺激条件だったが, 怒りを強く認識した参加者が多くみられた. このため, Disinformation-control 刺激として, 自己意識的な感情が多く組み込まれたものを再作成した．これは，他者を非難する感情語を多く含む虚偽の噂は真実の投稿よりも拡散し，自己意識的な感情語を多く含む虚偽の噂は拡散しにくいという先行研究の知見に基づく[353]. 世代間対立テーマの Disinformation-viral (older) 刺激は，新型コロナ禍における若者の問題行動に関する内容であったが，新型コロナの収束という社会情勢の変化に応じて，別の内容へと変更した.

(3) 介入デザインの作成

Disinformation の強い怒りによる感情的な共有を減らすという目的が達成されるよう，Disinformation の強い怒りに対して有効性が見込める介入デザインとして情動調節ナッジを作成した．情動調節ナッジは，投稿コンテンツに対する感情評価において，怒りのスコアが高いものをユーザが共有しようとした時に表示される仕組みとした．怒りのスコアが高い投稿コンテンツという条件は，先行研究 [276] 及び「5.2.本実験」の結果に基づいて設定された．このような条件設定の場合，怒りのスコアが高い真実の投稿コンテンツに対しても，ユーザが共有しようとした時に情動調節ナッジが表示される．ここで重要なことは，本研究で問題視しているのは「怒りによってユーザが有害な意思決定を誘

発」[282] され，「自身の幸福につながる行動が阻害される」[31] ことである．投稿コンテンツが表現する強い怒りに対して注意を向けることは，ユーザが自身の感情や行動を振り返ることにつながり，自分の状況を踏まえた上での理性的な行動（共有/キャンセル）をするのに役立つと考えられる．すなわち，投稿コンテンツが真実であっても，情動調節ナッジは自身の強い怒りを振り返り，幸福につながる理性的な行動を選択するのに役立つことが考えられる．あくまでもユーザにとって望ましい意思決定を導くことが重要であり，ユーザによる「投稿コンテンツを共有したい」「情動調節ナッジを無視したい」という自律性や表現の自由を尊重するものとした．

本研究では X に投稿された Disinformation への対処を想定していることから，情動調節ナッジは X が 2020 年 6 月に導入した機能 [137] をベースにして作成された．この機能は，ユーザがニュース記事を読まずにリポスト（共有）ボタンを押した時にポップアップウィンドウが表示されるという介入デザインだった．ユーザへ共有する前に元の記事を読むよう促した上で，再度リポスト（共有）と引用（コメントを追加して共有）の選択ボタンが提供された．キャンセルの選択ボタンはなく，ポップアップウィンドウの画面外をタップすると共有を中止できるという仕組みだった．実験において回答選択肢等の目新しさや理解の難しさがユーザに与える影響を最小限にするため，この既存の機能を模して情動調節ナッジが作成された．情動調節ナッジは，感情情報の円グラフ（図 6-1①），情動調節メッセージ（図 6-1②），及び回答選択肢ボタン（図 6-1③）から構成される．

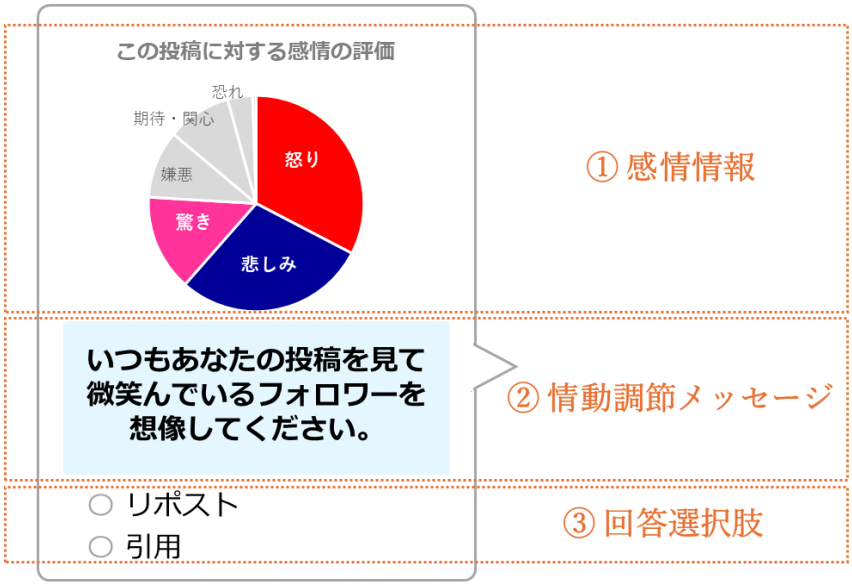


図 6-1 情動調節ナッジの構成要素

① 感情情報

感情情報は，強い怒りにユーザの注意を向けさせることを目的としていたことから，投稿コンテンツに含まれる感情の種類と割合を円グラフで視覚化したものとした．これは，投稿コンテンツ内の感情的な要素をより認識しやすくするだけで，情動調節が実現する可能性があることが示唆されているからである[291]．

本研究では，先行研究における既存の情動調節を促すデザインを応用し，新たに強い怒りを対象としたナッジを作成した．Kiskola ら[291,292]と Syrjämäki ら[293]の先行研究は，テキストに感情表現が含まれているということをユーザに認識させるだけであり，感情の詳細については明確に示されていなかった．ユーザに投稿コンテンツの強い怒りに気付かせるためには，感情の種類とその割合を具体的にユーザへ示す必要がある．このため，本研究では投稿コンテンツに含まれる感情の種類と割合を，説得力が高いとされる円グラフ[329]でユーザへ提示することとした．これは，ユーザの共有行動に関連する技術を構築する際には，視覚的手法を優先することが推奨されている[190]からである．

感情情報の元となるデータは，「5.2.本実験」において参加者 300 人から得られた回答データが使用された（表 6-1）．円グラフでは，参加者の回答のうち多かった上位 3 つの感情の種類がその感情を表現する色[334]でハイライト表示された．例えば，各感情に割り当てられた色は，怒り＝赤，嫌悪＝濃い黄色，悲しみ＝藍色，驚き＝明るいピンクだった．上位 4 つ目以降の感情はグレーアウトで表示された．

表 6-1 各 Disinformation-viral 刺激の感情情報

テキスト投稿刺激	1 番目に多い感情	2 番目に多い感情	3 番目に多い感情
Disinformation-viral (men)	驚き（29.1%）	怒り（23.0%）	嫌悪（17.9%）
Disinformation-viral (women)	怒り（32.6%）	悲しみ（28.9%）	驚き（14.4%）

② 情動調節メッセージ

より怒りに対する効果を高めるため，感情情報と共に怒りに対する有効性が見込まれる情動調節戦略を用いたメッセージを提示した．情動調節メッセージは，情動調節戦略に基づいて理性的な状態をユーザに思い出させるナッジであり，ユーザが感情のままに投稿コンテンツを共有してしまうことを防ぐ可能性がある．

本研究では，先行研究において無礼なユーザ投稿行動を減らした情動調節メッセージを応用し，新たにユーザの共有行動を対象とした怒りに効果があるテキストのナッジを作成した．このアプローチは，グラフィカルな視覚化は共にテキストメッセージを提供することで強化できるという考えに基づいている [330,335]．Syrjämäki らによる先行研究において，ニュースサイトのコメント欄にユーザが無礼なコメントを書き込んだ時に，情動調節メッセージを表示することで無礼なコメントの投稿が減ったことが明らかになっている [293]．このため，ユーザの共有行動に対しても情動調節メッセージは共有を減らす可能性が見込まれた．さらに，本研究では，より怒りへの効果を高めるために，怒りに有効な情動調節戦略を用いた情動調節メッセージを用いた．

情動調節メッセージの種類と内容は先行研究の実験において使用された文章を参考に，3 個の気晴らし，2 個の再評価，2 個の視点取得，1 個の共感的対応メッセージが作成された（付録 3）．これら 8 個の情動調節メッセージに感情情報のみを表示する 1 個を加えて，計 9 個の情動調節ナッジが作成された（表 6-2）．なお，情動調節戦略における行動の抑制は，ユーザが共有しようとした時にポップアップウィンドウが表示されるというフリクションによる共有行動の一時停止として実装された．情動調節ナッジにより，感情情報と自分の感情基準との不一致（例えば，強すぎる感情や誤った感情の種類が想起された場合 [292]）に気づくと，情動調節メッセージをヒントにしてユーザは共有を再考する可能性がある．

表 6-2 情動調節メッセージの種類と概要

情動調節戦略	ID	情動調節メッセージの概要
感情情報のみ	A1	（メッセージなし）
気晴らし	D1	ポジティブな状況を想像する
	D2	他者（投稿者）への思いやりを想像する
	D3	他者（フォロワー）への思いやりを想像する
再評価	R1	状況をポジティブな変化につなげる
	R2	感情刺激そのものの解釈を変える
視点取得	P1	他者に自己を取り入れる
	P2	客観的・分析的視点で見る
共感的対応	E1	閲覧者自身への思いやりを伝える

(4) 手続き

実験は，Google フォームを使用したアンケート形式で行われた．参加者には，実験中

に提示された投稿は全て実験者によって作成された架空のものであること，感情がおさまってから実験を開始することが教示された．これは，実験以外に起因する感情（例えば，実験前にイライラする出来事を体験した等）による回答への影響を可能な限り減らすことを意図していた．

実験を開始すると，参加者はいつものようにXにログインした際に表示されるホーム画面のおすすめタイムラインを見ている状況をイメージするよう求められた．実験者が再作成した男女間対立テーマの Disinformation-control 刺激 2 個と Disinformation-viral (older) 刺激 2 個の計 4 個が順番に提示され，提示されるたびに参加者は以下 2 つの質問に回答するよう求められた．

1) 感情の種類：テキスト投稿刺激に対して認識した感情の種類について，Plutchik の「感情の輪」[352] から最も近い感情を 1 つ回答するよう求められた（回答選択肢：期待，喜び，信頼，恐怖，驚き，悲しみ，嫌悪，怒り，又は感情なし）．
2) 感情の強さ：何らかの感情を認識した場合，その強さを回答するよう求められた（回答選択肢：1.弱い〜4.強い）．なお，1) 感情の種類において「感情なし」と回答した参加者には当該質問項目は表示されなかった．

次に，参加者は「テキスト投稿刺激を見て強い怒りからリポストボタンを押した山田さん」という架空の同性の人物になりきって回答するよう教示された．男性の参加者には Disinformation-viral (men) 刺激が提示され，女性の参加者には Disinformation-viral (women) 刺激が提示された．続けて，参加者は「リポストボタンを押すとポップアップ画面が表示されました」という説明と共に，実験者が作成したナッジデザイン 9 個が昇順又は降順のいずれかの順番で提示された．提示されるたびに，参加者は以下の 2 つの質問に回答するよう求められた．

1) 共有意思の変化：参加者は「投稿文を読んで強い怒りを感じていた山田さんは，この表示を見た後にどのボタンをクリックすると思いますか？」という質問に対して回答するよう求められた（回答選択肢：リポスト/引用/キャンセル）．
2) 感情の強さの変化：参加者は「この表示を見た時，山田さんの怒りは投稿文を読んですぐの状態と比較して変化したと思いますか？投稿文を読んだ時に感じた怒りを 5 としてお答えください」という質問に対して回答するよう求められた（回答選択肢：0. 怒りが弱まった〜10. 怒りが強まった）．

実験の最後に，参加者は年代，Xの利用頻度，及び実験後の感情（1= ポジティブ〜5

= ネガティブ）を回答するよう求められた．

6.1.2. 結果

(1) 情動調節ナッジによる効果の検証

Disinformation-viral 刺激の共有を減らす効果が高いと評価されたナッジデザインは，客観的・分析的視点で見る視点取得ナッジ（P2）だった．各ナッジデザインに対する共有意思の回答傾向をみるため，ナッジ提示後の各回答（リポスト/引用/キャンセル）の割合を算出した（表 6-3）．キャンセルの回答割合が最も高かったのは，客観的・分析的視点で見る視点取得ナッジ（P2）（66.0%）とフォロワーへの思いやりを想像する気晴らしナッジ（D3）（48.9%）だった．反対に，ナッジ提示後もリポストする回答の割合が最も高かったのは感情情報のみのナッジ（A1）（61.7%）だった．Disinformation-viral 刺激の共有を減らす効果が高いナッジデザインを明らかにするために，ナッジ提示後の共有意思（リポスト/引用/キャンセル）を独立変数とし，それぞれの回答人数を従属変数とするカイ二乗検定を行った．その結果，客観的・分析的視点で見る視点取得ナッジ（P2）は，キャンセルすると回答した参加者がリポストすると回答した参加者よりも有意に多かった（$\chi^2(2) = 24.553, p < .01$）．

表 6-3 各ナッジ提示後の共有意思の回答傾向

情動調節戦略	ID	リポスト（%）	引用（%）	キャンセル（%）
なし	A1	61.7	19.1	19.1
気晴らし	D1	44.7	17.0	38.3
	D2	51.1	21.3	27.7
	D3	42.6	8.5	48.9
再評価	R1	27.7	42.6	29.8
	R2	55.3	27.7	17.0
視点取得	P1	53.2	17.0	29.8
	P2	25.5	8.5	66.0
共感的対応	E1	40.4	34.0	25.5

Disinformation-viral 刺激に対して認識した強い怒りを弱める効果が高いと評価されたナッジは，客観的・分析的視点で見る視点取得ナッジ（P2）とフォロワーへの思いやりを想像する気晴らしナッジ（D3）だった．各ナッジデザインにおけるナッジ提示後の感情の強さの変化（0-10）の回答傾向をみるため，ナッジデザインごとに中央値と四分位偏

差を算出した（図 6-2）．提示後に怒りが弱まったのは，客観的・分析的視点で見る視点取得ナッジ（P2）（*Md* = 4, *QD* = 2.0-5.0）とフォロワーへの思いやりを想像する気晴らしナッジ（D3）（*Md* = 4, *QD* = 2.0-5.5）だった．反対に，提示後に怒りが強まったのは，閲覧者自身への思いやりを伝える共感的対応ナッジ（E1）だった（*Md* = 5, *QD* = 5-6）．Disinformation-viral 刺激に対する強い怒りを弱める効果が高いナッジデザインを明らかにするために，ナッジ提示後の感情の強さの変化（弱まった/強まった）を独立変数とし，それぞれの回答人数を従属変数とするカイ二乗検定を行った．その結果，客観的・分析的視点で見る視点取得ナッジ（P2）とフォロワーへの思いやりを想像する気晴らしナッジ（D3）では，感情の強さが弱まったと回答した参加者が，強まったと回答した参加者よりも有意に多かった（P2: $\chi^2(1) = 18.778, p < .001$; D3: $\chi^2(1) = 4.568, p < .05$）．

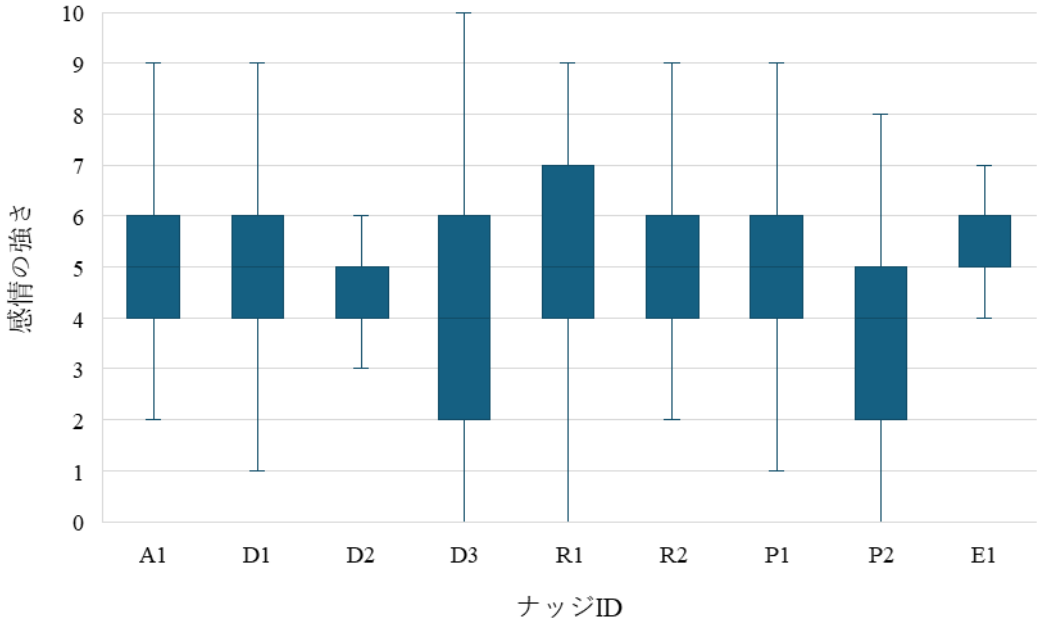


図 6-2 ナッジ提示後に参加者が回答した感情の強さ（ナッジ提示前＝5）

(2) 再作成したテキスト投稿刺激の回答傾向

再作成した Disinformation-viral (older) 刺激は，不当性から怒りを認識した参加者がみられ，バイラル性においては Disinformation-control 刺激よりも感情が強く認識された．感情の種類の割合について算出したところ，男女間対立テーマの Disinformation-control 刺激は嫌悪（14.9-17.0%）を認識した参加者が多かった（図 6-3）．世代間対立テーマの

Disinformation-viral (older) 刺激は悲しみ(38.3-51.1%)が最も多かったが, 怒り(4.3-6.4%)を認識した参加者もみられた．刺激条件のバイラル性による違いを確認するために感情の強さ（0-4）の中央値と四分位偏差を算出して比較した結果，Disinformation-viral (older) 刺激（$Md = 2.0$）は，Disinformation-control 刺激（$Md = 1.0$）よりも強い感情の回答が多い傾向がみられた．感情を認識した参加者の人数と感情なしと回答した参加者の人数に有意な差があるかどうかを検証することとした．各テキスト投稿刺激に対する感情の強さ(1-4/0)を独立変数とし, それぞれの回答人数を従属変数とするカイ二乗検定を行った．その結果，バイラル性がある Disinformation-viral (older) 刺激は感情を認識した参加者の人数が，感情なしと回答した参加者の人数よりも有意に多かった（いずれも $p < .001$）(表6-4)．

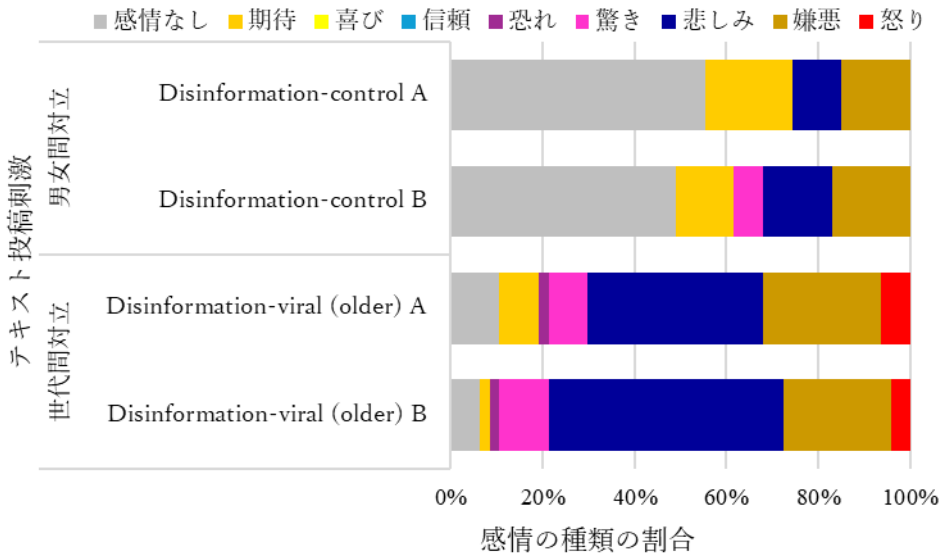


図 6-3 再作成した各テキスト投稿刺激で認識された感情の種類の割合

表 6-4 再作成したテキスト投稿刺激の中央値（四分位偏差）とカイニ乗値

テーマ	テキスト投稿刺激	A	B
男女間対立	Disinformation-control	1.0 (0.00-1.00)	1.0 (0.00-2.00)
		0.53	0.02
世代間対立	Disinformation-viral (older)	2.0 (1.00-2.00)	2.0 (1.00-3.00)
		29.13***	35.77***

* $p < .05$, ** $p < .01$, *** $p < .001$

6.1.3. 考察

(1) 情動調節ナッジによる効果の検証

感情情報と情動調節メッセージを組み合わせたナッジデザインは，感情情報のみのナッジデザイン（A1）よりも，Disinformation-viral 刺激の共有を減らすのに効果的であると評価された．具体的には，客観的・分析的視点で見る視点取得ナッジ（P2）が最も Disinformation-viral 刺激の共有を減らす効果があると評価された．さらに，フォロワーへの思いやりを想像する気晴らしナッジ（D3）と客観的・分析的視点で見る視点取得ナッジ（P2）は，感情の強さが弱まると評価されたことから，強い怒りに対して効果的である可能性がある．これらの知見は，情動調節方略のプロセス間効果量に基づいて感情を管理するための効果的な方法を特定した Webb ら [299] の知見と一致している．

しかし，予備実験にはいくつかの限界があった．第一に，予備実験の回答は，参加者が強い怒りを認識した第三者の人物を想像することに依存しており，参加者自身の怒りに伴う共有への介入効果は検証されなかった．第二に，強い怒りをイメージするよう強調した指示が，感情の強さの結果に何らかのバイアスをもたらした可能性がある．これらの限界に対処するためには，効果がみられた 2 個のナッジデザインが，参加者自身の感情に伴う共有にも効果的であるかどうかを明らかにする必要がある．

(2) 再作成したテキスト投稿刺激の選定

再作成したテキスト投稿刺激はいずれも刺激条件に合致していたため，2 種類（A/B）のいずれか 1 個を本実験で使用することとした．男女間対立テーマの Disinformation-control 刺激は，感情の強さの中央値等が低かった方を選定した．世代間対立テーマの Disinformation-older (older) 刺激は，刺激条件が対象とする属性において感情の強さの中央値等が高かった方を採用することとした．

6.2. 本実験

本実験の目的は，作成した情動調節ナッジが，従来のナッジよりも Disinformation の共有を減らす効果が高いか比較評価により明らかにすることだった．「2.4.5.ナッジ」にて前述した通り，従来のナッジは情報の正確さに注目させる [50]，あるいは共有行動を一時停止させる [138] ことによって，ユーザの共有を抑制しようとするものである．これらは怒りに関する対策ではないため，強い怒りによる Disinformation の共有に対抗するには不十

分である可能性がある．そこで，予備実験で強い怒りへの効果を確認した情動調節ナッジを用いて，従来のナッジよりも Disinformation の共有を減らすかどうかを本実験で確認した．ナッジの効果は，テキスト投稿刺激の提示後（すなわちナッジ提示前）とナッジ提示後に参加者の反応がどのように変化したかを比較評価することで示された．

6.2.1. 仮説

情動調節ナッジの有効性を評価するために，Disinformation の共有を減らす効果について既存ナッジと比較評価する 2 つの仮説を設けた．比較に用いた既存ナッジは，X が導入した「リポストよりも引用を奨励する機能」[138] をモデルに，実験者によって作成されたものだった．この機能は，リポストボタンを押すとポップアップウィンドウが表示されるというフリクションによってユーザの共有行動を一時停止し，コメントを加える理由や内容について考える時間を提供するものだった．X は，この機能によりリポストが 23%減少し，引用が 26%増加したと報告している [140]．

第一の仮説は，情動調節ナッジ及び既存ナッジの提示後は，ナッジの提示前よりも Disinformation の共有が少ないというものである．情動調節ナッジ及び既存ナッジは，ユーザがリポストボタンを押して共有しようとした際に，ポップアップウィンドウが表示されるというフリクションが用いられていた．このフリクションの効果により，ナッジ提示後は提示前よりもリポストが減ることが予測された．

第二の仮説は，情動調節ナッジは，既存ナッジよりも Disinformation の共有を減らす効果が高いというものである．既存ナッジがフリクションによって共有を減らす効果を持つのに対し，情動調節ナッジにはフリクションと怒りに有効な情動調節戦略の両方が組み込まれていた．フリクションは再考時間を提供するだけだが，情動調節ナッジは感情情報と情動調節メッセージを用いて再考に必要な熟慮を促す．このため，情動調節ナッジは，フリクションのみの既存ナッジよりも，情動調節効果によって Disinformation の共有を減らす効果が高い可能性がある．

6.2.2. 方法

(1) 参加者

参加者は日本在住の 400 人（男性 200 人，女性 200 人；年齢 $M = 44.5, SD = 13.6$）であり，2024 年 1 月に Web 調査会社を通じてオンライン調査モニターとして募集された．参

加者には報酬としてリワードプログラムのポイントが付与された．スクリーニングにより，参加者はXを週3回以上利用し，Xのログイン時に表示されるおすすめタイムラインでコンテンツをリポストしたことがあるという2つの条件を満たしていた．1つ目の条件は，総務省の調査結果[349]を参考に設定された．2つ目の条件は，実験で評価するナッジのデザインがXのインタフェースに基づいて作成されていたため，参加者はXを日常的に利用しており，ボタン名等の用語とそれに関連する動作を理解している必要があった．実験はデータ収集前に研究倫理委員会の承認を得ており，学術目的の調査であることを説明した上で参加者から同意を得た．

(2) テキスト投稿刺激

「5.2.本実験」で使用されたテキスト投稿刺激8個と「6.1.予備実験」で選定したテキスト投稿刺激2個の計10個が使用された（付録4）．テキスト投稿刺激のうち，Disinformation-viral 刺激をリポストするとナッジが表示されるという仕組みを表現するため，テキスト投稿刺激はXのポスト画面を模した画像で表示された．画像の表示要素は，投稿者のアイコン，投稿者名，ユーザID，投稿日，及びテキスト投稿文により構成された．投稿者に関する情報の影響を最小限にするため，投稿者のアイコンはフリー素材の動物の絵で統一され，投稿者名とユーザIDは日本人に多い名字の平仮名とアルファベット表記とした．投稿日は全て1日前に統一された．実際のXのタイムラインにはそれ以外の表示要素として，リツイートアイコン，いいねアイコン，インプレッション数があるが，これらの数値はメッセージの信憑性と共有意思へ強く影響し[354]，仮説検証に影響を及ぼす可能性があることから除外された．

(3) ナッジデザイン

実験で表示される情動調節ナッジには，フォロワーへの思いやりを想像する気晴らしナッジ（D3）（以下，気晴らしナッジとする）と客観的・分析的視点で見る視点取得ナッジ（P2）（以下，視点取得ナッジとする）の2個が使用された．これらは，「6.1.予備実験」において，Disinformation-viral 刺激に認識した怒りを弱めると評価されたことから，Disinformationの怒りを生み出す要因に対して効果的である可能性があった．感情情報として円グラフで表示される内容は，「5.2.本実験」の参加者300人から得られたデータが使用され，「6.1.予備実験」で再作成されたDisinformation-viral (older) 刺激は予備実験の

参加者 47 人から得られたデータが使用された（表 6-5）．円グラフでは，参加者の回答のうち多かった上位 3 つの感情の種類がその感情を表現する色 [334] でハイライト表示された．例えば，各感情に割り当てられた色は，怒り＝赤，嫌悪＝濃い黄色，悲しみ＝藍色，驚き＝明るいピンクだった．上位 4 つ目以降の感情はグレーアウトで表示された．

表 6-5 各 Disinformation-viral 刺激の感情情報

テキスト投稿刺激	1 番目に多い感情	2 番目に多い感情	3 番目に多い感情
Disinformation-viral (men)	驚き（29.1%）	怒り（23.0%）	嫌悪（17.9%）
Disinformation-viral (women)	怒り（32.6%）	悲しみ（28.9%）	驚き（14.4%）
Disinformation-viral (older)	怒り（39.3%）	嫌悪（24.1%）	悲しみ（15.7%）
Disinformation-viral (younger)	悲しみ（30.7%）	怒り（30.3%）	驚き（17.3%）

既存ナッジは X の機能 [138] をモデルに実験者によって作成されたものであり，リポストボタンを押すと「コメントを追加」というメッセージが書かれた投稿画面がポップアップ表示されるというデザインだった．提供される選択ボタンはリポストのみであり，キャンセルはポップアップウィンドウの画面外をタップするという仕組みだった（付録 5）．

(4) 手続き

実験は，Web ベースの調査会社によるアンケートを用いて実施された．参加者には，実験中に提示された投稿は全て実験者によって作成された架空のものであること，感情がおさまってから実験を開始することが教示された．これは，実験以外に起因する感情（例えば，実験前にイライラする出来事を体験した等）による回答への影響を可能な限り減らすことを意図していた．

実験を開始すると，参加者はいつものように X にログインした際に表示されるホーム画面のおすすめタイムラインを見ている状況をイメージするよう求められた．参加者は，男女間対立又は世代間対立の 2 テーマのいずれかに無作為に割り当てられ，5 個のテキスト投稿刺激が昇順又は降順のいずれかで提示された．提示順は，True information-viral 刺激，Disinformation-viral (men/older) 刺激，True information-control 刺激，Disinformation-viral (women/younger) 刺激，Disinformation-control 刺激，又はその逆の順序だった．参加

者はテキスト投稿刺激が提示されるたびに，以下 4 つの質問に回答するよう求められた．

1) 共有意思：参加者は「あなたのタイムラインにこの投稿が表示されていたらリポストすると思いますか？」という質問に対して回答するよう求められた（回答選択肢：リポストする/リポストしない）．
2) 感情の種類：テキスト投稿刺激に対して認識した感情の種類について，Plutchik の「感情の輪」[352] から最も近い感情を 1 つ回答するよう求められた（回答選択肢：期待，喜び，信頼，恐怖，驚き，悲しみ，嫌悪，怒り，又は感情なし）．
3) 感情の強さ：何らかの感情を認識した場合，その強さを回答するよう求められた（回答選択肢：1.弱い～10.強い）．なお，2) 感情の種類において「感情なし」と回答した参加者には当該質問項目は表示されなかった．
4) 信憑性判断：参加者は「もしこの投稿があなたのタイムラインに表示されたら，『実際の出来事』だと信じますか」という質問に対して回答するよう求められた（回答選択肢：信じる/信じない）．

Disinformation-viral 刺激の後，参加者は気晴らしナッジ，視点取得ナッジ，又は既存ナッジの 3 つがランダムに提示された．参加者はナッジが提示されるたびに，以下 4 つの質問に回答するよう求められた．なお，Disinformation-viral 刺激を「リポストしない」と回答した参加者にもナッジ 3 つがランダムに提示され，リポストしていた場合を想定して質問に回答するよう指示された．

1) 共有意思の変化：参加者は「このポップアップ表示を見て，あなたはどのボタンをクリックすると思いますか？」という質問に対して回答するよう求められた（回答選択肢：リポスト/引用/キャンセル）．なお，既存ナッジは表示されている選択ボタンがリポストのみであるため，引用に該当する回答選択肢は「コメントを追加してリポストボタンを押す」とした．
2) 感情の強さの変化：参加者は「このポップアップ表示を見て、あなたの感情に変化はありましたか？」という質問に対して回答するよう求められた（回答選択肢：1.弱い～10.強い）．
3) 感情の種類の変化：参加者は「このポップアップ表示を見て，あなたの感情の種類に変化はありましたか？」という質問に対して Plutchik の「感情の輪」[352] から最も近い感情を 1 つ回答するよう求められた（回答選択肢：期待，喜び，信頼，恐怖，驚き，悲しみ，嫌悪，怒り，又は感情に変化なし）．

4) 信憑性判断の変化：参加者は「このポップアップ表示を見て、投稿文が『実際にあった出来事だ』と信じる気持ちに変化はありましたか？」という質問に回答するよう求められた（回答選択肢：信じる/信じない）.

実験の最後に，参加者は関心のある日本の社会問題，虚偽尺度項目，及び実験後の感情（1 = ポジティブ〜5 = ネガティブ）を回答するよう求められた．参加者の属性情報（性別，年齢等）については，Web 調査会社より提供された.

(5) 分析

① 分析のためのデータ準備

回答の精度を担保するため，スクリーニング項目を逆転した虚偽尺度が設けられていた．これらの項目で回答が一致しなかった参加者は 0 人であったため，全ての参加者の回答データを分析対象とした.

② テキスト投稿刺激の操作チェック

刺激の妥当性を確認するために，参加者が各テキスト投稿刺激に対して認識した感情の種類と強さを確認した．刺激条件の不当性は，Disinformation 刺激に対して怒りを認識した参加者が True information 刺激よりも多いことで確認することができる．そのため，Disinformation 刺激又は True information 刺激に対して参加者が回答した感情の種類の割合を算出した.

刺激条件のバイラル性は，viral 刺激が control 刺激よりも強い感情が認識されたことにより確認することができる．そのため，感情の強さについてテキスト投稿刺激ごとに中央値と四分位偏差を算出し，差があるか検定した. 感情の強さの差を比較するにあたり，Disinformation-viral 刺激と True information-viral 刺激を viral 刺激としてグループ化し，Disinformation-control 刺激と True information-control 刺激を control 刺激としてグループ化した．分析にはウィルコクソンの符号順位検定を行った．独立変数は刺激グループ（viral/control）であり，従属変数は感情の強さ（0.感情なし，1.弱い〜10.強い）だった.

③ 仮説検証のための分析

仮説 1 の解は，ナッジの提示前と後で Disinformation-viral 刺激を共有すると回答した人数をナッジ間で比較することで得られる．X の報告からフリクションによりリポスト

が 23%減少したことが明らかになっている [140] ため，仮説検証においても同様に，Disinformation-viral 刺激を共有する回答者数が，ナッジ（気晴らし/視点取得/既存）提示後に減ることが予測される．分析にあたり，ナッジ提示前の共有意思の回答はリポストする=1，リポストしない=0 という数値データに変換された．同様に，ナッジ提示後の共有意思の回答は，リポスト又は引用=1，キャンセル=0 という数値データに変換された．仮説検証は，コクランの Q 検定を行った．独立変数はナッジの提示前と後（気晴らし/視点取得/既存）における Disinformation-viral 刺激の共有意思であり，従属変数はその回答者数だった．

仮説 2 の解は，ナッジ提示後に Disinformation-viral 刺激を共有すると回答した人数をナッジ間で共有することで得られる．気晴らしナッジと視点取得ナッジは，フリクションに加えて情動調節戦略が組み込まれているため, 既存ナッジよりも Disinformation-viral 刺激を共有する回答者が減ることが予測された．分析にあたり，ナッジ提示後の共有意思の回答は，リポスト又は引用=1，キャンセル=0 という数値データに変換された．仮説検証は，コクランの Q 検定を行った．独立変数はナッジ提示後（気晴らし/視点取得/既存）における Disinformation-viral 刺激の共有意思であり，従属変数はその回答者数だった．

6.2.3. 結果

(1) 操作チェック

テキスト投稿刺激の不当性による違いにおいて Disinformation 刺激は True information 刺激よりも嫌悪と怒りが多く，バイラル性による違いにおいて viral 刺激は control 刺激よりも感情が強かった. テキスト投稿刺激に対して認識した感情の種類は, Disinformation 刺激では悲しみ（13.8%），嫌悪（11.3%）又は怒り（6.3%）の感情が多く認識され，True information 刺激では期待（12.0%）又は悲しみ（10.1%）の感情が多く認識された（図 6-4）．テキスト投稿刺激ごとに感情の強さの中央値と四分位偏差を算出したところ，viral 刺激（Md＝0-3）は control 刺激（Md＝0）よりも強い感情の回答が有意に多かった（p＜.001）（表 6-6）. 感情の種類によって認識された感情の強さは異なり，信頼（Md＝7, QD＝5.25-9）と怒り（Md＝7, QD＝5-9）が最も強く，驚き（Md＝5.5, QD＝4-7）が最も弱かった．

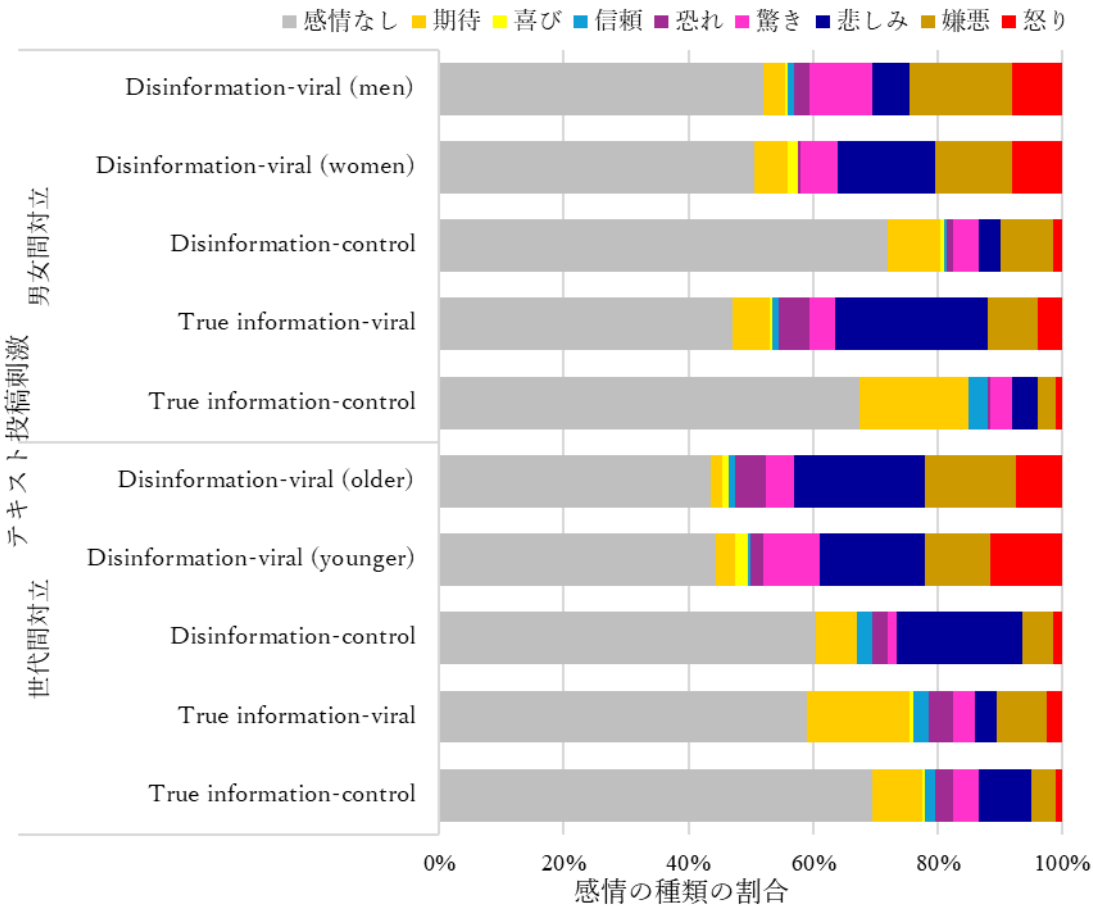


図 6-4 各テキスト投稿刺激で認識された感情の種類の割合

表 6-6 各テキスト投稿刺激の感情の強さ（中央値と四分位偏差）

テーマ	テキスト投稿刺激	*Md* (*QD*)
男女間対立	Disinformation-viral (men)	0.0 (0.00-6.00)
	Disinformation-viral (women)	0.0 (0.00-6.00)
	Disinformation-control	0.0 (0.00-3.00)
	True information-viral	1.5 (0.00-6.00)
	True information-control	0.0 (0.00-4.25)
世代間対立	Disinformation-viral (older)	3.0 (0.00-6.00)
	Disinformation-viral (younger)	3.0 (0.00-6.25)
	Disinformation-control	0.0 (0.00-5.00)
	True information-viral	0.0 (0.00-6.00)
	True information-control	0.0 (0.00-4.00)

実験後の感情（1=ポジティブ～5=ネガティブ）について中央値と四分位偏差を算出したところ，中央値は 3（*QD* = 3-4）であり極端な負への偏りはなかった．

(2) 全般的な回答傾向

Disinformation-viral 刺激が共有されやすいか確認するために, テキスト投稿刺激ごとに共有すると回答した参加者の割合を算出したところ, Disinformation-viral 刺激を共有する回答の割合は各テーマにおいて平均又は平均以上だった（表 6-7）. 男女間対立テーマで共有すると回答した人数の割合の平均は 6.8%（*SD*＝2.2）であり, そのうち Disinformation-viral (men) 刺激は 8.0%, Disinformation-viral (women) 刺激は 6.0%と平均程度だった. 世代間対立テーマで共有すると回答した人数の割合の平均は 14.9%（*SD* = 3.3）であり, Disinformation-viral (older) 刺激は 13.0%と平均に近く, Disinformation-viral (younger) 刺激は 21.2%と高かった.

表 6-7 テキスト投稿刺激ごとの信じる/共有する割合

テーマ	テキスト投稿刺激	信じる(%)	共有する(%)
男女間対立	Disinformation-viral (men)	40.0	8.0
(*n*=200)	Disinformation-viral (women)	50.0	6.0
	Disinformation-control	46.0	4.0
	True information-viral	50.0	5.5
	True information-control	50.0	10.5
世代間対立	Disinformation-viral (older)	47.0	13.0
(*n*=200)	Disinformation-viral (younger)	58.5	21.2
	Disinformation-control	61.0	13.0
	True information-viral	40.0	14.9
	True information-control	52.0	12.4

(3) フリクションが共有に及ぼす効果

全てのナッジにおいて, ナッジ提示後は提示前よりも Disinformation-viral 刺激を共有すると回答した人数が減ったことが分かった. ナッジ提示前に Disinformation-viral 刺激を共有すると回答した参加者のうち, ナッジ提示後にも共有（リポスト又は引用）すると回答した参加者の割合は, 気晴らしナッジが 66.3%, 視点取得ナッジが 69.8%, 及び既存ナッジが 77.9%だった.

仮説 1 を検証した結果, ナッジ提示前と後で共有すると回答した人数に有意な差があることが明らかになった（$Q = 50.9$, $df = 3$, $p < .001$）. どのナッジで有意な差がみられたのかを確認するため, ボンフェローニ法による多重比較をした. その結果, 全てのナッジが, ナッジ提示前と比較して Disinformation-viral 刺激を共有すると回答した人数を有

意に減らしたことが示された（$p < .001$）.

(4) 情動調節が共有に及ぼす効果

仮説 2 を検証した結果，気晴らしナッジが既存ナッジよりも Disinformation-viral 刺激の共有を減らす効果が高いことが明らかになった. Disinformation-viral 刺激を共有すると回答した人数は，ナッジ間で有意な差があることが分かった（$Q = 6.87$, $df = 2$, $p < .05$）. どのナッジが有意に共有を減らしたのかを確認するため，ボンフェローニ法による多重比較をした. その結果，気晴らしナッジが，既存ナッジよりも Disinformation-viral 刺激を共有すると回答した人数を有意に減らしたことが示された（$p < .05$）. 一方で，視点取得ナッジと既存ナッジの間に有意な差はなかった.

ナッジがどの Disinformation-viral 刺激に効果があったのか確認したところ，気晴らしナッジは全ての Disinformation-viral 刺激において共有すると回答した人数を減らしたことが分かった（図 6-5）. 仮説 1 の分析をテキスト投稿刺激ごとに実施した結果，気晴らしナッジは Disinformation-viral 刺激 4 個においてナッジ提示前よりも共有すると回答した人数を有意に減らした（いずれも $p < .05$）. 視点取得ナッジは，Disinformation-viral (older/younger) 刺激 2 個においてナッジ提示前よりも共有すると回答した人数を有意に減らした（いずれも $p < .05$）. 既存ナッジは，Disinformation-viral (younger) 刺激のみにおいてナッジ提示前よりも共有すると回答した人数を有意に減らした（$p < .05$）. これらの結果から，いずれのナッジも Disinformation-viral 刺激の共有を減らすが，その効果はユーザに提示されたナッジの表示内容によって異なることが示唆された.

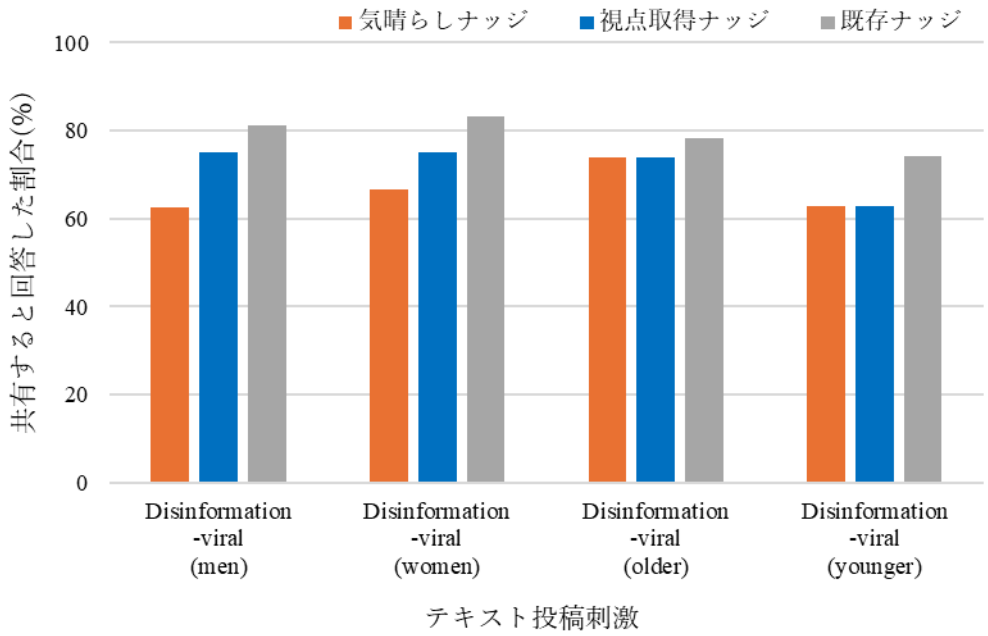


図 6-5 ナッジ提示前に共有すると回答した参加者のうち，ナッジ提示後も共有（リポスト又は引用）すると回答した参加者の Disinformation-viral 刺激ごとの割合

ナッジ提示前に Disinformation-viral 刺激を共有すると回答した参加者において，テキスト投稿刺激に対して認識していた感情の強さを弱めたのは気晴らしナッジと既存ナッジだった．Disinformation-viral 刺激を共有すると回答した参加者のナッジ提示前と後における感情の強さの回答について，中央値と四分位偏差を算出した．全てのナッジにおいて，Disinformation-viral 刺激に対して認識した感情の強さ（$Md = 7, QD = 6\text{-}8.75$）は，ナッジ提示後（$Md = 6, QD = 5\text{-}8$）に弱まっていた．ナッジ提示前と後での感情の強さに有意な差があるかどうかを確認するために，独立変数をナッジ提示前と後（気晴らし/視点取得/既存）における感情の強さ（0-10），従属変数をその回答者数としたフリードマン検定を行った．その結果，感情の強さに有意な差が認められた（$\chi^2(3) = 22.617, p < .001$）ため，ボンフェローニ法による多重比較をした．ナッジ提示前と比較して感情の強さを有意に弱めたのは，気晴らしナッジと既存ナッジだった（いずれも $p < .05$）．

ナッジによって Disinformation-viral 刺激に対して認識した感情の強さが弱まった参加者を対象に，感情が弱まっても共有（リポスト又は引用）すると回答する可能性について調べたところ，既存ナッジでは共有を継続する参加者が多いことが分かった．ナッジ提示後に感情が弱まった参加者が Disinformation-viral 刺激の共有をキャンセルすると回

答した割合は，気晴らしナッジ 41.5%，視点取得ナッジ 39.4%，及び既存ナッジ 19.5%だった（図 6-6）．ナッジごとに，独立変数をナッジ提示後の参加者の共有意思（リポスト又は引用/キャンセル），従属変数をその回答者数としたカイ二乗検定を行った．その結果，ナッジ提示後に感情が弱まった参加者において，気晴らしナッジと視点取得ナッジでは共有とキャンセルの回答者数に有意な差はなかった．しかし，既存ナッジはナッジ提示後に感情が弱まった参加者において，共有する回答者がキャンセルする回答者よりも有意に多かった（$\chi^2(2) = 5.308, p < .05$）．

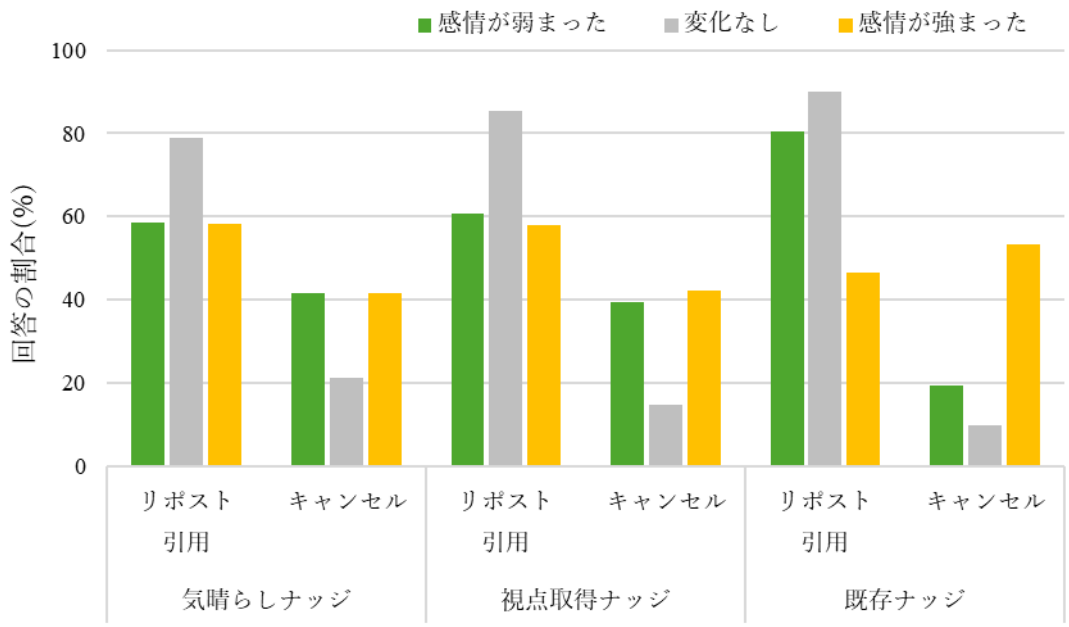


図 6-6 Disinformation-viral 刺激を共有する回答者のナッジ提示後の共有傾向

気晴らしナッジは, Disinformation-viral 刺激に対して認識した感情をポジティブな感情に変化させる傾向がややみられた. Disinformation-viral 刺激に対して恐怖, 悲しみ, 嫌悪, 又は怒りといったネガティブな感情を認識して共有すると回答した参加者のうち，ナッジ提示後に期待，喜び，又は信頼といったポジティブな感情に変化した回答の割合は，気晴らしナッジでは 1.5%だったのに対し，視点取得ナッジと既存ナッジではいずれも 0.8%だった．

6.2.4. 考察

実験者が作成した気晴らしナッジが, Disinformation-viral 刺激の共有を減らすのに最も有効であることが示された．仮説 1 が支持され，全てのナッジは Disinformation-viral 刺激の共有を減少させる効果があることが分かった．仮説 2 は部分的に支持され，Disinformation-viral 刺激の共有を減らす効果についてのナッジ間での比較では, 気晴らしナッジが既存ナッジよりも有意に共有を減らす効果がみられた．

気晴らしナッジは Disinformation-viral 刺激の共有を減らす効果が最も高く，その効果は 4 個全ての Disinformation-viral 刺激においてみられた．気晴らしナッジでは，フォロワーに対する思いやりを想起させるようにデザインされた情動調節メッセージが使用されていた．参加者は,「あなたの投稿を見ていつも微笑んでいるフォロワー」を想像するよう促されたため, Disinformation-viral 刺激の感情情報（怒り, 悲しみ, 嫌悪, 又は驚き）との不一致[292]が生じた．この不一致が情動調節の必要性を引き起こし[320], 参加者が情動調節メッセージをヒントに熟慮するよう促した可能性がある．また，参加者はこれらのネガティブな感情を共有することで，自分のフォロワーに不快な思いをさせるかもしれないといった想像が働いた可能性がある．全ての Disinformation-viral 刺激において共有を減らす効果がみられたという結果は，現実世界における様々な Disinformation に対しても同様の効果が期待できる可能性がある．

既存ナッジは Disinformation-viral 刺激の共有を減らす効果があるものの，その効果は気晴らしナッジよりも小さかった．既存ナッジも含め，全てのナッジでみられた Disinformation-viral 刺激の共有を減らす効果は，フリクション[215]による効果である可能性がある．このことから, 気晴らしナッジの追加的な効果は, フリクションだけでなく, 感情情報と情動調節メッセージが含まれていることから生じてた可能性が示唆される．さらに，既存ナッジではナッジ提示後に感情が弱まった参加者の間で，Disinformation-viral 刺激の共有を継続する傾向がみられた．この傾向を二重過程理論（直感的認知プロセス/熟慮的認知プロセス）で説明すると，気晴らしナッジは情動調節により熟慮的認知プロセスが促されることで[355]共有が再考されたのに対し，既存ナッジは感情に伴う直感的認知プロセスに依存したまま[286,287]共有した可能性がある．

一方，視点取得ナッジは Disinformation-viral 刺激の共有を減らす効果があるものの, その効果は既存ナッジと差はなかった．視点取得ナッジでは，客観的で分析的な視点により投稿内容の背後にある意図について考えさせるようにデザインされた情動調節メッ

セージが使用されていた. Disinformation-viral 刺激の意図に参加者の注意を向けさせることは，参加者が自分の感情に集中するのを妨げ，感情の不一致を検出しにくくした可能性がある [292]．その結果，情動調節の必要性 [320] が活性化されず，情動調節メッセージを用いた熟慮が促されなかったと考えられる．また，視点取得ナッジでは，共有するかどうかを決定する際に他者のことを考慮する必要はなかった．気晴らしナッジと視点取得ナッジの効果の差から推測すると，他者に対するユーザの意識を高め，Disinformation を共有することの潜在的な影響を想像するよう促すことが，共有を減らすために効果的な戦略である可能性がある．

6.3. 小括

本章では，Disinformation の怒りに着目した情動調節ナッジを作成し，従来のナッジよりも Disinformation の共有を減らす効果が高いことを明らかにした．情動調節ナッジは，投稿コンテンツに対する感情評価において，怒りのスコアが高いものをユーザが共有しようとした時に表示される仕組みとした. X の既存のデザインをモデルに, Disinformation の感情情報，情動調節メッセージ，及び回答選択肢ボタンから構成される情動調節ナッジを作成した．情動調節メッセージを 9 個作成して予備実験を行ったところ，客観的・分析的視点で見る「視点取得ナッジ」とフォロワーへの思いやりを想像する「気晴らしナッジ」の 2 個が Disinformation の強い怒りに効果があると評価された．この 2 個の情動調節ナッジが Disinformation の共有を減らす効果について，X のナッジをモデルに作成した既存ナッジと比較評価をした．

第一に，いずれのナッジも Disinformation の共有を減らしたが，その中でも気晴らしナッジが Disinformation の共有を減らすのに最も効果的であることが分かった．

第二に，同じ情動調節ナッジである視点取得ナッジは既存ナッジと有意な差がなかったことから，他者と社会的影響に対する認識を高めることが Disinformation の共有を効果的に減少させる可能性があることが示唆された．

以上の結果から，共有行動を対象に介入するナッジ同士において，新たに提案する怒りに着目した介入策である気晴らしナッジが現対策よりも効果的であることが示された．

7. 情動調節ナッジと教育の比較評価

本章では，Disinformation の怒りを生み出す要因に対して，情動調節ナッジが既存のメディア情報リテラシー教育（以下，教育とする）を補完する介入策として有効であることを実験により明らかにする．5 章において，Disinformation の強い怒りは信憑性判断よりも共有意思への影響が大きいことが分かった．このため，Disinformation の強い怒りに着目することが Disinformation の共有を抑制する上で有用である可能性があると考え，6 章で強い怒りに対して有効な情動調節ナッジを作成した．この怒りに着目した情動調節ナッジは，Disinformation の怒りによって効果が限定し得る現対策を補完する策として有用な可能性がある．6 章では現対策の 1 つであるナッジ同士で比較評価をすることによって，強い怒りに対して有効な情動調節が Disinformation の共有を減らす効果があることを確認した．本章では，怒りに関する Disinformation 対策として現在最も使用されている教育と比較評価をすることにより，Disinformation の怒りの共有メカニズムに対する対策としての限界と有効性を確認する．

7.1. 仮説

情動調節ナッジと教育が Disinformation の共有を減らす効果を比較評価するために，検証モデル（図 7-1）を設計して二つの仮説を設けた．仮説検証モデルはこれまでの実験結果をもとに構築されたものであり，Disinformation の強い怒りから共有を決定するまでのプロセスの中で介入策がどのように作用するか予測した．情動調節ナッジ，教育，及びいずれの介入もなし（統制条件）の場合に，独立変数である感情と信憑性判断，従属変数である共有意思がどのように異なるか仮説を用いて検証された．

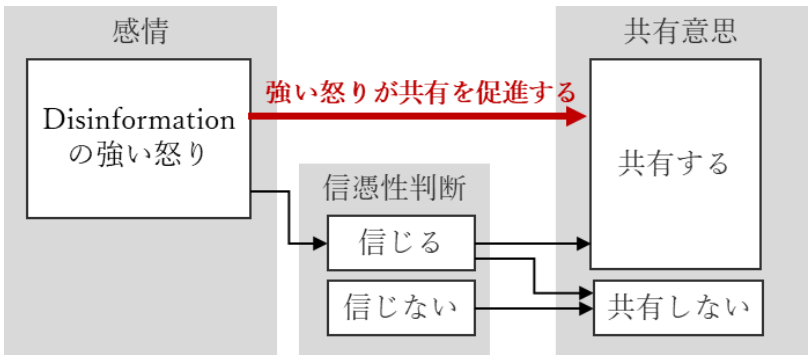


図 7-1 仮説の検証モデル（介入なしの場合）

第一の仮説は，情動調節ナッジは教育よりも Disinformation の共有を減らし，教育は介入なしよりも Disinformation の共有を減らすというものである．情動調節ナッジは，怒りのスコアが高い投稿コンテンツをユーザが共有しようとした時に表示される．すなわち，Disinformation の怒りを生み出す要因が共有を促進した時に，情動調節ナッジはユーザの共有行動に介入する（図 7-2）．強い怒りが信憑性判断に必要な熟慮を妨げて共有を促進するが，情動調節ナッジはユーザを感情に注目させて振り返ることにより熟慮を促すため，信憑性判断に基づく共有の再考につながる可能性がある．一方，教育は教育効果によって信憑性判断が促進される場合と，Disinformation の怒りを生み出す要因によって共有が促進される場合の両方がある可能性がある（図 7-3）．強い怒りは信憑性判断に必要な熟慮を妨げる可能性があり，その場合教育効果が十分に発揮されず強い怒りのままに共有する可能性がある．いずれの介入もない（統制条件）場合は，Disinformation の怒りを生み出す要因が信憑性判断に必要な熟慮を妨げることによって共有が促進されることが「5.2.本実験」の結果から考えられる（図 7-1）．したがって，教育は効果が発揮される場合があることから Disinformation の共有を減らす効果は介入なしよりも大きく，強い怒りによって効果が発揮されない場合があることから Disinformation の共有を減らす効果は情動調節ナッジよりも小さいと予測した．

第二の仮説は，情動調節ナッジでは信憑性判断が感情よりも共有意思への影響が大きく，教育では感情が信憑性判断よりも共有意思への影響が大きいというものである．この予測は「5.2.本実験」の結果に基づいており，強い怒りを認識させた Disinformation-viral 刺激では，感情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．情動調節ナッジは怒りに効果がある情動調節戦略を用いているため，情動調節効果によって Disinformation の怒りを生み出す要因による影響が小さくなる可能性がある．この感情による共有意思への影響が小さくなることによって，信憑性判断による共有意思への影響が大きくなることが考えられる．これに対し，教育は Disinformation に接触する前に行われるものであり，Disinformation の怒りを生み出す要因が共有を促進した時には介入は行われない．教育効果によって信憑性判断が促進される場合と，Disinformation の怒りを生み出す要因によって共有が促進される場合があり，前者の場合は信憑性判断により共有しない可能性がある．後者の場合，Disinformation の怒りを生み出す要因が熟慮を妨げることで信憑性判断が十分に行われず感情のままに共有してしまう可能性がある．この場合，感情が信憑性判断よりも共有意思への影響が大きくなると考えられる．

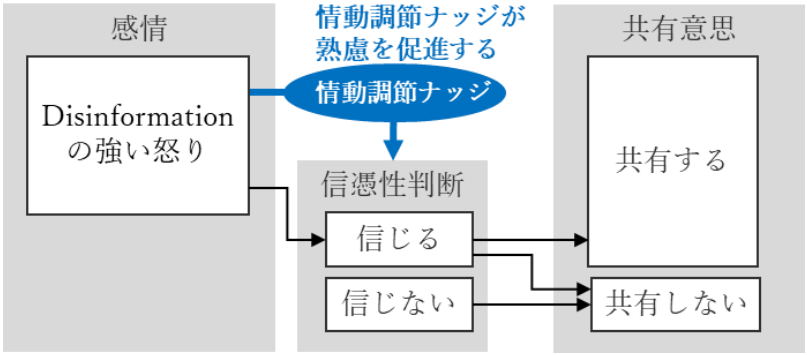


図 7-2 仮説の検証モデル（情動調節ナッジの場合）

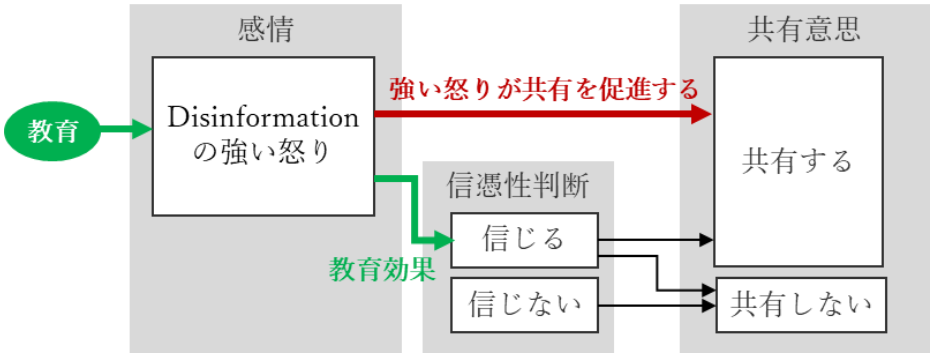


図 7-3 仮説の検証モデル（教育の場合）

7.2. 方法

(1) 参加者

参加者は日本在住の 300 人（男性 150 人，女性 150 人；年齢 $M = 44.6, SD = 13.5$）であり，2025 年 2 月に Web 調査会社を通じてオンライン調査モニターとして募集された．参加者には報酬としてリワードプログラムのポイントが付与された．スクリーニングにより，参加者は X のログイン時に表示されるおすすめタイムラインでコンテンツをリポストしたことがあり，これまでリテラシーや情報モラルについて学んだことがないという という 2 つの条件を満たしていた．1 つ目の条件は，実験で評価するナッジのデザインが X のインタフェースに基づいて作成されていたため，参加者は X を日常的に利用しており，ボタン名等の用語とそれに関連する動作を理解している必要があった．2 つ目の条件は，教育及びナッジによる効果を正確に測定するために，参加者の事前知識レベルを統制することを目的に設定された．実験はデータ収集前に研究倫理委員会の承認を得ており，学術目的の調査であることを説明した上で参加者から同意を得た．

(2) 実験条件

3 条件（ナッジ/教育/統制）の実験条件が設けられた．第一に，ナッジ条件では，「6.2.本実験」で有効性が評価された気晴らしナッジが使用された．実験は「6.2.本実験」と同じく，ユーザが怒りのスコアが高い投稿，すなわち Disinformation-viral 刺激を共有した時にポップアップ表示される仕組みが想定された．表示される要素は，熟慮を促すための感情情報，情動調節メッセージ，及び回答選択肢ボタン（リポスト/引用）だった．感情情報は円グラフで感情の種類と割合が視覚化されたものであり，「6.2.本実験」と同じデータが使用された（表 7-1）．感情の種類は Plutchik の感情の輪 [352] に従って，期待，喜び，信頼，恐怖，驚き，悲しみ，嫌悪，怒り，又は感情なしに分類されており，参加者の回答で多かった上位 3 つの感情の種類がその感情を表現する色 [334] でハイライト表示された．例えば，各感情に割り当てられた色は，怒り＝赤，嫌悪＝濃い黄色，悲しみ＝藍色，驚き＝明るいピンクだった．上位 4 つ目以降の感情はグレーアウトで表示された．情動調節メッセージは，「あなたの投稿を見ていつも微笑むフォロワーを想像してください」という気晴らしメッセージが表示された．

表 7-1 各 Disinformation-viral 刺激の感情情報

テキスト投稿刺激	1 番目に多い感情	2 番目に多い感情	3 番目に多い感情
Disinformation-viral (men)	驚き（29.1%）	怒り（23.0%）	嫌悪（17.9%）
Disinformation-viral (women)	怒り（32.6%）	悲しみ（28.9%）	驚き（14.4%）
Disinformation-viral (older)	怒り（39.3%）	嫌悪（24.1%）	悲しみ（15.7%）
Disinformation-viral (younger)	悲しみ（30.7%）	怒り（30.3%）	驚き（17.3%）

第二に，教育条件では，実験前に日本の総務省が公開した教育教材「インターネットとの向き合い方～ニセ・誤情報にだまされないために～」の初版 [246] から抜粋したパワーポイントスライドと解説文のセットが参加者へ提示された．この教材は，欧州委員会による教材「Spot and Fight Disinformation」[244] とベルギーの非営利団体による教材「GET YOUR FACTS STRAIGHT!」[240] を参照して作成された．事例は日本人になじみ深いものへ変更され，有識者検討会での意見が効果的に反映されていた．総務省主導による教材の効果検証テストでは，講座後は講座前よりもテストの平均点が上がったことが報告さ

れた[247]．全 66 ページの教材のうち実験で使用されたのは，用語の定義（P.7, 10, 13, 15），チェックの基本（P.40-44），チェックの応用（P.46-48, P.50-51），及びまとめ（P.55-56）の計 16 ページだった．

最後に，統制条件ではナッジも教育資料も提示されなかった．この条件は，ナッジ及び教育条件の効果を評価するためのベースラインとして設定された．

(3) テキスト投稿刺激

「6.2.本実験」と同じテキスト投稿刺激 10 個が使用された（付録 4）．

(4) 手続き

実験は，Web ベースのリサーチ会社によるアンケートを用いて実施された．参加者には，実験中に提示された投稿は全て実験者によって作成された架空のものであること，感情がおさまってから実験を開始することが教示された．これは，実験以外に起因する感情による回答への影響を可能な限り減らすことを意図していた．

実験を開始すると，参加者はいつものように X にログインした際に表示されるホーム画面のおすすめタイムラインを見ている状況をイメージするよう求められた．参加者は 3 条件（ナッジ/教育/統制）に無作為に割り当てられ，教育条件の参加者へは教育教材が 1 ページずつ順番に提示された（図 7-4）．次に，各条件の参加者は男女間対立又は世代間対立の 2 テーマのいずれかのテキスト投稿刺激 5 個がランダムに提示された．参加者はテキスト投稿刺激が提示されるたびに，以下 4 つの質問に回答するよう求められた．

1) 共有意思：参加者は「あなたのタイムラインにこの投稿が表示されていたらリポストすると思いますか？」という質問に対して回答するよう求められた（回答選択肢：リポストする/リポストしない）．
2) 感情の種類：テキスト投稿刺激に対して認識した感情の種類について，Plutchik の「感情の輪」[352] から最も近い感情を 1 つ回答するよう求められた（回答選択肢：期待，喜び，信頼，恐怖，驚き，悲しみ，嫌悪，怒り，又は感情なし）．
3) 感情の強さ：何らかの感情を認識した場合，その強さを回答するよう求められた（回答選択肢：1.弱い〜10.強い）．なお，2) 感情の種類において「感情なし」と回答した参加者には当該質問項目は表示されなかった．
4) 信憑性判断：参加者は「もしこの投稿があなたのタイムラインに表示されたら，

『実際の出来事』だと信じますか」という質問に対して回答するよう求められた（回答選択肢：信じる/信じない）.

ナッジ条件に割り当てられた参加者のうち, Disinformation-viral 刺激を共有すると回答した場合のみ，続けて気晴らしナッジが提示された．参加者はナッジが提示されるたびに，以下 4 つの質問に回答するよう求められた.

1) 共有意思の変化：参加者は「このポップアップ表示を見て，あなたはどのボタンをクリックすると思いますか？」という質問に対して回答するよう求められた（回答選択肢：リポスト/引用/キャンセル）.
2) 感情の強さの変化：参加者は「このポップアップ表示を見て、あなたの感情に変化はありましたか？」という質問に対して回答するよう求められた（回答選択肢：0.感情なし～10.強い）.
3) 感情の種類の変化：参加者は「このポップアップ表示を見て，あなたの感情の種類に変化はありましたか？」という質問に対して Plutchik の「感情の輪」[352] から最も近い感情を 1 つ回答するよう求められた（回答選択肢：期待，喜び，信頼，恐怖，驚き，悲しみ，嫌悪，怒り，又は感情に変化なし）.
4) 信憑性判断の変化：参加者は「このポップアップ表示を見て、投稿文が『実際にあった出来事だ』と信じる気持ちに変化はありましたか？」という質問に回答するよう求められた（回答選択肢：信じる/信じない）.

実験の最後に，参加者は教育資料の効果検証テスト [247]，関心のある日本の社会問題，虚偽尺度項目，及び実験後の感情（1=ポジティブ～5=ネガティブ）を回答するよう求められた．参加者の属性情報（性別，年齢等）については，Web 調査会社より提供された.

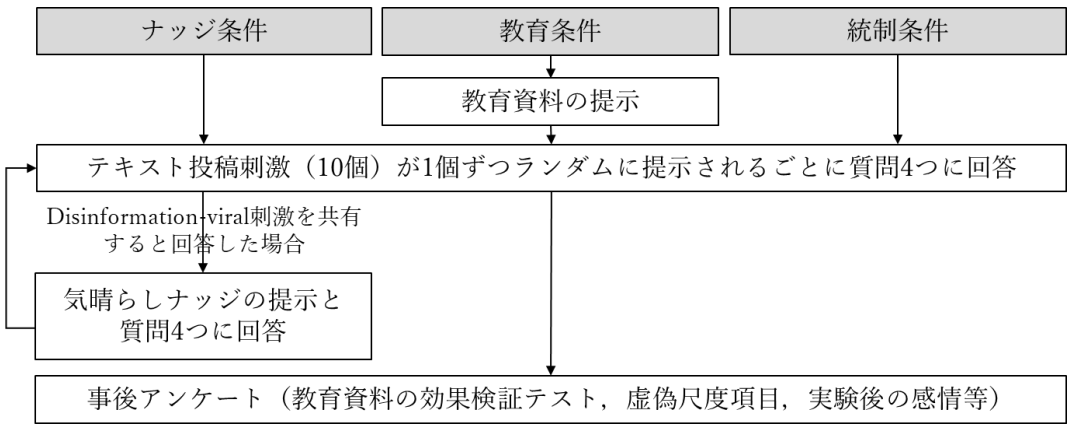


図 7-4 実験手続きの流れ

(5) 分析

① 分析のためのデータ準備

回答の精度を担保するため，スクリーニング項目を逆転した虚偽尺度が設けられていた．これらの項目で回答が一致しなかった参加者 10 人分のデータを除外し，最終的な有効サンプルは 290 人（男性 144 人，女性 146 人；年齢 $M = 44.2$，$SD = 13.5$）の参加者で構成された．

② テキスト投稿刺激の操作チェック

刺激の妥当性を確認するために，参加者が各テキスト投稿刺激に対して認識した感情の種類と強さを確認した．刺激条件の不当性は，Disinformation 刺激に対して怒りを認識した参加者が True information 刺激よりも多いことで確認することができる．そのため，Disinformation 刺激又は True information 刺激に対して参加者が回答した感情の種類の割合を算出した．

刺激条件のバイラル性は，viral 刺激が control 刺激よりも強い感情が認識されたことにより確認することができる．そのため，感情の強さについてテキスト投稿刺激ごとに中央値と四分位偏差を算出し，差があるか検定した．感情の強さの差を比較するにあたり，Disinformation-viral 刺激と True information-viral 刺激を viral 刺激としてグループ化し，Disinformation-control 刺激と True information-control 刺激を control 刺激としてグループ化した．分析にはウィルコクソンの符号順位検定を行った．独立変数は刺激グループ（viral/control）であり，従属変数は感情の強さ（0.感情なし，1.弱い〜10.強い）だった．

③ 教育の効果測定

事後アンケートで取得した教育資料の効果検証テストの正答率を実験条件（ナッジ/教育/統制）ごとに算出した．これは，教育条件の正答率と総務省の効果検証テストの結果[247]とを比較して考察するためだった．教育条件の効果を測定するためには，参加者が教育資料の内容を理解した上で Disinformation-viral 刺激の共有意思を回答している必要があった．教育条件の正答率が，ナッジ及び統制条件よりも高ければ，教育の効果により Disinformation-viral 刺激の共有が減った可能性がある．条件間での比較は，クロス集計（3×2）のカイ二乗検定を行った．独立変数は各条件における教育資料の効果検証テスト回答の正/誤であり，従属変数はその回答者数だった．

④ ナッジの効果測定

気晴らしナッジの提示前と後において, Disinformation-viral 刺激を共有する回答者の割合を算出した．これは，「6.2.本実験」でみられた気晴らしナッジの効果と同様の結果が得られたかを確認するためだった．気晴らしナッジの提示後に Disinformation-viral 刺激を共有すると回答した参加者の割合は 66.3%，すなわち 33.7%の参加者が共有をキャンセルすると回答していた．再度類似した数値が結果において得られた場合，実験における気晴らしナッジの効果の再現性を示すことができる．

⑤ 仮説検証のための分析

仮説 1 の解は，Disinformation-viral 刺激を共有すると回答した人数を条件間（ナッジ/教育/統制）で比較することで得られる．Disinformation-viral 刺激を共有すると回答した人数は，ナッジ条件が教育条件よりも有意に少なく，教育条件は統制条件よりも有意に少ないと予測した．分析にあたり，ナッジ条件は提示後の共有意思を分析対象とし，リポスト又は引用の回答を「共有する」，キャンセルの回答を「共有しない」に分類した．仮説検証は，クロス集計（3×2）のカイ二乗検定を行った．独立変数は各条件における Disinformation-viral 刺激の共有意思（共有する/共有しない）であり，従属変数はその回答者数だった．

仮説 2 の解は，Disinformation-viral 刺激の共有意思の回答において，感情の強さと信憑性判断のどちらの影響が大きいかを条件ごとに確認することで得られる．ナッジ条件は情動調節により熟慮が促されるため，信憑性判断が感情よりも共有意思への影響が大きい可能性がある．反対に，教育条件と統制条件は，Disinformation の怒りを生み出す要因が熟慮を妨げるため, 感情が信憑性判断よりも共有意思への影響が大きい可能性がある．分析するにあたり，ナッジ条件は提示後の共有意思を分析対象とし，リポスト又は引用の回答を「共有する」，キャンセルの回答を「共有しない」に分類した．仮説検証のため，ロジスティック回帰分析を実施した．ロジスティック回帰分析は，従属変数が 2 値（共有する/共有しない）である場合に，複数の独立変数から従属変数である 2 値の結果に及ぼす影響を予測することができる．独立変数は感情の強さ（0.感情なし, 1.弱い〜10.強い）と信憑性判断（信じる/信じない）であり，従属変数は Disinformation-viral 刺激の共有意思（共有する/共有しない）だった．

7.3. 結果

(1) 操作チェック

テキスト投稿刺激の不当性による違いにおいて Disinformation 刺激は True information 刺激よりも嫌悪と怒りが多く，バイラル性による違いにおいて viral 刺激は control 刺激よりも感情が強かった. テキスト投稿刺激に対して認識した感情の種類は, Disinformation 刺激では悲しみ（14.6%），嫌悪（8.5%）又は怒り（7.5%）の感情が多く認識され，True information 刺激では期待（8.1%）又は悲しみ（7.9%）の感情が多く認識された（図 7-5）. テキスト投稿刺激ごとに感情の強さの中央値と四分位偏差を算出したところ，viral 刺激（Md＝0-5）は control 刺激（Md＝0）よりも強い感情の回答が有意に多かった（p＜.001）（表 7-2）. 感情の種類によって認識された感情の強さは異なり，嫌悪と怒り（Md＝7, QD＝5-8）が最も強く，驚き（Md＝6, QD＝5-6）が最も弱かった.

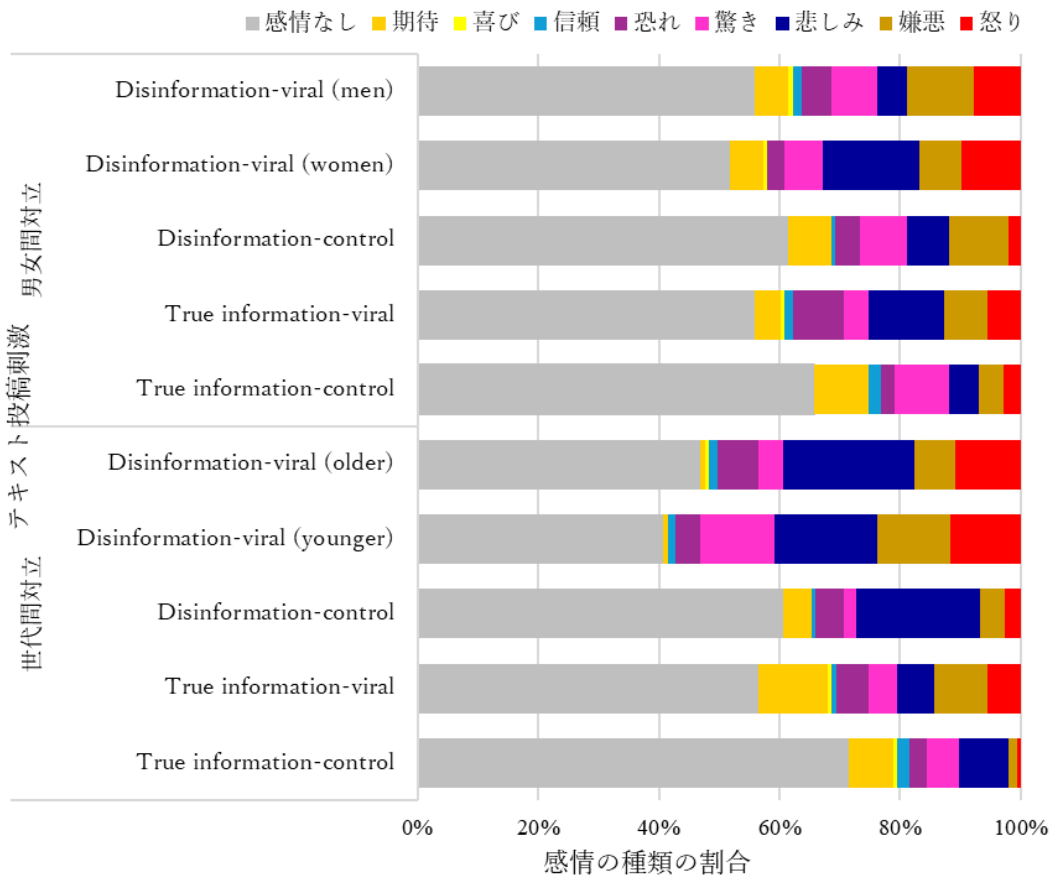


図 7-5 各テキスト投稿刺激で認識された感情の種類の割合

表 7-2 各テキスト投稿刺激の感情の強さ（中央値と四分位偏差）

テーマ	テキスト投稿刺激	*Md* (*QD*)
男女間対立	Disinformation-viral (men)	0.0 (0.00-6.00)
	Disinformation-viral (women)	0.0 (0.00-6.00)
	Disinformation-control	0.0 (0.00-5.00)
	True information-viral	0.0 (0.00-6.00)
	True information-control	0.0 (0.00-5.00)
世代間対立	Disinformation-viral (older)	4.0 (0.00-6.00)
	Disinformation-viral (younger)	5.0 (0.00-7.00)
	Disinformation-control	0.0 (0.00-5.00)
	True information-viral	0.0 (0.00-6.00)
	True information-control	0.0 (0.00-4.00)

教育効果検証テストの正答率は，統制条件（16.2%），教育条件（16.0%），ナッジ条件（14.4%）の順に高かった．実験後の感情（1=ポジティブ〜5=ネガティブ）について中央値と四分位偏差を算出したところ，中央値は 3（*QD*＝3-3）であり極端な負への偏りはなかった．

(2) 全般的な回答傾向

各テキスト投稿刺激において共有すると回答した参加者の割合を算出した結果，Disinformation-viral 刺激を共有する割合は概ね平均以上であり，条件によっては平均以下のものがあった（表 7-3）．ナッジ条件は提示前と後の共有する割合をそれぞれ算出し，ナッジの提示後についてはリポスト又は引用の回答を「共有する」として集計した．男女間対立テーマで共有すると回答した人数の割合の平均は 10.0%（*SD* = 4.7）であり，Disinformation-viral (women) 刺激は 10.0〜23.4%と平均以上だった．世代間対立テーマで共有すると回答した人数の割合の平均は 9.1%（*SD*＝4.1）であり，Disinformation-viral 刺激の両方とも概ね平均以上だった．

表 7-3 テキスト投稿刺激ごとの共有する割合

テーマ	テキスト投稿刺激	ナッジ(*n*=97)		教育	統制
		提示前	提示後	(*n*=94)	(*n*=99)
男女間対立	Disinformation-viral (men)	10.6	8.5	4.3	4.0
(*n*=143)	Disinformation-viral (women)	23.4	12.8	13.0	10.0
	Disinformation-control	6.4	-	6.5	10.0
	True information-viral	12.8	-	2.2	10.0
	True information-control	12.8	-	13.0	10.0
世代間対立	Disinformation-viral (older)	14.0	10.0	4.2	10.2
(*n*=147)	Disinformation-viral (younger)	12.0	8.0	14.6	18.4
	Disinformation-control	6.0	-	6.3	8.2
	True information-viral	10.0	-	10.4	0.0
	True information-control	8.0	-	6.3	8.2

(3) Disinformation の共有を減らす効果

Disinformation-viral 刺激を共有する割合が多かったのは，統制条件，ナッジ条件，教育条件の順だった．ナッジ条件は提示後の共有意思を分析対象とし，リポスト又は引用の回答を「共有する」，キャンセルの回答を「共有しない」に分類して集計した．各条件で Disinformation-viral 刺激を共有すると回答した参加者の割合は，ナッジ条件で 9.8%，教育条件で 9.0%，統制条件で 10.6%だった．また，ナッジ条件では，ナッジ提示後にキャンセルすると回答した参加者の割合は 34.5%だった．

仮説 1 を分析した結果，条件間で Disinformation-viral 刺激を共有すると回答した人数に有意な差はなかった．その理由を明らかにするために，教育の効果検証テストの正答率に注目した．Disinformation-viral 刺激を共有すると回答した人数の割合とテストの正答率について z スコアで標準化し，3 条件間で比較をした（図 7-6）．その結果，ナッジ条件は効果検証テストの正答率が低いにも関わらず Disinformation-viral 刺激を共有すると回答した人数の割合は中程度だった．教育条件と統制条件は，いずれも効果検証テストの正答率が高かった．しかし，Disinformation-viral 刺激を共有すると回答した人数の割合は，教育条件では最も少なかったのに対し，統制条件は最も多かった．

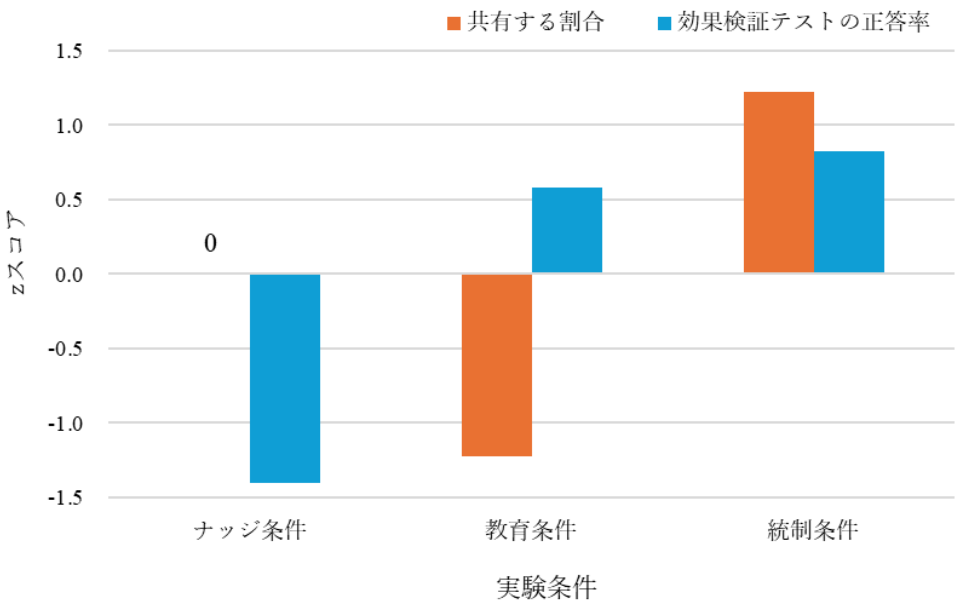


図 7-6 効果検証テストの正答率と Disinformation-viral 刺激を共有する割合

(4) 共有意思に及ぼす影響の違い

仮説 2 を検証した結果，ナッジ条件と統制条件では感情よりも信憑性判断が Disinformation-viral 刺激の共有意思に及ぼす影響が大きかった（表 7-4）．ナッジ条件と統制条件では，共有意思へ有意に影響を及ぼす独立変数は感情の強さと信憑性判断の両方だった（いずれも $p<.05$）．一方で，教育条件では共有意思に有意に影響を及ぼす独立変数は感情の強さのみであり（$p<.01$），信憑性判断は共有意思に影響を及ぼす変数ではなかった．さらに，ナッジ条件と統制条件において，感情の強さと信憑性判断のどちらがより共有意思に及ぼす影響が大きいか独立変数間で比較するために，標準化偏回帰係数（β）を算出した．共有意思に有意に影響を及ぼす独立変数のうち，標準化偏回帰係数の数値が大きい独立変数の方が，共有意思に及ぼす影響がより大きいことを意味している．ナッジ条件では，信憑性判断（1.354）が感情の強さ（0.283）よりも標準化偏回帰係数が大きく，共有意思に及ぼす影響の大きさは信憑性判断の方が大きいことが分かった．統制条件も，信憑性判断（1.723）が感情の強さ（0.426）よりも標準化偏回帰係数が大きく，共有意思に及ぼす影響の大きさは信憑性判断の方が大きいことが分かった．なお，教育条件のオッズ比は，独立変数「信じない」と従属変数「共有する」と回答した参加者が 0 人であり，特定カテゴリーの値がデータセットに存在しないことから無限大（Infinity）という結果となった可能性がある．

表 7-4 共有意思に影響を与える独立変数

条件	独立変数	β	OR (95%CI)	Hosmer-Lemeshow
ナッジ条件	感情の強さ	0.283**	1.3 (1.1-1.6)	0.067
	信憑性判断	1.354*	3.9 (1.2-12.8)	
教育条件	感情の強さ	0.377**	1.5 (1.1-1.9)	0.951
	信憑性判断	18.428	100682600 (0-Inf)	
統制条件	感情の強さ	0.426***	1.5 (1.2-1.9)	0.381
	信憑性判断	1.723*	5.6 (1.2-26.4)	

* $p < .05$, ** $p < .01$, *** $p < .001$

ナッジ条件では情動調節を用いた熟慮により感情の強さが弱まったかどうか確認したところ，ナッジ提示後は他の条件よりも感情が弱まっていたことが分かった（図 7-7）．各条件において, Disinformation-viral 刺激を共有すると回答した参加者が認識していた感情の強さ（0-10）について中央値と四分位偏差を算出した．ナッジ条件は，提示後の共有意思を分析対象とし，リポスト又は引用の回答を「共有する」，キャンセルの回答を「共有しない」に分類して集計した．その結果，Disinformation-viral 刺激を共有すると回答した参加者の感情の強さは，ナッジ条件（$Md = 5, QD = 5\text{-}7$）が教育条件と統制条件（いずれも $Md = 7, QD = 7\text{-}8$）よりも弱かった．

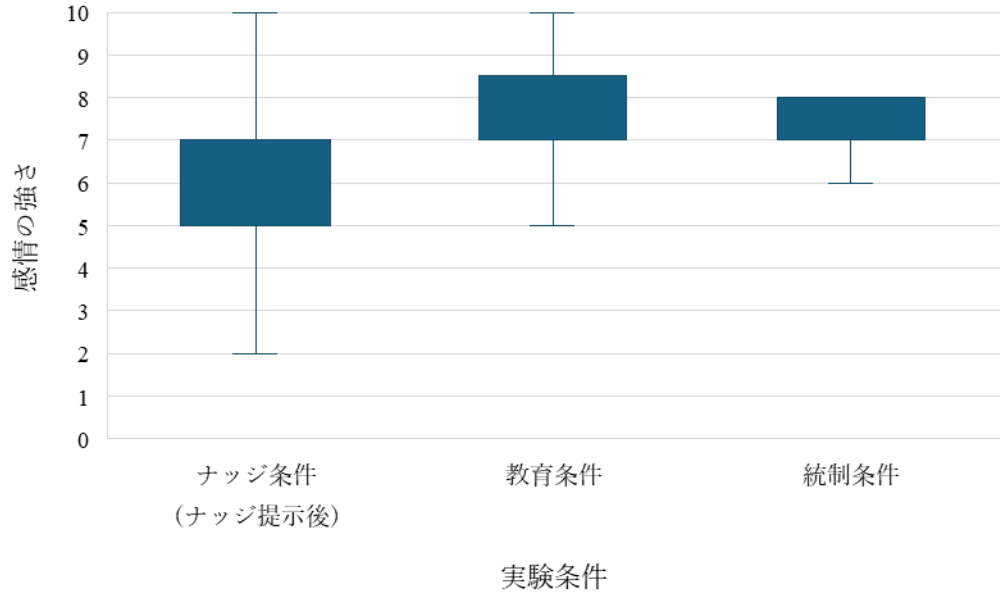


図 7-7 Disinformation-viral 刺激を共有する回答者の感情の強さ

7.4. 考察

Disinformation の怒りを生み出す要因が熟慮を妨げるという予測に基づき，情動調節ナッジと教育が共有を減らす効果を比較評価した結果，情動調節ナッジは教育を補完する介入策として有効であることが示された．仮説 1 は支持されなかったが，情動調節ナッジは教育の効果と同じく，Disinformation-viral 刺激の共有を減らす一定程度の効果があることが分かった．仮説 2 は支持され，情動調節ナッジでは信憑性判断が感情よりも共有意思に及ぼす影響がより大きく，教育は感情の強さのみが共有意思に有意に影響を及ぼしていた．情動調節ナッジは Disinformation の強い怒りが共有を促進した時に介入し，情動調節効果によって感情の影響が小さくなることで，信憑性判断の影響が大きくなった可能性がある（図 7-8）．これに対し，教育は Disinformation の強い怒りが共有を促進した時には教育の効果が発揮されず，感情の影響が大きいままに共有していた可能性がある（図 7-9）．これは，事前に教育をしても Disinformation の怒りを生み出す要因によって共有が促進されてしまうため，情動調節ナッジを併用して熟慮に促すことが信憑性判断の影響を強める可能性を示唆している．

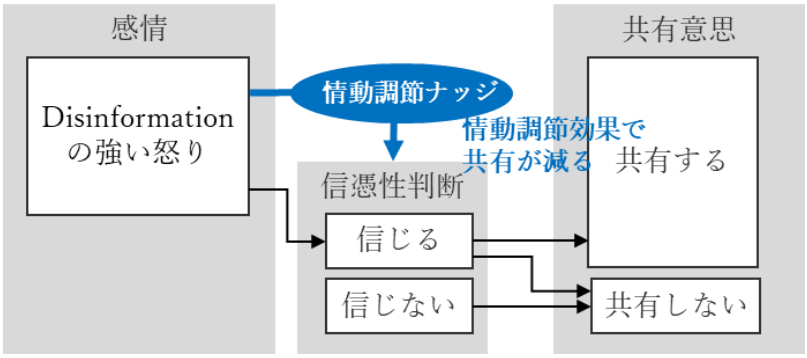


図 7-8 情動調節ナッジの認知プロセスの考察

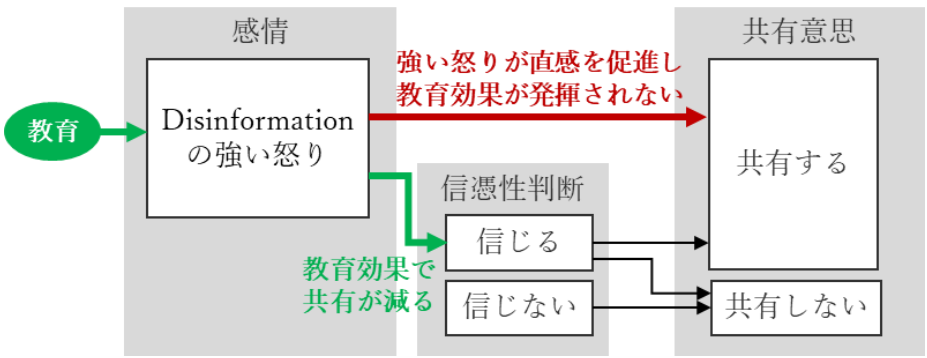


図 7-9 教育の認知プロセスの考察

(1) Disinformation の共有を減らす効果

仮説 1 の検証結果において条件間で Disinformation を共有する人数に有意な差がなかったことから，教育でみられた減少効果と同じく，情動調節を用いて熟慮を促す気晴らしナッジにおいても一定の減少効果が確認された. 気晴らしナッジが Disinformation-viral 刺激の共有を減らす割合は 34.5%であり，これは「6.2.本実験」で得られた 33.7%という結果と近似していた．参加者は，Disinformation-viral 刺激に対して，悲しみ，怒り，又は嫌悪の感情を最も多く認識していた（図 7-5）．これは，気晴らしナッジで表示された Disinformation-viral 刺激の感情情報（表 7-1）と類似している．一方で，同時に表示された情動調節メッセージを通じて，気晴らしナッジは参加者に微笑んでいるフォロワーを想像するよう促した. この感情の種類の不一致[292]が情動調節の必要性[320]を引き起こし，参加者は熟慮が促されたことで Disinformation-viral 刺激の共有が減った可能性がある．

しかし，この結果には留意すべき事項が 2 点ある．第一に，教育条件に割り当てられた参加者の効果検証テストの正答率が低かったことである. 総務省の公開資料によると，同じ教材を用いた教育受講者の平均正答率は 65%だった[247]．これに対し，本研究の教育条件に割り当てられた参加者の正答率は 16%と低かった．このような差異が発生した理由として，使用した教材は本来セミナー形式で学ぶことを想定して作成されたものであるため，単に読むだけでは理解が深まらなかった可能性が考えられる．第二に，統制条件に割り当てられた参加者の効果検証テストの正答率が，条件間で最も高い傾向がみられたことである（図 7-6）．参加者は，事前知識レベルが結果に与える影響を最小限にするために，スクリーニングによって情報リテラシーや情報モラルについて学んだことがない人のみに絞り込んだ．その上で，参加者は各条件へランダムに割り当てられたが，統制条件の参加者は効果検証テストの正答率が 16.2%であり，教育条件で教材を読んだ参加者の正答率と近似していた．統制条件と教育条件では同程度の知識・スキルを有していることが考えられるが，それにも関わらず条件間で Disinformation-viral 刺激を共有する割合は異なっていた．これは，教育条件では実験前に教材が提示されたことで参加者の意識が高まり，Disinformation-viral 刺激の共有が抑制された可能性を示唆している．

(2) 共有意思に及ぼす影響の違い

ナッジ条件では，共有意思に及ぼす影響は信憑性判断が感情の強さよりもより大きく（表 7-4），これは気晴らしナッジにより参加者の感情の強さが弱まった（図 7-7）こと

が関係している可能性がある．先行研究において強いネガティブな感情は熟慮プロセスを妨げる可能性があることが示唆されいている [290]．その考えに従えば，参加者は Disinformation-viral 刺激に対して強い怒りを認識したが，気晴らしナッジが感情のバランスを取り戻すのに役立ち，それが信憑性判断に基づく理性的な共有につながったと考えられる．信憑性判断に基づく判断は，直感的認知プロセスを促す感情 [286,287] ではないため，熟慮的認知プロセスによるものである可能性がある．

一方，教育条件では，Disinformation-viral 刺激の共有意思に有意に影響を与えたのは感情の強さだけだった．参加者は，信憑性判断に関わらず強い感情を認識した Disinformation-viral 刺激を共有していることが分かった．予想した通り，Disinformation に対して認識した強い怒りが熟慮的認知プロセスを妨げ，それによって教育の効果が発揮されなかった可能性がある．これは，事前に教育によって知識や能力を習得しても，ソーシャルメディアを利用している時に Disinformation に遭遇して強い感情を認識すると，直感的認知プロセスによって Disinformation を感情的に共有してしまう可能性があることを示唆している．

統制条件は，ナッジ条件と同じく，信憑性判断が感情の強さよりも共有意思に及ぼす影響がより大きかった．しかし，Disinformation-viral 刺激を共有する参加者の感情の強さは，教育条件と同じく強い傾向がみられた（図 7-7）．ナッジ条件と異なり，統制条件の参加者は Disinformation-viral 刺激に関連する感情を強く認識し，信じて共有している可能性がある．つまり，統制条件での Disinformation-viral 刺激の共有は，感情による直感的認知プロセスによって，鵜呑みに信じた共有であった可能性がある．

7.5. 小括

本章では，Disinformation の怒りを生み出す要因に対して，情動調節ナッジが既存の教育を補完する介入策として有効であることを実験により明らかにした．6 章で既存ナッジよりも Disinformation の共有を減らす効果が高かった気晴らしナッジと教育の比較評価を行った．教育は，総務省が公開している教育教材の抜粋が使用された．

第一に，ナッジは Disinformation の強い怒りが共有を促進した時に介入することによって，信憑性判断に基づく共有を促した可能性が示唆された．

第二に，教育はその効果がみられたものの，Disinformation の強い怒りを認識した場合には，信憑性判断に関わらず共有が促進されてしまう可能性が示唆された．

以上の結果から，事前に教育によって知識や能力を習得しても，Disinformation に遭遇して強い感情を認識すると感情的に共有してしまう可能性がある．このため，教育の効果を十分に発揮させるための補完策としてナッジは有用であると考えられる．

8. 総括

本章では，これまでに得られた調査及び実験の結果を総括し，本研究の貢献，限界，及び提言と今後の課題について述べる．本研究では，Disinformation の共有を減らす効果的なユーザ向け対策を提案することを目指した．Disinformation の怒りが共有を促進している可能性があることから，現在実施又は検討されている Disinformation 対策を調査し，その十分性について関連研究の調査をもとに考察した．その結果，現対策において感情に言及するものがあったが，その効果は怒りの共有メカニズムによって十分に発揮されない可能性が考えられた．

そこで，本研究では Disinformation の怒りを生み出す要因が共有に及ぼす影響を明らかにし，Disinformation の共有を減らすために怒りに着目した有効策を提案することを目的とした．この目的を達成するために，Disinformation 対策及び関連研究の調査から明らかになった検討すべき重要な 3 つの課題を解決する実験を行った．

第一の実験では，Disinformation の怒りによる共有メカニズムが実証されていないという課題に対し，Disinformation の怒りを生み出す要因が信憑性に関わらず共有を促進していることを明らかにした．仮説検証モデルを作成して検証した結果，Disinformation は怒りから「信じて共有するルート」と強い怒りから「信憑性に関わらず共有するルート」の 2 つの経路で共有が促進されていることが分かった．

第二の実験では，ユーザが強い怒りから Disinformation を共有しようとした時に介入する怒りに着目した有効策がないという課題に対し，情動調節ナッジを作成し，既存ナッジよりも Disinformation の共有を減らす効果が高いことを明らかにした．特に，フォロワーへの思いやりを想像する気晴らしの情動調節ナッジが最も効果が高かった．

第三の実験では，Disinformation の怒りに着目した介入策が現対策と比較しても有用かどうか明らかではないという課題に対し，情動調節ナッジは怒りに関する対策として最も使用されている教育よりも Disinformation の怒りの共有メカニズムに対して有効であることを明らかにした．Disinformation の共有を減らす効果を比較評価した結果，情動調節ナッジは感情の影響を減らすことによって信憑性判断に基づく共有を促進したのに対し，教育には感情に基づく共有を抑制する効果はみられなかった．Disinformation の強い怒りが熟慮を妨げることによって教育の効果が十分に発揮されなかった可能性があるため，情動調節ナッジは教育を補完する効果が見込まれる．

8.1. 本研究の貢献

本研究の第一の新規性は，先行研究では検証されていなかった「感情が共有に及ぼす影響」を明らかにしたことである（図 8-1 の赤色）．Disinformation の怒りを生み出す要因は，信憑性に関わらず Disinformation の共有を促進した．この傾向は，教育を事前に受けた参加者においてもみられた．これは，ユーザがソーシャルメディアの利用中に強い怒りを表現する Disinformation に遭遇すると，ユーザの直感的認知プロセスが促進され，教育で習得した知識やスキルを活用して Disinformation の真偽について吟味することなく，感情のままに共有してしまう可能性があるということを意味している．

本研究の第二の新規性は，Disinformation 対策の新たなアプローチとして，真偽ではなく感情，特に Disinformation が悪用する強い怒りに注意を向けさせるユーザ介入策として情動調節ナッジを提案したことである（図 8-1 の青色）．Disinformation の強い怒りが抑制する教育の効果を十分に活用するためには，強い怒りにより促進された直感的認知プロセスを熟慮的認知プロセスに切り替えることが有用であると考える．その認知プロセスの切り替えを試みる新しいアプローチとして，本研究では情動調節ナッジを提案し，その有効性を示した．情動調節ナッジには，投稿コンテンツに含まれる感情の種類とその割合を示す感情情報と，熟慮を促すヒントとなる情動調節メッセージが組み込まれていた．このような仕組みによって Disinformation の強い怒りにユーザの注意を向けさせ熟慮を促すことが，直感的認知プロセスから熟慮的認知プロセスへの切り替えを支援した可能性がある．

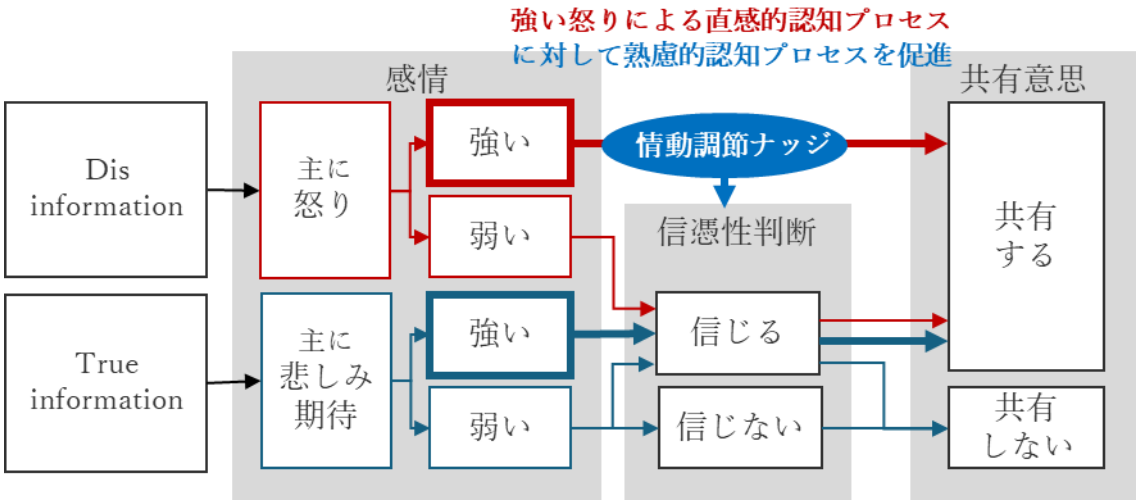


図 8-1 Disinformation の怒りを生み出す要因に対する情動調節ナッジの効果

本研究の貢献は，従来の真偽に着目した対策では拡散を抑制することが難しいDisinformationに対し，新たな手法として強い怒りに着目することの有用性を示したことである．Disinformationには，違法性がないことから現行法による法的対処が困難なもの，完全なる虚偽ではないことから技術的な検出が難しいものが存在する．このため，ソーシャルメディアに残存したDisinformationが拡散するかどうかは，それらを共有するユーザ個々の能力・判断に依存している．しかし，Disinformationには事実又は事実と虚偽が混在したものもあることから，ユーザが真偽の観点からDisinformationであるか否かを見分けるのは困難である．このような課題に対し，本研究ではDisinformationの特徴の1つである強い怒りに着目した情動調節ナッジという新たな対処方法を提案した．

情動調節ナッジのうち，特に最も効果が高いことが示された気晴らしナッジは，共有すると回答していた参加者の平均34.1%（実験②33.7%，実験③34.5%）が共有を中止した．2021年の日本での調査において，偽情報を気付かずに拡散した5,991人のフォロワーを合計すると，拡散数は7,805,450人だった[347]．このうち気晴らしナッジによって34.1%の人が共有を中止した場合，約266万人への影響が減ることとなる．実際のソーシャルメディアでは共有の連鎖により拡散が生じているため，連鎖の過程における共有の減少に累積的な効果をもたらすことが期待される．情動調節ナッジはDisinformationの共有を減らすことを目的としているが，それだけでなく教育で習得した知識やスキルを有効に機能させることに貢献する可能性がある．また，ソーシャルメディアを利用している全てのユーザが，教育機会に恵まれているわけではない可能性がある．情動調節ナッジはソーシャルメディア利用時に直接介入するため，教育機会を得にくいユーザへのアウトリーチに役立つ．

Disinformationの感情に着目することの有用な点は，第一にプラットフォーム事業者が懸念する「真実の裁定者になる」[46,47]必要がないことである．プラットフォーム事業者は違法又は虚偽コンテンツの検出及び削除等の対応を要請されているが，いずれにも該当しないDisinformationは表現の自由と公衆衛生の保護との間にトレードオフの関係がある[114]．現在プラットフォーム事業者はコンテキスト（情報源，他ユーザによる注釈等）を提供しているが，Disinformationの怒りが直感的認知プロセスを促進した場合，それらのコンテキストを読むことなく共有行動が促進される可能性がある．怒りに着目した対策は，投稿コンテンツに対する真偽性判断を仰ぐものではなく，コンテンツ内に含まれ

る感情情報を視覚化するものである．投稿コンテンツの表現の自由を尊重しながら，ユーザの熟慮に基づく判断を支援することは，民主主義国家における「真実の裁定者」は国民であるという考えに寄与する．

第二に，判定基準が真偽ではないことから事実の検証を要さないことである．ファクトチェック団体は実施体制や資金面においてリソースを確保することが難しいという現状があり，検証できる記事数の量的課題がある[186]．また，Disinformation は必ずしも虚偽ではないことから，ファクトチェック対象とならず拡散して大きな影響を及ぼす可能性もある．このため，Disinformation が悪用する強い怒りを起点にユーザ自身が自分の行動を再考できるシステムにより，ユーザ 1 人 1 人の理性的な行動から健全な情報空間の構築を目指すことが重要であると考える．

情動調節ナッジをより活用していくためには，教育と併用することでその効果を相乗効果的に高めていくことが重要である．実験結果より情動調節ナッジが教育の補完策として有効であることが示唆されたが，教育によって情動調節ナッジの効果を高めることもできる．プラットフォーム事業者は新しい機能を導入した際に，当該機能の目的や意図についてブログや動画で説明している．しかし，全ての利用ユーザが機能に関する説明を事前に読む可能性は高くないと考えられる．フリクションによるナッジは，一時停止させることでユーザの行動を阻害するため，リアクタンスを生じさせるリスクがある[216]．このため，教育において Disinformation が悪用する技法（極端な意見，声を増幅させる，感情に働きかける等[243]）を知ることで，情動調節ナッジがプラットフォーム事業者において導入・活用される背景への理解が促進されると考えられる．

情動調節ナッジの応用方法としては，ユーザの感情的な共有行動に介入するという仕組みを，ユーザによる感情的なコメント投稿行動への介入に活用できる可能性がある．Kiskola ら[291,292]及び Syrjämäki ら[293]はオンラインニュースサイトのニュース記事に対するユーザの無礼なコメント投稿に対して，情動調節介入をするユーザインタフェースを研究していた．ニュースサイトのコメント投稿欄やソーシャルメディアの投稿欄に書き込んだ文章の感情評価がカラーバー等で視覚化されると，人格攻撃（アドホミネム攻撃）や誹謗中傷等の感情的なコメントの投稿を再考することにつながる可能性がある．

8.2. 本研究の限界

本研究では仮説に基づく検証により結果を示したが，これらの結果には大きく 3 つの

限界があった．それは，実験室実験による限界，情動調節ナッジの実装に関する限界，及びナッジという介入策そのものの効果の限界である．

(1) 実験室実験による限界

本研究で得られた知見は，管理された実験環境で得られた結果に基づいたものであり，実際のプラットフォーム環境上での効果については検証されていない．第一に，本研究で示された Disinformation に対する効果は，実験者が作成したテキスト投稿刺激に対する反応への効果に限られている．Disinformation 刺激は実験者が創作したものであり，それらが Disinformation であるという刺激の妥当性を検証することは難しいと考えられる．また，怒りに着目した有効策を提案することを目的に，本研究では Disinformation が社会的な意見の対立・分断を狙うために怒りを悪用する状況を想定した．つまり，情動調節ナッジによる効果を検証したのは，強い怒りを表現する Disinformation のみである．しかし，実環境での Disinformation は多様であり，ワクチン接種に関連するもの，気候変動等の環境に関するもの，災害に関連するもの，株価操作を狙ったもの等もある．このような Disinformation に対しても同様の効果が得られるか，又は逆に悪影響を及ぼす可能性があるかどうかは更なる検証が必要である．例えば，ワクチン接種や災害に関連する Disinformation では恐ろしさが拡散要因である可能性がある[356]ため，怒りではなく恐怖のスコアを使用する等，Disinformation の種類に応じた情動調節ナッジの拡張が考えられる．

第二に，実験デザインが結果に影響を与えた可能性は完全には否定できない．Web 調査会社を用いたアンケート形式での実験は，参加者の実験環境を統制することができないため，テキスト投稿刺激と参加者が回答した感情の因果関係を証明することはできない．また，テキスト投稿刺激，ナッジ，及び教育以外の要因が回答へ影響を与えた可能性を排除することはできない．

第三に，情動調節ナッジが Disinformation の共有を減らす効果は，実験室環境で測定された効果に留まる．「6.2.本実験」において X の機能を模した既存ナッジが共有を減らした割合（22.1%）は実測値（23%）[140] と近似していたが，先行研究において実環境ではナッジの効果が実験室実験で得られた効果よりも減少したことが報告されている [345]．このため，実環境における多様な Disinformation にも有効であるかどうか，情動調節ナッジの効果が現実のプラットフォーム上で再現できるかは更なる検証が必要である．

(2) 情動調節ナッジの実装に関する限界

実環境に情動調節ナッジを実装するにあたっては，リアルタイムでの感情分析と怒りのスコア判定に基づくラベル付けが前提条件となる（図 8-2）．本研究では，実験の参加者から得られた感情の種類の回答とその割合が，情動調節ナッジの感情情報として組み込まれた．これを実環境に実装するにあたっては，プラットフォームに投稿されたコンテンツのリアルタイムでの感情分析が必要となる．X は投稿データの情報抽出において感情分析の有用性を示しており[357]，また投稿者が過去 7 日間の自身の投稿を感情分析してスコア（ポジティブ/ネガティブ/中立）を表示する方法を提供している[358]．海外では X の投稿コンテンツをリアルタイムで感情分析するプロジェクトが進められていたり[359]，感情分析ツールが GitHub 上でオープンソースとして公開されていたりする[360,361]．日本においても，Yahoo! Japan が「Yahoo!リアルタイム検索」というサービスにて，検索した文字列を含む投稿コンテンツに対して自動推定された感情分析結果（ポジティブ/ネガティブ/中立）を表示している[362]．ただし，これら多くの感情分析サービスによる分類は 3 カテゴリー（ポジティブ/ネガティブ/中立）に留まる．このため，感情分析でネガティブに分類された投稿コンテンツのみ, さらに感情の種類ごとに分類する感情分析ツール（例えば，日本語であれば「pymlask」[363]等）を併用することで，怒りが多い投稿コンテンツを分類することが可能になる．この怒りが強いかどうかは投稿コンテンツに「怒り」が含まれていると閲覧者が認識する可能性を確率スコアとして算出し（例えば,「Perspective API」[364]），事前に指定した怒りスコアの閾値を超えた場合に投稿コンテンツへラベルを付与する．このラベル付き投稿コンテンツを閲覧者が共有しようとリポストボタンを押した時に，感情の種類ごとの感情分析結果を感情情報とした情動調節ナッジが，閲覧者の画面にポップアップ表示される仕組みが考えられる．

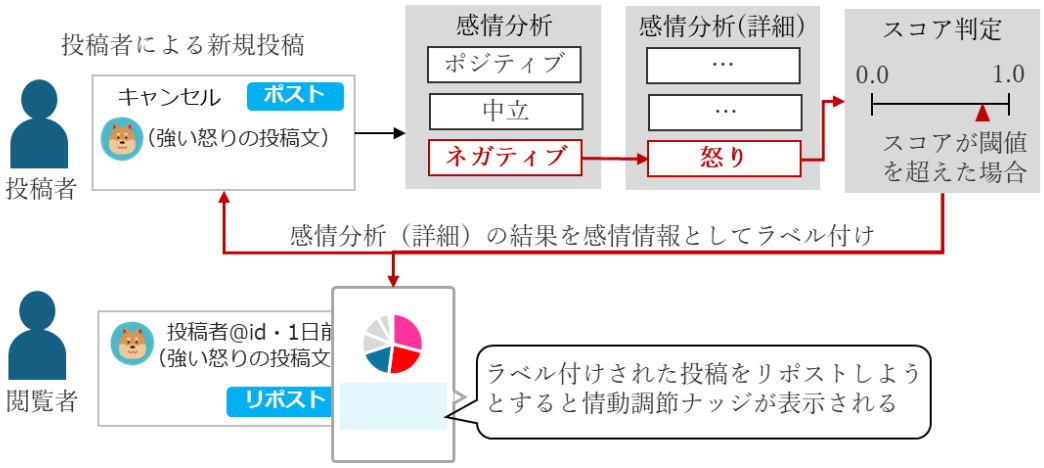


図 8-2 情動調節ナッジの実装例

しかし，この仕組みによる情動調節ナッジの実装には 3 つの課題がある．第一に，情動調節ナッジが表示される判定基準は投稿コンテンツの怒りスコアが高いか否かであるため，強い怒りを認識させる True information にも表示されることになるが，その影響については本研究では検証されていない．本研究は実験室実験ではあるものの，可能な限り実際のソーシャルメディア環境に近い状況を設定し，その状況での情動調節ナッジの効果を測定することが重要であると考えた．このため，2006 年～2017 年に Twitter で拡散したデータを分析した先行研究 [276] の結果に類似する状況での検証を試みた．先行研究の分析結果では，True information は悲しみ，期待，喜び，及び信頼の感情を引き起こすことが多かった．本研究でも同様の傾向が結果として得られており，全ての実験において True information-viral 刺激に対して認識された感情は期待又は悲しみが多かった．このような実環境に近い状況で効果を測定することが重要であると考えるが，頻度が少ないながらも強い怒りを認識させる True information に対しても情動調節ナッジが表示されるという仕組みは，抑制すべきではない True information の共有まで減らしてしまう可能性がある．情動調節ナッジと教育を比較評価した結果では，教育では怒りを認識させることが多い Disinformation だけでなく，期待又は悲しみを認識しやすい True information の共有も減らすケースがみられた（表 7-3）．同様に，情動調節ナッジが怒りを認識させる True information の共有を減らす可能性があるため，本研究は検証が不足していたと考える．怒りを認識させる True information に対して情動調節ナッジが表示されることは，アラブの春のような民主化運動の事例を抑制し得る可能性がある．True information に対して情動調節ナッジが表示されることの影響については，Disinformation 対策としての効果

との兼ね合いも含め実環境での検証が必要である．

第二に，情動調節ナッジを実装するプラットフォーム事業者において，導入及び運用に掛かるコスト増を受容する必要がある．情動調節ナッジは情報の真偽を判定するものではないため「真実の裁定者にならない」というスタンスに沿いつつ Disinformation の蔓延を抑制できるが, 実装及び運用にあたっては機器と人員リソースが追加で必要となる．例えば，X は 1 日あたり 1 億件のオリジナル投稿がある[365]ことから，1 秒あたり約 1157 件の新規投稿があることが想定される．このうち，日本に関しては X の利用率が X 全体の 11.21%[366]であることから，1 秒あたり約 130 件の投稿があると仮定する．おそらく，処理性能が高いマシンを用いることで実装できる可能性があるが，あくまでも仮定に基づいた概算であるため導入及び運用コストに関しては検証が必要である．また，感情分析についてもいくつかの課題があり，皮肉や風刺，短いテキスト，Bot，言語の多様性とスラングについて正確に検出・解釈することが難しいとされている [359]．特に，日本語はハイコンテクスト文化であるため，分類精度を高める必要があると考えられる．これらのコスト増に対しては，例えば当該サービスを有料で提供するといったことが考えられる．投稿コンテンツの感情分析は，これまでも企業のマーケティング戦略の中で活用されてきた[367]．昨今はそれだけでなく，ソーシャルメディア上での炎上等が企業にとってのレピュテーションリスクであると認識されつつある．危機管理の観点から，企業に関連するトピックにおいて怒りスコアが高い投稿が増えているといったようなレポート結果を，X プレミアム[368]等の有料サービスで提供することで有料会員の増加を図るといったことが考えられる．

第三に，ソーシャルメディアユーザ及びプラットフォーム事業者が，Disinformation の共有を減らす仕組みとして情動調節を用いることを受け入れない可能性がある．本研究では, Disinformation の怒りによる有害な行動（例えば, 選挙のボイコット）を問題視し，情動調節によりユーザが自身の幸福につながる理性的な行動（例えば，意思表示のための投票行動）につなげる仕組みを提案した．しかし，Disinformation を意図的に共有したいユーザ，又は理性よりも感情や衝動を重視して行動するユーザも存在することが考えられる．情動調節ナッジで表示される選択肢は「リポスト」「引用」ボタンであり，共有したいという表現の自由を尊重している．これに加えて，情動調節ナッジを使用するかどうかについて，ユーザが設定で選択できるようにすることが望ましい．また，プラットフォーム事業者は強い怒りを認識させる True information にも情動調節ナッジが表示さ

れること，及びユーザの共有行動が減ることを許容できない可能性がある．情動調節ナッジは，情動調節効果によりソーシャルメディアでの怒りの連鎖・増幅を抑制することにも貢献する可能性があるが，プラットフォーム事業者が強い怒りも含む様々な感情が蔓延することでソーシャルメディア上での議論が活発化することを望んでいる場合には情動調節ナッジは適合しない．

(3) ナッジによる介入策の効果の限界

「3.2.4.ナッジに関する留意事項」にて前述した通り，ナッジには長期的効果について限界があることが指摘されている[327]．これは，ナッジには教育的な効果がない[342]ことや慣れによる効果の減衰[344]によるものである．情動調節ナッジの設計ではこのような慣れを考慮し，感情情報は Disinformation-viral 刺激ごとに異なる色と割合からなる円グラフが表示される仕組みとした．しかし，実環境では強い怒りが認識される投稿コンテンツのみに情動調節ナッジが表示される仕組みであるため，感情情報は怒りの割合が多い円グラフばかり表示されることが想定される．情動調節ナッジが長期的な効果を持つかどうかを判断するには，実環境で表示される頻度やバリエーション，ユーザの利用頻度によるナッジの受け止め方の違い等も含めて更なる検証が必要である．

また, 情動調節ナッジが表示されることによる逆効果についての懸念がある. 例えば, 情動調節ナッジが表示されることにより, あえて怒りを増幅するために Disinformation を共有することを選択するユーザがいる可能性がある．本研究では，ユーザによる共有という表現の自由を尊重する観点から，当該逆効果については防ぐことができない．

8.3. 提言と今後の課題

Disinformation がソーシャルメディア上で拡散するのは，ソーシャル Bot 以上に，ソーシャルメディアを利用しているユーザが Disinformation を共有しているからである．Disinformation 対策には，法による規制，モニタリングと検出，ファクトチェックによるユーザ支援，及び教育によるユーザの能力向上等があるが，本研究ではいずれの対策でも対処が難しい Disinformation を対象に怒りに着目した新たな有効策を提案した．

ユーザによる Disinformation の共有を減らすためには，Disinformation を共有するユーザの認知的脆弱性や認知プロセスを考慮したコグニティブセキュリティの観点から対策を検討することが有用であることを本研究は示唆した．これまでの Disinformation 対策

の多くは，真偽を判断基準とし，それを見分けるための能力開発やナッジ介入が行われてきた．しかし，Disinformation は必ずしも虚偽ではなく，事実と虚偽の混在や事実（不都合な真実）が含まれる．このように，従来の情報の真偽を起点とした判断では難しいケースが存在するため，本研究では Disinformation の特徴の 1 つである怒りに着目し，感情を起点としてユーザによる Disinformation の共有を減らす対策を提案した．Disinformation の怒りを生み出す要因が共有を促進していることから，Disinformation が悪用する感情的側面へユーザの意識を向けることが，ユーザ自身の気付きや振り返りにつながる可能性がある．このように，Disinformation が悪用する認知的脆弱性と，その脆弱性によって生じる認知プロセスへ対処することが，ユーザによる Disinformation の共有を減らす介入策として有用であると考える．

日本では，ソーシャルメディアをはじめとする「デジタル空間における情報流通の健全性を確保すること」が重要な課題となっている．本研究はこの課題を解決する 1 つの方法を示すものだが，プラットフォーム事業者が本提案手法を実環境に導入することを前提としている．技術的な観点において，日本語を対象とした精度の高い感情分析ツールが必要であり，プラットフォーム事業者又は学術・研究機関はそのための研究開発が必要である．このような取組みに対して，①情報流通の健全性確保を目的とした研究開発を促進するための助成金，②研究開発したツール・製品等の評価に使用するための Disinformation データセット等が含まれる研究開発用の共有データベース基盤，③健全性確保に寄与するプラットフォーム事業者の取組みに対するインセンティブといったものが，本研究をはじめとする Disinformation 対策の推進を後押しする．

今後の発展課題として，Disinformation の怒りを生み出す要因が，誤った信念の構築に及ぼす影響とその対策について検討する必要があると考える．本研究の目的はユーザが Disinformation を共有するという行動を減らすことであり，Disinformation によって誤った信念が構築されるのを防ぐことまでは対象としていなかった．情動調節ナッジが Disinformation の共有を減らすということは，共有の連鎖過程において Disinformation に遭遇して信念に影響を受ける可能性のあったユーザも減ることを意味する．しかし，最初に Disinformation に遭遇したユーザは，情動調節ナッジが表示されて強い怒りが弱まることで共有を再考する可能性があるが，怒りは弱まっても信念に影響を与えている可能性がある．このため，今後は Disinformation を信じるユーザを減らす効果があるユーザ介入策についても研究していく所存である．

謝辞

本研究の遂行にあたり，多くの方々にご指導ご鞭撻を賜りました．
指導教官である情報セキュリティ大学院大学 稲葉緑教授には，博士前期課程より長きに渡り多大なご指導を賜りました．ここに深謝の意を表します．
同大学院 後藤厚宏教授には，主査，博士演習の教官，並びにメンター教官として，多くの場面で適切なご助言を賜りました．深く感謝申し上げます．
同大学院 大久保隆夫教授，村上康二郎教授には，副査として本論文の作成にあたり多くのご助言を賜りました．心より感謝申し上げます．
香川大学創造工学部 橋本正樹准教授には，博士演習の教官として適切なご指導賜りました．心より感謝申し上げます．
予備実験にご協力頂いた株式会社ラック サイバー・グリッド・ジャパンの皆様，並びに本実験にご参加頂いた皆様に，心よりお礼申し上げます．
研究活動の遂行にあたり多くの示唆を頂いた国立研究開発法人情報通信研究機構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の皆様にお礼申し上げます．
稲葉研究室の皆様には，本研究の遂行にあたり多くのご助言，ご協力頂きました．ここに誠意の意を表します．
最後に，大学院生活を支えてくれたパートナーと子どもたちに感謝いたします．
あらためて皆様のご支援に心より感謝いたします．

研究業績

１．査読付き学術論文

- Haruka Nakajima Suzuki and Midori Inaba, “Digital Nudges Using Emotion Regulation to Reduce Online Disinformation Sharing,” in IPSJ Journal of Information Processing (JIP) - Special issue of “Applications and the internet” in conjunction with the main topics of COMPSAC 2024, vol. 33, 2025. ※2025 年 7 月 3 日採択，10 月出版予定

２．査読付き国際会議（プロシーディングとして採録）

- Haruka Nakajima Suzuki and Midori Inaba, “Nudges to Reduce the Spread of Online Disinformation: A Comparison with the Educational Effect,” in Proceedings of the 20th International Workshop on Security (IWSEC 2025), 2025. ※2025 年 6 月 2 日採択，11 月 25～27 日発表予定
- Haruka Nakajima Suzuki and Midori Inaba, “Psychological Study on Judgment and Sharing of Online Disinformation,” in Proceedings of the 2023 IEEE 47th Annual Computer Software and Applications Conference (COMPSAC), pp. 1558–1563, 2023.

３．その他の研究実績

- 鈴木悠, “情報操作型サイバー攻撃における認知的側面,” 日本セキュリティ・マネジメント学会解説論文, vol. 38, no. 1, pp. 13–20, 2024.
- 鈴木悠・稲葉緑, “Disinformation の共有を減らすための情動調節戦略を用いたナッジ,” 情報処理学会研究報告, vol. 2024-SPT-55, no. 4, pp. 1–7, 2024.
- 鈴木悠・稲葉緑, “Disinformation の人による拡散を低減するための有効策の考察,” 情報処理学会研究報告, vol. 2024-SPT-55, no. 3, pp. 1–8, 2024.
- 中嶋悠, “感情の喚起が Disinformation の二次的な社会的共有に及ぼす影響,” 情報セキュリティ大学院大学情報セキュリティ研究科修士論文, 2022.
- 中嶋悠・稲葉緑, “Disinformation による情動伝染が二次的な社会的共有に及ぼす影響,” 情報処理学会研究報告, vol. 2021-SPT-45, no. 13, pp. 1–6, 2021.

付録

付録 1. 調査対象とした教育プログラム

No.	名称	提供元	主な教育内容
1	Internet literacy handbook (2017)	欧州委員会	デジタルスキル
2	Resilience Series Graphic Novels	米国土安全保障省 CISA	Disinformation 関連
3	Spot and fight disinformation	欧州委員会	Disinformation 関連
4	Media and information literate citizens: think critically, click wisely!	UNESCO	メディアリテラシー
5	インターネットとの向き合い方〜ニセ・誤情報にだまされないために〜	総務省	Disinformation 関連
6	Misinformation and disinformation	OECD	メディアリテラシー
7	NATO's approach to countering disinformation	NATO	Disinformation 関連
8	SELMA	AISBL (Belgium)	ヘイトスピーチ
9	Social Media Literacy for Change	AISBL (Belgium)	メディアリテラシー
10	Media Literacy for Living Together	ルゾフォナ大学 (Portugal)	メディアリテラシー
11	GET YOUR FACTS STRAIGHT!	All Digital (Belgium)	Disinformation 関連
12	START2THINK	Centre for International Relations (Poland)他	Disinformation 関連
13	SMaRT-EU	COFAC (Portugal) 他	Disinformation 関連
14	YouVerify!	AFP (France) 他	Disinformation 関連

15	e-Media: Media Literacy and Digital Citizenship for All	All Digital (Belgium)	メディアリテラシー
16	Barefoot Computing	BCS (UK) 他	デジタルスキル
17	Media in Action	Pontydysgu (UK) 他	メディアリテラシー
18	Common Sense Education Digital Citizenship	Common Sense とハーバード大学大学院 (USA)	デジタルスキル
19	Mind Over Media	Media Education Lab (USA)	対プロパガンダ
20	Check, Please! Starter Cours	ワシントン州立大学 (USA)	Disinformation 関連
21	Checkology	News Literacy Project (USA)	ニュースリテラシー
22	CTRL-F	CIVIX (Canada)	デジタルスキル

付録 2. 文章のテキスト投稿刺激（10 個）

(1) 男女間対立テーマ

刺激条件	内容
Disinformation-viral (men)	職場のイケメン新人君が突然会社を辞めたんだけど、退職後に「女上司から逆セクハラによる精神的苦痛を受けた」って女上司と会社を訴えて裁判沙汰になったんだよ！でも、裁判では女から男へのセクハラ行為は考えにくい、親身な指導の一貫に過ぎないと認められなかったって。そもそも判例が少ないからって…女尊男卑反対！
Disinformation-viral (women)	結婚・妊娠したら女は寿退社、子育てが落ち着いたら職場復帰すればいいって言うじゃない？私は毎年正規雇用を希望しているのにず～っとパートのまま。正規雇用と同じ責任ある業務をしているのに低賃金で期限付き。男だらけの経営者が女性の労働力を安く買い叩いてる。日本女性の平均所得は男性よりも 43.7%も低いんだって！
Disinformation-control	日本の男女差別をなくすために活動してる NPO 法人が、発足直後から SNS で大炎上したとニュースになってる。団体代表が、過去に SNS で性差別的な発言をしていたことが次々と発覚した様子。慌てて団体代表が謝罪して、その取り巻きも擁護してたけど、偽善的な活動なのがミエミエ。
True information-viral	新型コロナウイルスの感染拡大で、女性の非正規労働者や母子世帯など弱い立場にある人が最も影響を受ける「女性不況」であることが指摘された。在宅勤務の増加に伴い DV 被害も深刻化したほか、主婦や高校生などの自殺者が増加したことが挙げられている。日本のジェンダー不平等の実態があらためて浮き彫りになった。 （抜粋元：日本経済新聞[369]）
True information-control	現在の会社は男女平等であると感じるか聞いたところ、「感じる」は 51.4%、「感じない」は 48.7%となった。男女平等ではないと感じている人は男性 45.3%、女性 53.8%と、男女ともに格差を感じていることがうかがえる。今後企業は、すべての人が平等に評価されるシステムを作ることが課題となる。 （抜粋元：PR TIMES[370]）

(2) 世代間対立テーマ

刺激条件	内容
Disinformation-viral (older)	若者のワクチン接種率が低いと聞いてるのに、外ではマスクなしで騒いでる若者がビックリするほど多い！感染者数がようやっと減ってきたものの、またいつ感染者が増えるか分からないからと多くの人ができるだけ活動を自粛して経済も停滞しているというのに。身勝手な若者が感染の脅威をまき散らしていると思うと非常に腹立たしい。
Disinformation-viral (younger)	保育園に子どもを預けて、4 月から仕事に復帰する予定でいたのに本当に信じられない。新しくできる保育所に受かったから安心しきってた。まさか、近隣のジジババからの猛反対で開園中止になるなんて。「騒音が」って、自分の子育てのときは？静かだったの？ほんと 4 月からどうすればいいの…。
Disinformation-control	定年が 65 歳まで延長されたが、60 歳以降は給与も上がらないし現状維持ができてればいいと考えている。年下の上司も仕事を頼みにくそうだし、後進育成といっても今の仕事はパソコンが中心でデジタルに疎い自分が教えられるようなこともない。高齢者の雇用拡大が若者採用を抑制すると批判もあるが、年金も 65 歳からだし生活のためだ。
True information-viral	「若者の一票を高齢者の一票よりも重くすべき。」このコロナ禍が変化のチャンスと市長が語っている。若者は自分たちが弱者であることにすら気づいていない。カギを握るのは年金制度を支えている現役世代の行動である。若者は自分たちが相当まずい状況に追い込まれていることに一刻も早く気づくべきだ。 （抜粋元：Yahoo！JAPAN ニュース[371]）
True information-control	コロナ禍で、ネットに不慣れな人たちが取り残される「デジタル格差」が広がっている。スマホを持っていない高齢者が多くいることから、スマホを無償貸与してネット利用の促進をはかる自治体もでてきた。一方で、予算や人員の制約から、デジタル格差はやむを得ないとする立場の自治体もあるようだ。 （抜粋元：朝日新聞デジタル[372]）

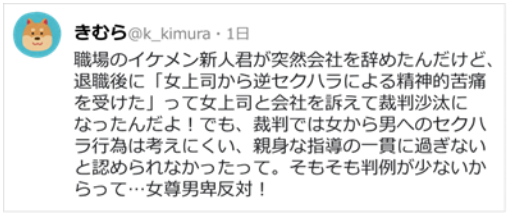
付録 3. 情動調節メッセージ（9 個）

情動調節戦略	ID	メッセージ文	参考元
感情情報のみ	A1	（なし）	-
気晴らし	D1	誰かの話ではなく、あなたに最近あった楽しい出来事を投稿しませんか？	300
	D2	投稿者に対する温かく思いやりのある言葉を想像してください。その言葉は共有で伝えられますか？	301
	D3	いつもあなたの投稿を見て微笑んでいるフォロワーを想像してください。	
再評価	R1	何に心を揺さぶられましたか？あなたの感情や考えを、より前向きな変化と捉えて言葉にしてみませんか。	306 307 308
	R2	この感情は投稿者の感情です。あなたの感情ではありません。	309
視点取得	P1	投稿者の感情は、あなたが共有することで解消されるでしょうか？	303 312
	P2	この投稿は、閲覧者の感情を刺激することを意図している可能性があります。	313 314
共感的対応	E1	あなたの感情はあなたのものです。投稿者の感情ではなく、あなたの感情を大事にしてください。	316

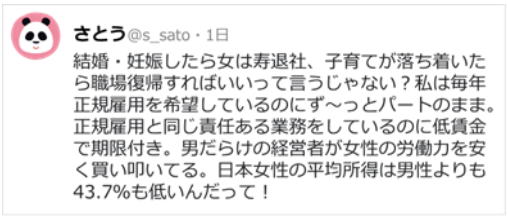
付録 4. 画像のテキスト投稿刺激（10 個）

(1) 男女間対立テーマ

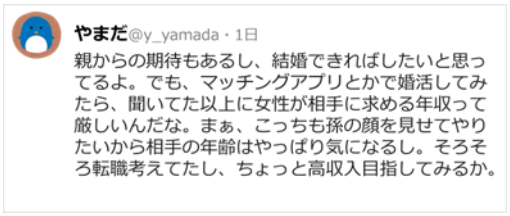
Disinformation-viral (men)



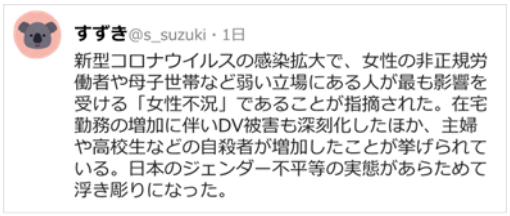
Disinformation-viral (women)



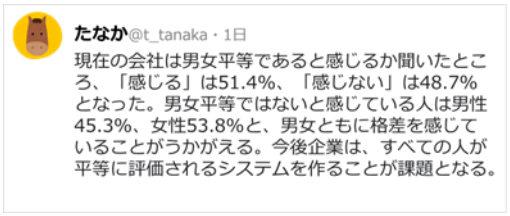
Disinformation-control



True information-viral

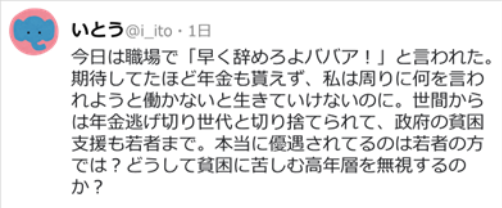


True information-control

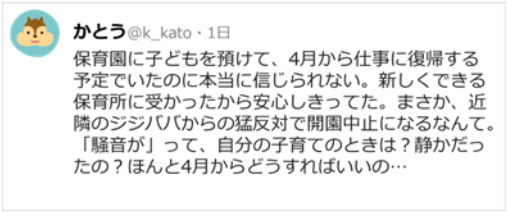


(2) 世代間対立テーマ

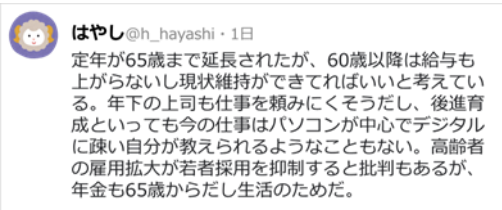
Disinformation-viral (older)



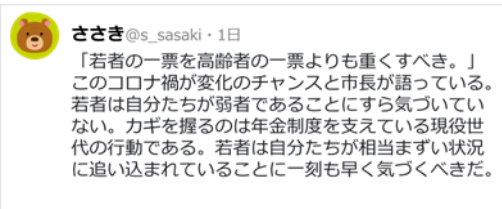
Disinformation-viral (younger)



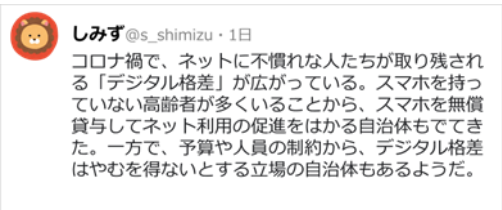
Disinformation-control



True information-viral

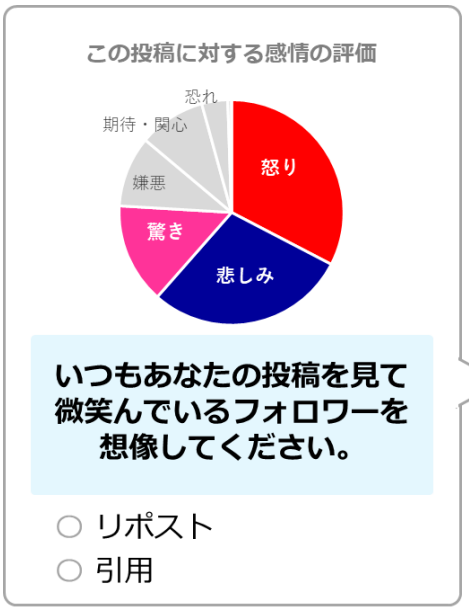


True information-control

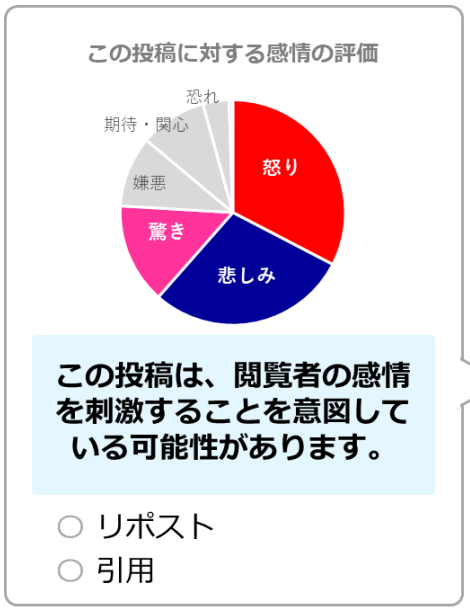


付録 5. ナッジデザイン（3 種）

(1) 気晴らしナッジ



(2) 視点取得ナッジ



(3) 既存ナッジ

引用文献

[1] 総務省, “平成 29 年度版 情報通信白書,” 2017. https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h29/html/nc111130.html

[2] 総務省, “平成 27 年度版 情報通信白書,” 2015. https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/pdf/n4200000.pdf

[3] 総務省, “令和元年度版 情報通信白書,” 2019. https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r01/html/nd114120.html

[4] W. Claire and D. Hossein, “Information disorder: Toward an interdisciplinary framework for research and policymaking,” Council of Europe, 2017. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

[5] M. E. Bonfanti, “An Intelligence-based approach to countering social media influence operations,” Romanian Intelligence Studies Review, no. 19–20, pp. 47–68, 2019. https://www.ceeol.com/search/article-detail?id=874123

[6] B. M. Pierce, “Protecting people from disinformation requires a cognitive security proving ground,” C4ISRNET, 2021. https://www.c4isrnet.com/opinion/2021/02/10/protecting-people-from-disinformation-requires-a-cognitive-security-proving-ground/

[7] Early use of the word, “disinformation,” Medicine Lodge Cresset, 17 February 1887, pp.3. https://www.newspapers.com/paper/medicine-lodge-cresset/3084/

[8] B. Albert and J. Hendrik, “Krym Nash: An Analysis of Modern Russian Deception Warfare,” Utrecht University Repository, 2020. https://dspace.library.uu.nl/handle/1874/400504

[9] A. Mahairas and M. Dvilyanski, “Disinformation–Дезинформация (Dezinformatsiya),” The Cyber Defense Review, vol. 3, no. 3, pp. 21–28, 2018. https://www.jstor.org/stable/26554993

[10] European External Action Service (EEAS) , “EUvsDisinfo,” 2015. https://euvsdisinfo.eu/

[11] Directorate-General for Communications Networks, Content and Technology (European Commission), “A multi-dimensional approach to disinformation,” European Union, 2018. https://op.europa.eu/en/publication-detail/-/publication/6ef4df8b-4cea-11e8-be1d-01aa75ed71a1/language-en

[12] European Commission, “Tackling online disinformation: a European Approach,” 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236

[13] House of Commons, “Disinformation and ‘fake news’: Final Report,” 2019. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

[14] T. Rid, “Active Measures: The Secret History of Disinformation and Political Warfare,” Farrar Straus & Giroux, 2020.

[15] T. Nagasako, “Global Disinformation Campaigns and Legal Challenges,” International Cyber Security Law Review, vol. 1, pp. 125–136, 2020. https://doi.org/10.1365/s43439-020-00010-7

[16] Disinformation 対策フォーラム, “Disinformation 対策フォーラム 中間とりまとめ,” 2021. https://www.saferinternet.or.jp/anti-disinformation/disinformation_interim_report/

[17] 総務省, “平成 24 年度版 情報通信白書,” 2012. https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h24/html/nc1212c0.html

[18] P. N. Howard and M. M. Hussain, “Democracy’s Fourth Wave?: Digital Media and the Arab Spring,” Oxford University Press, 2013. https://doi.org/10.1093/acprof:oso/9780199936953.003.0001

[19] A. Goldenberg and J. J. Gross, “Digital Emotion Contagion,” Trends in Cognitive Sciences, vol. 24, pp. 316–328, 2020. https://doi.org/10.1016/j.tics.2020.01.009

[20] Валерий Герасимов, “Ценность науки в предвидении,” Военно-промышленный курьер, no. 8 (476), 2013. https://vpk.name/news/85159_cennost_nauki_v_predvidenii.html

[21] P. N. Howard, B. Ganesh, D. Liotsiou, J. Kelly, and C. François, “The IRA, Social Media and Political Polarization in the United States, 2012-2018,” University of Oxford, 2018. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2018/12/The-IRA-Social-Media-and-

Political-Polarization.pdf
[22] B. Kaiser, “Targeted: My Inside Story of Cambridge Analytica and How Trump, Brexit and Facebook Broke Democracy,” HarperCollins, 2019.
[23] Information Commissioner’s Office, “Investigation into the use of data analytics in political campaigns: Investigation update,” 2018. https://ico.org.uk/media2/migrated/2259371/investigation-into-data-analytics-for-political-purposes-update.pdf
[24] C. Wylie, “Mindf*ck: Inside Cambridge Analytica’s Plot to Break the World,” Profile Books Ltd, 2019.
[25] A. Stamos, “An Update On Information Operations On Facebook,” Meta blog, 2017. https://about.fb.com/news/2017/09/information-operations-update/
[26] R. DiResta, K. Shaffer, B. Ruppel, D. Sullivan, R. Matney, R. Fox, J. Albright, and B. Johnson, “The Tactics & Tropes of the Internet Research Agency,” New Knowledge Organization, 2018. https://int.nyt.com/data/documenthelper/533-read-report-internet-research-agency/7871ea6d5b7bedafbf19/optimized/full.pdf
[27] 長迫智子, 小谷賢, 大澤淳, “SNS 時代の戦略兵器 陰謀論,” ウェッジ, 2025.
[28] A. Erlich and C. Garner, “Is pro-Kremlin disinformation effective? Evidence from Ukraine,” The International Journal of Press/Politics, vol. 28, pp. 5–28, 2023. https://doi.org/10.1177/19401612211045221
[29] R. Fan, K. Xu, and J. Zhao, “Higher contagion and weaker ties mean anger spreads faster than joy in social media,” 2016. (Preprint) https://doi.org/10.48550/arXiv.1608.03656
[30] R. L. Nabi, “Exploring the Framing Effects of Emotion: Do Discrete Emotions Differentially Influence Accessibility, Information Seeking, and Policy Preference?” Communication Research, vol. 30, pp. 224–247, 2003. https://doi.org/10.1177/0093650202250881
[31] S. Steinert and M. J. Dennis, “Emotions and Digital Well-Being: on Social Media’s Emotional Affordances,” Philosophy & Technology, vol. 35, article no. 36, 2022. https://doi.org/10.1007/s13347-022-00530-6
[32] Poynter, “A guide to anti-misinformation actions around the world,” 2018. https://www.poynter.org/ifcn/anti-misinformation-actions/
[33] C. von der Weth, J. Vachery, and M. Kankanhalli, “Nudging Users to Slow Down the Spread of Fake News in Social Media,” in Proceedings of the 2020 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–6, 2020. https://doi.org/10.1109/ICMEW46912.2020.9106003
[34] 総務省, “インターネット上の偽・誤情報の流通・拡散に適用され得る既存の法制度（例）,” デジタル空間における情報流通の健全性確保の在り方に関する検討会ワーキンググループ（第 20 回）配付資料 WG20-2-4, 2024. https://www.soumu.go.jp/main_content/000945918.pdf
[35] 総務省, “デジタル空間における情報流通の健全性確保の在り方に関する検討会 とりまとめ,” 2024. https://www.soumu.go.jp/main_content/000966997.pdf
[36] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. LiuShu, “FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media,” Big data, vol. 8, no. 3, pp. 171–188, 2020. https://doi.org/10.1089/big.2020.0062
[37] S. Ippa, T. Okubo, and M. Hashimoto, “An Analysis of Social Bot Activity on X in Modern Japan,” IEEE Access, vol. 12, pp. 125800–125808, 2024. https://doi.org/10.1109/ACCESS.2024.3454536
[38] S. Ippa, T. Okubo, and M. Hashimoto, “The Relationship Between Emotional and Other Factors in Information Diffusion,” IEEE Access, vol. 13, pp. 21249–21264, 2025. https://doi.org/10.1109/ACCESS.2025.3535547
[39] An OSoMe project, “Botometer X,” 2014. https://botometer.osome.iu.edu/
[40] U.S. Department of Justice, “Justice Department Leads Efforts Among Federal, International, and Private Sector Partners to Disrupt Covert Russian Government-Operated Social Media Bot

Farm," 2024. https://www.justice.gov/archives/opa/pr/justice-department-leads-efforts-among-federal-international-and-private-sector-partners
[41] TinEye, "Reverse Image Search," 2008. https://tineye.com/
[42] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS 2018), pp. 1–7, 2018. https://doi.org/10.1109/WIFS.2018.8630761
[43] Meta, "Here's how we're using AI to help detect misinformation," 2020. https://ai.meta.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/
[44] Stanford University, "Human Writer or AI? Scholars Build a Detection Tool," 2023. https://hai.stanford.edu/news/human-writer-or-ai-scholars-build-detection-tool
[45] Open AI, "New AI classifier for indicating AI-written text," 2023. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/
[46] The Guardian, "Zuckerberg says Facebook won't be 'arbiters of truth' after Trump threat," 2020. https://www.theguardian.com/technology/2020/may/28/zuckerberg-facebook-police-online-speech-trump
[47] jack [@jack], X/Twitter, 2020/05/28. https://x.com/jack/status/1265837139360485376 # Former CEO of Twitter.
[48] K. Coleman, "Introducing Birdwatch, a community-based approach to misinformation," X blog, 2021. https://blog.x.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation
[49] J. Allen, C. Martel, and D. G. Rand, "Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program," in Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 245, pp.1–19, 2022. https://doi.org/10.1145/3491102.3502040
[50] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand, "Shifting attention to accuracy can reduce misinformation online," Nature, vol. 592, pp. 590–595, 2021. https://doi.org/10.1038/s41586-021-03344-2
[51] F. Jahanbakhsh, A. X. Zhang, A. J. Berinsky, G. Pennycook, D. G. Rand, and D. R. Karger, "Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media," in Proceedings of the ACM on Human-Computer Interaction (HCI), vol. 5, pp. 1–42, 2021. https://doi.org/10.1145/3449092
[52] U. K. H. Ecker, S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, and M. A. Amazeen, "The psychological drivers of misinformation belief and its resistance to correction," Nature Reviews Psychology, vol. 1, pp. 13–29, 2022. https://doi.org/10.1038/s44159-021-00006-y
[53] A. Ibrahim, J. Ye, and C. Hoffner, "Diffusion of news of the Shuttle Columbia Disaster: The role of emotional responses and motives for interpersonal communication," Communication Research Reports, vol. 25, pp. 91–101, 2008. https://doi.org/10.1080/08824090802021970
[54] J. Berger and K. L. Milkman, "What makes online content viral?" Journal of Marketing Research, vol. 49, pp. 192–205, 2012. https://doi.org/10.1509/jmr.10.0353
[55] S. Lewandowsky, U. K. H. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and its correction: Continued influence and successful debiasing," Psychological Science in the Public Interest, vol. 13, pp. 106–131, 2012. https://doi.org/10.1177/1529100612451018
[56] A. Kozyreva, S. Lewandowsky, and R. Hertwig, "Citizens versus the internet: Confronting digital challenges with cognitive tools," Psychological Science in the Public Interest, vol. 21, pp. 103–156, 2020. https://doi.org/10.1177/1529100620946707
[57] J. Roozenbeek, E. Culloty, and J. Suiter, "Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions," European Psychologist, vol. 28, no. 3, pp. 189–205, 2023. https://doi.org/10.1027/1016-9040/a000492
[58] A. Kozyreva, P. Lorenz-Spreen, S. M. Herzog, U. K. H. Ecker, S. Lewandowsky, R. Hertwig, A. Ali, J. Bak-Coleman, S. Barzilai, M. Basol, A. J. Berinsky, C. Betsch, J. Cook, L. K. Fazio, M. Geers, A. M. Guess, H. Huang, H. Larreguy, R. Maertens, F. Panizza, G. Pennycook, D. G. Rand, S. Rathje, J. Reifler, P. Schmid, M. Smith, B. Swire-Thompson, P. Szewach, S. van der Linden, and S. Wineburg, "Toolbox of individual-level interventions against online misinformation," Nature Human Behaviour, vol. 8, pp. 1044–1052, 2024. https://doi.org/10.1038/s41562-024-

01881-0

[59] T. Nagasako, “Global Disinformation Campaigns and Strategic Challenges –Case Study and Consideration of National Strategies as the Countermeasures–,” 情報セキュリティ大学院大学博士論文, 2024. https://lab.iisec.ac.jp/degrees/d/theses/iisec_d50_thesis.pdf

[60] 公益財団法人笹川平和財団安全保障研究グループ, “外国からのディスインフォメーションに備えを！～サイバー空間の情報操作の脅威～,” 2022. https://www.spf.org/cyber/publications/20220207_cyber.html

[61] Economist Intelligence, “Democracy Index 2024,” 2025. https://www.eiu.com/n/campaigns/democracy-index-2024/

[62] European Court of Auditors, “Special Report 09/2021: Disinformation affecting the EU: tackled but not tamed,” 2021. https://www.eca.europa.eu/en/publications?did=58682

[63] European Parliament, “Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance),” 2022. http://data.europa.eu/eli/reg/2022/2065/oj

[64] European Commission, “Questions and answers on the Digital Services Act*,” 2024. https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348

[65] European Commission, “The Code of Conduct on Disinformation,” 2025. https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation

[66] Directorate-General for Communication, “European Democracy Action Plan: making EU democracies stronger,” 2020. https://commission.europa.eu/publications/documents-european-democracy-action-plan_en

[67] European Parliament, “Foreign interference in all democratic processes in the European Union,” 2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022IP0064&qid=1749648835731

[68] Department for Science, Innovation & Technology, “Counter-Disinformation Unit – open source information collection and analysis: privacy notice,” GOV.UK, 2023. https://www.gov.uk/government/publications/counter-disinformation-unit-open-source-information-collection-and-analysis-privacy-notice/counter-disinformation-unit-open-source-information-collection-and-analysis-privacy-notice

[69] House of Commons, “Disinformation and ‘fake news’: Interim Report: Government Response to the Committee’s Fifth Report of Session 2017–19,” 2018. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1630/1630.pdf

[70] House of Commons, “Disinformation and ‘fake news’: Interim Report,” 2018. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/363.pdf

[71] Department for Digital, Culture, Media & Sport and Home Office, “Online Harms White Paper,” GOV.UK, 2019. https://www.gov.uk/government/consultations/online-harms-white-paper

[72] Department for Digital, Culture, Media & Sport and The Rt Hon Karen Bradley MP, “Internet Safety Strategy green paper,” GOV.UK, 2017. https://www.gov.uk/government/consultations/internet-safety-strategy-green-paper

[73] Department for Science, Innovation and Technology, Department for Digital, Culture, Media & Sport and The Rt Hon Matt Hancock, “Digital Charter,” GOV.UK, 2018. https://www.gov.uk/government/publications/digital-charter

[74] J. Pamment, “RESIST 2 Counter Disinformation Toolkit,” Government Communication Service, 2021. https://gcs.civilservice.gov.uk/publications/resist-2-counter-disinformation-toolkit/

[75] Department for Digital, Culture, Media & Sport and Home Office, “Online Harms White Paper: Full government response to the consultation,” GOV.UK, 2020. https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response

[76] UK Parliament, “Online Safety Act 2023,” legislation.gov.uk, 2023. https://www.legislation.gov.uk/ukpga/2023/50

[77] Ofcom, “Enforcing the Online Safety Act: Ofcom opens nine new investigations,” 2025.

https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/enforcing-the-online-safety-act-ofcom-opens-9-new-investigations

[78] J. Woodhouse, L. Conway, and S. Lipscombe, “Online Safety Bill: progress of the Bill,” UK Parliament, 2023. https://commonslibrary.parliament.uk/research-briefings/cbp-9579/

[79] UK Parliament, “National Security Act 2023,” legislation.gov.uk, 2023. https://www.legislation.gov.uk/ukpga/2023/32/contents

[80] Home Office, “Foreign interference: National Security Bill factsheet,” GOV.UK, 2025. https://www.gov.uk/government/publications/national-security-bill-factsheets/foreign-interference-national-security-bill-factsheet

[81] The White House, “RESTORING FREEDOM OF SPEECH AND ENDING FEDERAL CENSORSHIP,” 2025. https://www.whitehouse.gov/presidential-actions/2025/01/restoring-freedom-of-speech-and-ending-federal-censorship/

[82] 楽原響子, “トランプ政権の「偽情報対策」廃止政策の全貌,” 日米同盟研究会コメンタリー, no. 63, 2025.

https://www.npi.or.jp/research/data/npi_commentary_kuwahara_20250529.pdf

[83] D. B. Johnson, “State Department’s disinformation office to close after funding nixed in NDAA,” CyberScoop, 2025. https://cyberscoop.com/state-departments-disinformation-office-to-close-after-funding-nixed-in-ndaa/

[84] Cybersecurity and Infrastructure Security Agency, “Building Resilience to Foreign Interference, Misinformation Activities,” 2019. https://www.cisa.gov/news-events/alerts/2019/07/22/building-resilience-foreign-interference-misinformation-activities

[85] D. DiMolfetta, “DHS Secretary Kristi Noem has committed to rescoping the Cybersecurity and Infrastructure Security Agency so that it pivots away from disinformation matters,” NextGov/FCS, 2025. https://www.nextgov.com/people/2025/02/cisa-staff-focused-disinformation-and-influence-operations-put-leave/402958/

[86] G. Sands, “DHS shuts down disinformation board months after its efforts were paused,” CNN, 2022. https://edition.cnn.com/2022/08/24/politics/dhs-disinformation-board-shut-down/index.html

[87] C. Tan, “Regulating disinformation on Twitter and Facebook,” Griffith Law Review, vol. 31, pp. 513–536, 2022. https://doi.org/10.1080/10383441.2022.2138140

[88] 水谷瑛嗣郎, “SNS と法の交錯点―表現の自由、民主政治の視点から―,” ソーシャルメディアの動向と課題：科学技術に関する調査プロジェクト報告書, 国立国会図書館調査及び立法考査局, 2020.

https://dl.ndl.go.jp/view/download/digidepo_11472869_po_20190504.pdf?contentNo=1

[89] 水谷瑛嗣郎, “偽情報にどう向き合うか―憲法・メディア法の視点から,” 法学研究, 97 巻, 12 号, pp. 132–133, 2024. https://aslp.law.keio.ac.jp/pdf/AN00224504-20241228-0132.pdf

[90] Bloomberg News, “ツイッター、トランプ氏投稿に「誤解を招く恐れある」と警告ラベル,” 2020. https://www.bloomberg.co.jp/news/articles/2020-11-04/QJ9D9RT1UM10

[91] The White House, “Executive Order on Preventing Online Censorship,” 2020. https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-preventing-online-censorship/

[92] 株式会社三菱総合研究所, “インターネット上の違法・有害情報を巡る米国の動向,” 総務省プラットフォームサービスに関する研究会（第 24 回）配布資料 4, 2021.

https://www.soumu.go.jp/main_content/000739937.pdf

[93] Ministry of Home Affairs, “Written Answer by Minister for Law, Mr K Shanmugam, to Parliamentary Question on the Responsibility of Social Media Platforms and Internet Service Providers in Addressing Fake News,” 2017. https://www.mlaw.gov.sg/news/parliamentary-speeches/written-answer-by-minister-for-law--mr-k-shanmugam-to-parliament/

[94] Government of Singapore, “Deliberate Online Falsehoods: Challenges and Implications,” National Archives Singapore, 2018.

https://www.nas.gov.sg/archivesonline/government_records/record-details/6797717d-f25b-11e7-bafc-001a4a5ba61b
[95] Parliament of Singapore, “Report of the Select Committee on Deliberate Online Falsehoods – Causes, Consequences and Countermeasures,” 2018. https://sprs.parl.gov.sg/selectcommittee/selectcommittee/download?id=1&type=subReport
[96] POFMA Office, n.d. https://www.pofmaoffice.gov.sg/
[97] Ministry of Home Affairs, “Protection from Online Falsehoods and Manipulation Act,” Singapore Statutes Online, 2019. https://sso.agc.gov.sg/Act/POFMA2019
[98] Ministry of Home Affairs, “Foreign Interference (Countermeasures) Act,” Singapore Statutes Online, 2021. https://sso.agc.gov.sg/Act/FICA2021
[99] Ministry of Home Affairs, “Online Safety (Miscellaneous Amendments) Act,” Singapore Statutes Online, 2022. https://sso.agc.gov.sg/Acts-Supp/38-2022/Published/20221221?DocDate=20221221&WholeDoc=1
[100] Ministry of Home Affairs, “Summary Factsheet on FICA,” Government of Singapore, 2021. https://www.mha.gov.sg/docs/default-source/default-document-library/summary-factsheet-on-fica.pdf
[101] 総務省, “インターネット上のフェイクニュースや偽情報への対策,” n.d. https://www.soumu.go.jp/main_sosiki/joho_tsusin/d_syohi/ihoyugai_05.html
[102] 総務省, “プラットフォームサービスに関する研究会 中間報告書,” 2019. https://www.soumu.go.jp/main_content/000613197.pdf
[103] 総務省, “プラットフォームサービスに関する研究会 最終報告書,” 2020. https://www.soumu.go.jp/main_content/000668595.pdf
[104] 一般社団法人セーファーインターネット協会, “Disinformation 対策フォーラム,” 2020. https://www.saferinternet.or.jp/anti-disinformation/
[105] 内閣官房, “国家安全保障戦略について,” 2022. https://www.cas.go.jp/jp/siryou/221216anzenhoshou.html
[106] 内閣官房サイバー安全保障体制整備準備室, “サイバー安全保障分野での対応能力の向上に向けた有識者会議 これまでの議論の整理（案）概要,” 2024. https://www.cas.go.jp/jp/seisaku/cyber_anzen_hosyo/dai3/siryou2-1.pdf
[107] NHK, “「国家サイバー統括室」7 月 1 日発足を閣議決定,” 2025. https://www3.nhk.or.jp/news/html/20250620/k10014839811000.html
[108] 内閣官房サイバー安全保障体制整備準備室, “サイバー対処能力強化法及び同整備法について,” 2025. https://www.cas.go.jp/jp/seisaku/cyber_anzen_hosyo_torikumi/pdf/setsumei.pdf
[109] 日本経済新聞, “警察・自衛隊、能動的サイバー防御で合同拠点設置へ,” 2025. https://www.nikkei.com/article/DGXZQOUA29CLE0Z20C25A1000000/
[110] 日本経済新聞, “外国からの選挙介入監視　参院選で大量の自動投稿,” 日経電子版, 2025. https://www.nikkei.com/article/DGKKZO90344210Z20C25A7PD0000/
[111] Disinformation 対策フォーラム, “Disinformation 対策フォーラム 報告書,” 2022. https://www.saferinternet.or.jp/wordpress/wp-content/uploads/Disinformation_report.pdf
[112] 外務省, “偽情報の拡散を含む情報操作への対応,” 2025. https://www.mofa.go.jp/mofaj/gaiko/pagew_000001_00550.html
[113] 曽我部真裕, “「情報流通の健全性」と憲法,” デジタル空間における情報流通の健全性確保の在り方に関する検討会ワーキンググループ（第 14 回）配付資料 WG14－2, 2024. https://www.soumu.go.jp/main_content/000942293.pdf
[114] A. Kozyreva, S. M. Herzog, S. Lewandowsky, R. Hertwig, P. Lorenz-Spreen, M. Leiser, and J. Reifler, “Resolving content moderation dilemmas between free speech and harmful misinformation,” in Proceedings of the National Academy of Sciences of the United States of

America (PNAS), vol. 120, no. 7, article no. e2210666120, 2023. https://doi.org/10.1073/pnas.2210666120

[115] 総務省, “プラットフォーム事業者ヒアリングの結果（案）,” デジタル空間における情報流通の健全性確保の在り方に関する検討会（第 22 回）配付資料 22-1-2, 2024. https://www.soumu.go.jp/main_content/000951298.pdf

[116] キャス・サンスティーン, “インターネットは民主主義の敵か,” 石川幸憲（訳）, 毎日新聞社, 2003.

[117] J. Suler, “The online disinhibition effect,” Cyberpsychology and Behavior, vol. 7, pp. 321–326, 2004. https://doi.org/10.1089/1094931041291295

[118] M. J. Metzger, A. J. Flanagin, and R. B. Medders, “Social and Heuristic Approaches to Credibility Evaluation Online,” Journal of Communication, vol. 60, pp. 413–439, 2010. https://doi.org/10.1111/j.1460-2466.2010.01488.x

[119] イーライ・パリサー, “閉じこもるインターネット―グーグル・パーソナライズ・民主主義,” 井口耕二（訳）, 早川書房, 2012.

[120] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, “Experimental evidence of massive-scale emotional contagion through social networks,” in Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 111, no. 24, pp. 8788–8790, 2014. https://doi.org/10.1073/pnas.1320040111

[121] K. Riddell, “Using fake news against opposing views,” The Washington Times, 2016. https://www.washingtontimes.com/news/2016/nov/24/using-fake-news-against-opposing-views/

[122] M. Zuckerberg, Facebook post, 2016/11/13. https://www.facebook.com/zuck/posts/10103253901916271

[123] M. Zuckerberg, Facebook post, 2016/11/18. https://www.facebook.com/zuck/posts/10103269806149061

[124] S. T. Dennis, “Google, Facebook, Twitter Asked to Appear at Second Senate Panel,” Bloomberg, 2017. https://www.bloomberg.com/news/articles/2017-10-25/google-facebook-twitter-asked-to-appear-at-second-senate-panel (現在参照不可)

[125] M. Reardon, T. Collins, L. Hautala, R. Nieva, and A. Ng, “Congress grills Facebook, Twitter, Google over Russian influence,” CNET, 2017. https://www.cnet.com/news/politics/congress-grills-facebook-twitter-google-over-russian-influence/

[126] European commission, “2018 Code of Practice on Disinformation,” 2022. https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation

[127] X, “X (Twitter Japan 株式会社)ヒアリングシート回答,” 総務省デジタル空間における情報流通の健全性確保の在り方に関する検討会（第 15 回）配布資料 15-2-2, 2024. https://www.soumu.go.jp/main_content/000938665.pdf

[128] S. Perez, “Twitter Releases New Suite Of Anti-Harassment Tools, Promises Faster Response Times For Dealing With Abuse,” TechCrunch, 2014. https://techcrunch.com/2014/12/02/twitter-releases-new-suite-of-anti-harassment-tools-promises-faster-response-times/

[129] X Blog, “Building a safer Twitter,” 2014. https://blog.x.com/official/en_us/a/2014/building-a-safer-twitter.html

[130] P. Cartes, “Announcing the Twitter Trust & Safety Council,” X Blog, 2016. https://blog.x.com/official/en_us/a/2016/announcing-the-twitter-trust-safety-council.html

[131] V. Gadde and B. Falck, “Increasing transparency for political campaigning ads on Twitter,” X Blog, 2018. https://blog.x.com/en_us/topics/company/2018/Increasing-Transparency-for-Political-Campaigning-Ads-on-Twitter

[132] V. Gadde and Y. Roth, “Enabling further research of information operations on Twitter,” X Blog, 2018. https://blog.x.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter

[133] Twitter, “Twitter’s Recommendation Algorithm,” X Engineering, 2023. https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

[134] K. Coleman, “Building a better Birdwatch,” X Blog, 2022.

https://blog.x.com/en_us/topics/company/2022/building-a-better-birdwatch
[135] Engineering [@XEng], X, 2024/06/12. https://x.com/XEng/status/1800634371906380067
[136] X Blog, “Defining public interest on Twitter,” 2019.
https://blog.x.com/en_us/topics/company/2019/publicinterest
[137] J. Vincent, “Twitter is bringing its ‘read before you retweet’ prompt to all users,” The Verge, 2020. https://www.theverge.com/2020/9/25/21455635/twitter-read-before-you-tweet-article-prompt-rolling-out-globally-soon
[138] V. Gadde and K. Beykpour, “Additional steps we’re taking ahead of the 2020 US Election,” X Blog, 2020. https://blog.x.com/en_us/topics/company/2020/2020-election-changes
[139] Y. Roth and N. Pickles, “Updating our approach to misleading information,” X Blog, 2020.
https://blog.x.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information
[140] V. Gadde and K. Beykpour, “An update on our work around the 2020 US Elections,” X Blog, 2020. https://blog.twitter.com/en_us/topics/company/2020/2020-election-update
[141] Meta Platforms, “Meta Platforms, Inc.ヒアリングシート回答,” 総務省デジタル空間における情報流通の健全性確保の在り方に関する検討会（第 14 回）配布資料 14-2-2, 2024.
https://www.soumu.go.jp/main_content/000938002.pdf
[142] WIRED, “Public Posting Now the Default on Facebook,” 2009.
https://www.wired.com/2009/12/facebook-privacy-update/
[143] K. ZETTER, “米国家情報長官：個人情報収集「PRISM」報道は誤解,” M. YAGURA and H. GOHARA/GALILEO（訳）, 2013. https://wired.jp/2013/06/11/prism-faq/
[144] The Guardian, “Facebook bows to pressure on privacy settings for new users,” 2014.
https://www.theguardian.com/technology/2014/may/22/facebook-privacy-settings-changes-users
[145] Meta, “Making It Easier to Share With Who You Want,” 2014.
https://about.fb.com/news/2014/05/making-it-easier-to-share-with-who-you-want/
[146] Meta, “Explaining Our Community Standards and Approach to Government Requests,” 2015.
https://about.fb.com/news/2015/03/explaining-our-community-standards-and-approach-to-government-requests/
[147] Meta, “Further Reducing Clickbait in Feed,” 2016. https://about.fb.com/news/2016/08/news-feed-fyi-further-reducing-clickbait-in-feed/
[148] A. Mosseri, “Working to Stop Misinformation and False News,” Meta, 2017.
https://about.fb.com/news/2017/04/working-to-stop-misinformation-and-false-news/
[149] Meta, “Hard Questions: How Is Facebook’s Fact-Checking Program Working?” 2018.
https://about.fb.com/news/2018/06/hard-questions-fact-checking/
[150] G. Rosen and T. Lyons, “Remove, Reduce, Inform: New Steps to Manage Problematic Content,” Meta, 2019. https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/
[151] J. Kaplan, “More Speech and Fewer Mistakes,” Meta, 2025.
https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/
[152] A. Anker, S. Su, and J. Smith, “New Test to Provide Context About Articles,” Meta, 2017.
https://about.fb.com/news/2017/10/news-feed-fyi-new-test-to-provide-context-about-articles/
[153] T. Hughes, J. Smith, and A. Leavitt, “Helping People Better Assess the Stories They See in News Feed with the Context Button,” Meta, 2018. https://about.fb.com/news/2018/04/news-feed-fyi-more-context/
[154] J. Hegeman, “Providing People With Additional Context About Content They Share,” Meta, 2020. https://about.fb.com/news/2020/06/more-context-for-news-articles-and-other-content/
[155] Meta, “Meta による Misinformation 及び Disinformation への対応について,” 総務省デジタル空間における情報流通の健全性確保の在り方に関する検討会（第 14 回）資料 14-2-3, 2024. https://www.soumu.go.jp/main_content/000938008.pdf
[156] Google Search, “Our approach to Search,” n.d.
https://www.google.com/intl/en_us/search/howsearchworks/our-approach/
[157] Google, “グーグル ヒアリングシート回答,” 総務省デジタル空間における情報流通の健全性確保の在り方に関する検討会（第 14 回）資料 14-1-2, 2024.
https://www.soumu.go.jp/main_content/000937867.pdf

[158] E. Schmidt, “Google Ideas Becomes Jigsaw,” Medium, 2016. https://medium.com/jigsaw/google-ideas-becomes-jigsaw-bcb5bd08c423
[159] Google News Initiative, “Google ニュースラボ,” 2015. https://newsinitiative.withgoogle.com/ja-jp/resources/google-news-lab/
[160] L. Junius, “More ways to find authoritative information in Europe,” Google, 2019. https://blog.google/around-the-globe/google-europe/more-ways-find-authoritative-information-europe/
[161] R. Gingras, “Labeling fact-check articles in Google News,” Google, 2016. https://blog.google/outreach-initiatives/google-news-initiative/labeling-fact-check-articles-google-news/
[162] C. Sinders, “Toxicity and Tone Are Not The Same Thing: analyzing the new Google API on toxicity, PerspectiveAPI,” Medium, 2017. https://medium.com/@carolinesinders/toxicity-and-tone-are-not-the-same-thing-analyzing-the-new-google-api-on-toxicity-perspectiveapi-14abe4e728b3
[163] Jigsaw, “Perspective API,” 2017. https://www.perspectiveapi.com/
[164] K. Walker and R. Salgado, “Security and disinformation in the U.S. 2016 election,” Google, 2017. https://blog.google/outreach-initiatives/public-policy/security-and-disinformation-us-2016-election/
[165] K. Canegallo, “Fighting disinformation across our products,” Google, 2019. https://blog.google/around-the-globe/google-europe/fighting-disinformation-across-our-products/
[166] P. Nayak, “New ways we’re helping you find high-quality information,” Google, 2022. https://blog.google/products/search/information-literacy/
[167] S. Huntley, “Updates about government-backed hacking and disinformation,” Google, 2020. https://blog.google/threat-analysis-group/updates-about-government-backed-hacking-and-disinformation/
[168] T. Kurian, “Google + Mandiant: Transforming Security Operations and Incident Response,” Google, 2022. https://cloud.google.com/blog/products/identity-security/google-completes-acquisition-of-mandiant?hl=en
[169] H. Cohen, “Bringing fact check information to Google Images,” Google, 2020. https://blog.google/products/search/bringing-fact-check-information-google-images/
[170] J. K. Kearns, “A quick way to learn more about your search results,” Google, 2021. https://blog.google/products/search/about-search-results/
[171] I. Snir and N. Hebbar, “Five new ways to verify info with Google Search,” Google, 2023. https://blog.google/products/search/google-search-new-fact-checking-misinformation/
[172] C. Dunton, “Get helpful context with About this image,” Google, 2023. https://blog.google/products/search/about-this-image-google-search/
[173] S. Gowal and P. Kohli, “Identifying AI-generated images with SynthID,” Google DeepMind, 2023. https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/
[174] P. Kohli, “SynthID Detector — a new portal to help identify AI-generated content,” Google, 2025. https://blog.google/technology/ai/google-synthid-ai-content-detector/
[175] International Fact-Checking Network, “The commitments of the Code of Principles,” 2015. https://ifcncodeofprinciples.poynter.org/the-commitments
[176] Reporters’ Lab, “Fact-Checking Sites Around the World,” Duke University, 2025. https://reporterslab.org/fact-checking/
[177] International Fact-Checking Network, “Signatories,” 2025. https://ifcncodeofprinciples.poynter.org/signatories
[178] Snopes, “About Us,” Snopes Media Group, 1995. https://www.snopes.com/about/
[179] B. Stelter, “Debunkers of Fictions Sift the Net,” The New York Times, 2010. https://www.nytimes.com/2010/04/05/technology/05snopes.html
[180] Snopes, “Fact Check Ratings,” Snopes Media Group, 1995. https://www.snopes.com/fact-check-ratings/
[181] PolitiFact, “The Principles of the Truth-O-Meter: PolitiFact’s methodology for independent fact-checking,” Poynter Institute, 2020. https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/

[182] 一般社団法人セーファーインターネット協会, “SIA、「日本ファクトチェックセンター」を設立,” 2022. https://www.saferinternet.or.jp/info/26099/

[183] 日本ファクトチェックセンター, “JFC への支援と会計,” 2024. https://www.factcheckcenter.jp/jfc-funding/

[184] 古田大輔, “ファクトチェックとは 定義・ルール・手法を解説,” 日本ファクトチェックセンター, 2024. https://www.factcheckcenter.jp/explainer/fact-check/jfc-fact-checking-101/

[185] 楊井人文, “ファクトチェックをとりまく世界と日本の状況・課題,” 総務省プラットフォームサービスに関する研究会（第 8 回）配布資料 2, 2019. https://www.soumu.go.jp/main_content/000621622.pdf

[186] 楊井人文, “日本におけるファクトチェック活動の現状と課題,” 総務省プラットフォームサービスに関する研究会（第 41 回）配布資料 6, 2023. https://www.soumu.go.jp/main_content/000861267.pdf

[187] E. Broda and J. Strömbäck, “Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review,” Annals of the International Communication Association, vol. 48, pp. 139–166, 2024. https://doi.org/10.1080/23808985.2024.2323736

[188] G. Pennycook, A. Bear, E. Collins, and D. G. Rand, “The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings,” Management Science, vol. 66, pp. 4921–5484, 2019. https://doi.org/10.1287/mnsc.2019.3478

[189] J. Ayoub, X. J. Yang, and F. Zhou, “Combat COVID-19 infodemic using explainable natural language processing models,” Information Processing and Management, vol. 58, article no. 102569, 2021. https://doi.org/10.1016/j.ipm.2021.102569

[190] Z. Amin, N. M. Ali, and A. F. Smeaton, “Visual Selective Attention System to Intervene User Attention in Sharing COVID-19 Misinformation,” International Journal of Advanced Computer Science and Applications, vol. 12, 2021. https://dx.doi.org/10.14569/IJACSA.2021.0121005

[191] P. Kim, Z. Fan, L. Fernando, J. Sham, C. Sun, Y. Sun, B. Wright, X. Yang, N. Ross, and D. M. Woodbridge, “Controversy Score Calculation for News Articles,” in Proceedings of the First International Conference on Transdisciplinary AI (TransAI), pp. 56–63, 2019. https://doi.org/10.1109/TransAI46475.2019.00018

[192] P. L. Moravec, R. Minas, and A. R. Dennis, “Fake news on social media: people believe what they want to believe when it makes no sense at all,” MIS Quarterly, vol. 43, no. 4, pp. 1343–1360, 2019. https://www.jstor.org/stable/26848107

[193] T. Celadin, V. Capraro, G. Pennycook, and D. G. Rand, “Displaying News Source Trustworthiness Ratings Reduces Sharing Intentions for False News Posts,” Journal of Online Trust and Safety, vol. 1, no. 5, 2023. https://doi.org/10.54501/jots.v1i5.100

[194] NewsGuard, “Website Rating Process and Criteria,” 2018. https://www.newsguardtech.com/ratings/rating-process-criteria/

[195] J. Lühring, H. Metzler, R. Lazzaroni, A. Shetty, and J. Lasser, “Best practices for source-based research on misinformation and news trustworthiness using NewsGuard,” Journal of Quantitative Description: Digital Media, vol. 5, 2025. https://doi.org/10.51685/jqd.2025.003

[196] M. Mensio and H. Alani, “News Source Credibility in the Eyes of Different Assessors,” in Proceedings of the Conference for Truth and Trust Online 2019, 2019. https://truthandtrustonline.com/wp-content/uploads/2019/09/paper_3.pdf

[197] H. M. Johnson and C. M. Seifert, “Sources of the continued influence effect: When misinformation in memory affects later inferences,” Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 20, no. 6, pp. 1420–1436, 1994. https://psycnet.apa.org/doi/10.1037/0278-7393.20.6.1420

[198] S. Lewandowsky, J. Cook, U. K. H. Ecker, D. Albarracín, M. A. Amazeen, P. Kendeou, D. Lombardi, E. J. Newman, G. Pennycook, E. Porter, D. G. Rand, D. N. Rapp, J. Reifler, J. Roozenbeek, P. Schmid, C. M. Seifert, G. M. Sinatra, B. Swire-Thompson, S. van der Linden, E. K. Vraga, T. J. Wood, and M. S. Zaragoza, “The Debunking Handbook 2020,” 2020.

https://doi.org/10.17910/b7.1182
[199] J. D. Featherstone and J. Zhang, “Feeling angry: the effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude,” Journal of Health Communication, vol. 25, pp. 692–702, 2020. https://doi.org/10.1080/10810730.2020.1838671
[200] J. Cook, “Countering Climate Science Denial and Communicating Scientific Consensus,” in H. Von Storch (Ed.), Oxford Research Encyclopedia: Climate Science Oxford University Press, 2016. https://doi.org/10.1093/acrefore/9780190228620.013.314
[201] T. Harjani, J. Roozenbeek, M. Biddlestone, S. van der Linden, A. Stuart, M. Iwahara, B. Piri, R. Xu, B. Goldberg, and M. Graham, “A Practical Guide to Prebunking Misinformation,” Jigsaw, 2022.
https://interventions.withgoogle.com/static/pdf/A_Practical_Guide_to_Prebunking_Misinformation.pdf
[202] J. Roozenbeek, S. van der Linden, B. Goldberg, S. Rathje, and S. Lewandowsky, “Psychological inoculation improves resilience against misinformation on social media,” Science Advances, vol. 8, article no. eabo6254, 2022. https://doi.org/10.1126/sciadv.abo6254
[203] Google, “Prebunking is a technique to preempt manipulation online,” 2024.
https://prebunking.withgoogle.com/
[204] Jigsaw, “Defanging Disinformation’s Threat to Ukrainian Refugees,” 2023.
https://medium.com/jigsaw/defanging-disinformations-threat-to-ukrainian-refugees-b164dbbc1c60
[205] Jigsaw, “Prebunking to Build Defenses Against Online Manipulation Tactics in Germany,” 2023. https://medium.com/jigsaw/prebunking-to-build-defenses-against-online-manipulation-tactics-in-germany-a1dbfbc67a1a
[206] UNESCO, “Social media campaign: UNESCO invites youth in South East Europe and Turkey to join #ThinkBeforeSharing campaign to celebrate Global Media and Information Literacy Week,” 2021. https://www.unesco.org/en/articles/social-media-campaign-unesco-invites-youth-south-east-europe-and-turkey-join-thinkbeforesharing
[207] Cambridge University’s Social Decision-Making Lab, “Bad News,” 2018.
https://www.getbadnews.com/en
[208] U.S. Department of State’s Global Engagement Center (GEC), U.S. Department of Homeland Security’s Cybersecurity and Infrastructure Security Agency (CISA), Tilt, and University of Cambridge’s Social Decision-Making Lab, “Breaking Harmony Square,” 2020.
https://www.harmonysquare.game/en
[209] F. Lewsey, “Cambridge game ‘pre-bunks’ coronavirus conspiracies,” 2020.
https://www.cam.ac.uk/stories/goviral
[210] M. Basol, J. Roozenbeek, M. Berriche, F. Uenal, W. P. McClanahan, and S. van der Linden, “Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation,” Big Data & Society, vol. 8, no. 1, 2021.
https://doi.org/10.1177/20539517211013868
[211] R. Maertens, J. Roozenbeek, M. Basol, and S. van der Linden, “Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments,” Journal of Experimental Psychology: Applied, vol. 27, pp. 1–16, 2021. https://doi.org/10.1037/xap0000315
[212] リチャード・セイラー, キャス・サンスティーン, “実践 行動経済学 健康, 富, 幸福への聡明な選択,” 遠藤真美（訳）, 日経 BP, 2009.
[213] キャス・R・サンスティーン, “スラッジ: 不合理をもたらすぬかるみ,” 土方奈美（訳）, 早川書房, 2023.
[214] B. Kaiser, J. Wei, E. Lucherini, K. Lee, J. N. Matias, and J. Mayer, “Adapting Security Warnings to Counter Online Disinformation,” in Proceedings of the 30th USENIX Security Symposium, pp. 1163–1180, 2021.
https://www.usenix.org/conference/usenixsecurity21/presentation/kaiser
[215] L. Fazio, “Pausing to consider why a headline is true or false can help reduce the sharing of false news,” Harvard Kennedy School Misinformation Review, vol. 1, 2020.
https://doi.org/10.37016/mr-2020-009

[216] Y. Wang, P. G. Leon, A. Acquisti, L. F. Cranor, A. Forget, and N. Sadeh, "A field trial of privacy nudges for facebook," in Proceedings of the SIGCHI conference on human factors in computing systems, pp. 2367–2376, 2014. https://doi.org/10.1145/2556288.2557413
[217] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D.G. Rand, "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention," Psychological Science, vol. 31, pp. 770–780, 2020. https://doi.org/10.1177/0956797620939054
[218] S. Rathje, J. Roozenbeek, J. J. Van Bavel, and S. van der Linden, "Accuracy and social motivations shape judgements of (mis)information," Nature Human Behaviour, vol. 7, pp. 892–903, 2023. https://doi.org/10.1038/s41562-023-01540-w
[219] S. Rathje, J. Roozenbeek, C. Steenbuch, J. J. Van Bavel, and S. van der Linden, "Letter to the Editors of Psychological Science: Meta-analysis Reveals That Accuracy Nudges Have Little to No Effect for U.S. Conservatives: Regarding Pennycook et al. (2020)," Psychological Science, 2022. https://journals.sagepub.com/page/pss/letters-to-the-eds
[220] G. Pennycook and D. G. Rand, "Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation," Nature Communications, vol. 13, article no. 2333, 2022. https://doi.org/10.1038/s41467-022-30073-5
[221] R. B. Cialdini and M. R. Trost, "Social influence: Social norms, conformity and compliance," in D. T. Gilbert, S. T. Fiske, and G. Lindzey (Eds.), The handbook of social psychology (4th ed., pp. 151–192), 1998.
[222] R. B. Cialdini and N. J. Goldstein, "Social Influence: Compliance and Conformity," Annual Review of Psychology, vol. 55, pp. 591–621, 2004. https://doi.org/10.1146/annurev.psych.55.090902.142015
[223] H. Gimpel, S. Heger, C. Olenberger, and L. Utz, "The effectiveness of social norms in fighting fake news on social media," Journal of Management Information Systems, vol. 38, pp. 196–221, 2021. https://doi.org/10.1080/07421222.2021.1870389
[224] S. Andı and J. Akesson, "Nudging away false news: Evidence from a social norms experiment," Digital Journalism, vol. 9, article no. 106125, 2021. https://doi.org/10.1080/21670811.2020.1847674
[225] A. M. Guess, M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, and N. Sircar, "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India," in Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 117, no. 27, pp. 15536–15545, 2020. https://doi.org/10.1073/pnas.1920498117
[226] OECD, "Misinformation and disinformation," 2022. https://www.oecd.org/en/publications/an-international-effort-using-behavioural-science-to-tackle-the-spread-of-misinformation_b7709d4f-en.html
[227] G. D. S. Martino, S. Shaar, Y. Zhang, S. Yu, A. Barrón-Cedeño, and P. Nakov, "Prta: A System to Support the Analysis of Propaganda Techniques in the News," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 287–293, 2020. https://doi.org/10.18653/v1/2020.acl-demos.32
[228] European Commission, "Co-Creating Misinformation-Resilient Societies," 2018. https://cordis.europa.eu/project/id/770302/results
[229] The Knowledge Media Institute, "Co-inform: Context Matters, Your Sources Too," European Commission, 2018. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5e0667c80&appId=PPGMS
[230] C. Chen and K. Shu, "Combating misinformation in the age of LLMs: Opportunities and challenges," AI Magazine, vol. 45, pp. 354–368, 2024. https://doi.org/10.1002/aaai.12188
[231] E. (G). Park, "I Trust You, but Let Me Talk to AI: The Role of the Chat Agents, Empathy, and Health Issues in Misinformation Guidance," International Journal of Strategic Communication, vol. 19, pp. 231–260, 2025. https://doi.org/10.1080/1553118X.2025.2462087
[232] Q. Lou and W. Xu, "Personality Modeling for Persuasion of Misinformation using AI Agent," 2025. (Preprint) https://doi.org/10.48550/arXiv.2501.08985
[233] R. Hertwig and T. Grüne-Yanoff, "Nudging and Boosting: Steering or Empowering Good

Decisions," Perspectives on Psychological Science, vol. 12, pp. 973–986, 2017. https://doi.org/10.1177/1745691617702496
[234] UNESCO, "Media and information literacy," 2018. https://unesdoc.unesco.org/ark:/48223/pf0000265509
[235] R. R. Tuazon, "UNESCO Media and Information Literacy Framework," UNESCO, 2020. https://www.unesco.gov.ph/wp-content/uploads/2020/03/UNESCO-Media-and-Information-Literacy-Framework-and-Recent-Initiatives.pdf
[236] S. M. Jones-Jang, T. Mortensen, and J. Liu, "Does media literacy help identification of fake news? Information literacy helps, but other literacies don't," American behavioral scientist, vol. 65, pp. 371–388, 2019. https://doi.org/10.1177/0002764219869406
[237] T. D. Adjin-Tettey, "Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education," Cogent Arts & Humanities, vol. 9, article no. 2037229, 2022. https://doi.org/10.1080/23311983.2022.2037229
[238] Media Education Lab, "Mind Over Media," 2017. https://propaganda.mediaeducationlab.com/
[239] Media Education Lab, "Mind Over Media: Analyzing Contemporary Propaganda Lesson Plans," 2018. https://mediaeducationlab.com/sites/default/files/FINAL%20Mind%20Over%20Media%209.17.18.pdf
[240] ALL DIGITAL AISBL, "GetFacts – Get Your Facts Straight!" 2019. https://all-digital.org/projects/get-your-facts-straight/
[241] P. Celot and F. L. EAVI, "GET YOUR FACTS STRAIGHT! (MEDIA LITERACY) TOOLKIT FOR EDUCATORS AND TRAINING PROVIDERS," All Digital, 2020. https://all-digital.org/wp-content/uploads/2020/11/GYFS-Toolkit_for_Educators_and_Training-Providers.pdf
[242] Centre for International Relations, "START2THINK," 2020. https://start2think.info/
[243] Centre for International Relations, "START2THINK - Disinformation techniques," 2020. https://start2think.info/disinformation-techniques/
[244] European Union, "Guidelines for teachers – How to spot and fight disinformation," 2021. https://south.euneighbours.eu/publication/guidelines-teachers-how-spot-and-fight-disinformation/
[245] European Union, "Staying vigilant online: can you spot information manipulation?" 2024. https://learning-corner.learning.europa.eu/learning-materials/staying-vigilant-online-can-you-spot-information-manipulation_en
[246] 総務省, "偽・誤情報に関する啓発教育教材「インターネットとの向き合い方～ニセ・誤情報に騙されないために～」等の公表," 2022. https://www.soumu.go.jp/menu_news/s-news/01ryutsu02_02000340.html
[247] 総務省, "メディア情報リテラシー向上施策の現状と課題等に関する調査結果報告," 2022. https://www.soumu.go.jp/main_content/000820476.pdf
[248] 総務省, "インターネットとの向き合い方～ニセ・誤情報にだまされないために～第2版," 2025. https://www.soumu.go.jp/use_the_internet_wisely/special/nisegojouhou/
[249] AISBL, "SELMA," 2017. https://hackinghate.eu/
[250] AISBL, "SELMA – Social and Emotional Learning," 2017. https://hackinghate.eu/toolkit/focus/social-and-emotional-learning/
[251] S. Wineburg and S. McGrew, "Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information," Teachers College Record, vol. 121, pp. 1–40, 2019. https://doi.org/10.1177/016146811912101102
[252] S. McGrew, "Learning to evaluate: An intervention in civic online reasoning," Computers & Education, vol. 145, article no. 103711, 2020. https://doi.org/10.1016/j.compedu.2019.103711
[253] CIVIX, "CTRL-F," n.d. https://ctrl-f.ca/en/
[254] C. D. Stavrositu and J. Kim, "Social media metrics: Third-person perceptions of health information," Computers in Human Behavior, vol. 35, pp. 61–67, 2014. https://doi.org/10.1016/j.chb.2014.02.025
[255] M. Chung, G. J. Munno, and B. Moritz, "Triggering participation: Exploring the effects of

third-person and hostile media perceptions on online participation," Computers in Human Behavior, vol. 53, pp. 452–461, 2015. https://doi.org/10.1016/j.chb.2015.06.037
[256] L. Ma, C. S. Lee, and D. H. Goh, "That's news to me: The influence of perceived Gratifications and personal experience on news sharing in social media," in Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, pp. 141–144, 2011. https://doi.org/10.1145/1998076.1998103
[257] A. Rudat, J. Buder, and F. W. Hesse, "Audience design in Twitter: Retweeting behavior between informational value and followers' interests," Computers in Human Behavior, vol. 35, pp. 132–139, 2014. https://doi.org/10.1016/j.chb.2014.03.006
[258] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer, "Political polarization on twitter," in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, vol. 5, no. 1, pp. 89–96, 2011. https://doi.org/10.1609/icwsm.v5i1.14126
[259] J. Fox, C. Cruz, and J. Y. Lee, "Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media," Computers in Human Behavior, vol. 52, pp. 436–442, 2015. https://doi.org/10.1016/j.chb.2015.06.024
[260] C. Martel, G. Pennycook, and D. G Rand, "Reliance on emotion promotes belief in fake News," Cognitive Research: Principles and Implications, vol. 5, article no. 47, 2020. https://doi.org/10.1186/s41235-020-00252-3
[261] W. James, "What is an emotion?" Mind, vol. 9, pp. 188–205, 1884. https://doi.org/10.1093/mind/os-IX.34.188
[262] 大平英樹, "感情心理学・入門," 有斐閣アルマ, 2010.
[263] 小川時洋, 飯田沙依亜, "なぜ概念・定義が問題となるのか," 感情心理学研究, 22 巻, 2 号, pp. 83–88, 2015. https://doi.org/10.4092/jsre.22.83
[264] A. Ortony, G. L. Clore, and A. Collins, "The cognitive structure of emotions," Cambridge University Press, 1988.
[265] C. Skurka and R. L. Nabi, "Perspectives on Emotion in the Digital Age," in R. L. Nabi and J. G. Myrick (Eds.) , Emotions in the Digital World: Exploring Affective Experience and Expression in Online Interactions (pp. 7–31), Oxford University Press, 2023.
[266] G. H. Bower, "Mood and memory," American Psychologist, vol. 36, no. 2, pp. 129–148, 1981. https://doi.org/10.1037/0003-066X.36.2.129
[267] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and Decision Making," Annual Review of Psychology, vol. 66, pp. 799–823, 2015. https://doi.org/10.1146/annurev-psych-010213-115043
[268] H. Bless and K. Fiedler, "Mood and the regulation of information processing and behavior," in J. P. Forgas (Ed.), Hearts and Minds: Affective Influences on Social Cognition and Behavior (pp. 65–84), New York: Psychology Press, 2006.
[269] B. Rimé, B. Mesquita, P. Philippot, and S. Boca, "Beyond the emotional event: Six studies on the social sharing of emotion," Cognition and Emotion, vol. 5, pp. 435–465, 1991. https://doi.org/10.1080/02699939108411052
[270] V. Christophe and B. Rimé, "Exposure to the social sharing of emotion: Emotional impact, listener responses and secondary social sharing," European Journal of Social Psychology, vol. 27, pp. 37–54, 1997. https://doi.org/10.1002/(SICI)1099-0992(199701)27:1<37::AID-EJSP806>3.0.CO;2-1
[271] 三浦麻子, 小森政嗣, 松村真宏, 平石界, "ソーシャルメディアにおけるネガティブ情動の社会的共有――東日本大震災関連ツイートの長期的変化――," エモーション・スタディーズ, 第 4 巻, pp. 26–32, 2019. https://doi.org/10.20797/ems.4.Si_26
[272] A. Al-Rawi, "Viral News on Social Media," Digital Journalism, vol. 7, pp. 63–79, 2017. https://doi.org/10.1080/21670811.2017.1387062
[273] J. Berger, "Contagious: Why Things Catch on," New York: Simon and Schuster, 2016.
[274] A. Dobele, A. Lindgreen, M. B. Beverland, J. Vanhamme, and R. van Wijk, "Why pass on viral messages? Because they connect emotionally," Business Horizons, vol. 50, no. 4, pp. 291–

304, 2007. https://doi.org/10.1016/j.bushor.2007.01.004
[275] J. Berger and K. Milkman, “Social transmission, emotion, and the virality of online content,” Marketing Science Institute Working Paper Series 2010, report no. 10–114, 2010. https://www.msi.org/?post_type=resources&p=1994
[276] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” Science, vol. 359, pp. 1146–1151, 2018. https://doi.org/10.1126/science.aap9559
[277] C. Guo, J. Cao, X. Zhang, K. Shu, and M. Yu, “Exploiting Emotions for Fake News Detection on Social Media,” 2019. (Preprint) https://arxiv.org/abs/1903.01728v1
[278] H. Kassinove and D. G. Sukhodolsky, “Anger disorders: Basic science and practice issues,” H. Kassinove (Ed.), Anger disorders: Definition, diagnosis, and treatment. Taylor and Francis, Philadelphia, 1995.
[279] J. Averill, “Anger and Aggression: An Essay on Emotion,” New York, NY: Springer-Verlag, 1982.
[280] E. Shuman, E. Halperin, and M. Reifen-Tagar, “Anger as a catalyst for change? Incremental beliefs and anger’s constructive effects in conflict,” Group Processes & Intergroup Relations, vol. 21, pp. 1092–1106, 2018. https://doi.org/10.1177/1368430217695442
[281] Y. Elsayed and A. B. Hollingshead, “Humor reduces online incivility,” Journal of Computer-Mediated Communication, vol. 27, zmac005, 2022. https://doi.org/10.1093/jcmc/zmac005
[282] W. J. Brady and M. J. Crockett, “How effective is online outrage?” Trends in Cognitive Sciences, vol. 23, pp. 79–80, 2019. https://doi.org/10.1016/j.tics.2018.11.004
[283] J. S. Lerner and L. Z. Tiedens, “Portrait of The Angry Decision Maker: How Appraisal Tendencies Shape Anger’s Influence on Cognition,” Journal of Behavioral Decision Making, vol. 19, pp. 115–137, 2006. https://doi.org/10.1002/bdm.515
[284] O. Luminet, P. Bouts, F. Delie, and A. S. R. Manstead, “Social sharing of emotion following exposure to a negatively valenced situation,” Cognition and Emotion, vol. 14, pp. 661–688, 2000. https://doi.org/10.1080/026999300501176666
[285] J. Han, M. Cha, and W. Lee, “Anger contributes to the spread of COVID-19 misinformation,” Harvard Kennedy School (HKS) Misinformation Review, 2020. https://doi.org/10.37016/mr-2020-39
[286] G. V. Bodenhausen, L. A. Sheppard, and G. P. Kramer, “Negative affect and social judgment: The differential impact of anger and sadness,” European Journal of Social Psychology, vol. 24, pp. 45–62, 1994. https://doi.org/10.1002/ejsp.2420240104
[287] 朴喜静, 大坊郁夫, “怒りと悲しみが真偽性判断の正答率に及ぼす影響,” 応用心理学研究, 40 巻, 1 号, pp. 1−10, 2014. http://id.ndl.go.jp/bib/026147220
[288] B. Bago, D. G. Rand, and G. Pennycook, “Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines,” Journal of Experimental Psychology: General, vol. 149, pp. 1608–1613, 2020. https://doi.org/10.1037/xge0000729
[289] ダニエル・カーネマン, “ファスト&スロー(上) あなたの意思はどのように決まるか?” 早川書房, 2014.
[290] R. R. Stains Jr. and J. Sarrouf, “Hard to say, hard to hear, heart to heart: Inviting and harnessing strong emotions in dialogue for deliberation,” Journal of Deliberative Democracy, vol. 18, 2022. https://doi.org/10.16997/jdd.979
[291] J. Kiskola, T. Olsson, H. Väätäjä, A. H. Syrjämäki, A. Rantasila, P. Isokoski, M. Ilves, and V. Surakka, “Applying critical voice in design of user interfaces for supporting self-reflection and emotion regulation in online news commenting,” in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI’21), article no. 88, pp. 1–13, 2021. https://doi.org/10.1145/3411764.3445783
[292] J. Kiskola, T. Olsson, A. H. Syrjämäki, A. Rantasila, M. Ilves, P. Isokoski, and V. Surakka, “Online Survey on Novel Designs for Supporting Self-Reflection and Emotion Regulation in Online News Commenting,” in Proceedings of the 25th International Academic Mindtrek Conference, pp. 278–312, 2022. https://doi.org/10.1145/3569219.3569411
[293] A. H. Syrjämäki, M. Ilves, J. Kiskola, A. Rantasila, P. Isokoski, T. Olsson, and V. Surakka, “Facilitating Implicit Emotion Regulation in Online News Commenting—An Experimental

Vignette Study," Interacting with Computers, vol. 34, pp. 129–136, 2022. https://doi.org/10.1093/iwc/iwad010
[294] J. J. Gross, "The emerging field of emotion regulation: An integrative review," Review of General Psychology, vol. 2, pp. 271–299, 1998. https://doi.org/10.1037/1089-2680.2.3.271
[295] D. J. Siegel, "The developing mind: Toward a neurobiology of interpersonal experience," Guilford Press, 1999.
[296] NHS Fife Psychology Department, "Emotion Regulation: Managing Emotions," NHS Fife, 2016. https://www.moodcafe.co.uk/media/fselmngo/er_handout_final_16_june_2016.pdf
[297] J. J. Gross and R. A. Thompson, "Emotion Regulation: Conceptual Foundations," in J. J. Gross (Ed.), Handbook of emotion regulation (pp. 3–24), The Guilford Press, 2007.
[298] C. L. Rusting and S. Nolen-Hoeksema, "Regulating responses to anger: Effects of rumination and distraction on angry mood," Journal of Personality and Social Psychology, vol. 74, pp. 790–803, 1998. https://doi.org/10.1037/0022-3514.74.3.790
[299] T. L. Webb, E. Miles, and P. Sheeran, "Dealing with feeling: A meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation," Psychological Bulletin, vol. 138, pp. 775–808, 2012. https://doi.org/10.1037/a0027600
[300] C. Watson and C. Purdon, "Attention training in the reduction and reappraisal of intrusive thoughts," Behavioural and Cognitive Psychotherapy, vol. 36, pp. 61–70, 2007. https://doi.org/10.1017/S1352465807003773
[301] C. A. Hutcherson, E. M. Seppala, and J. J. Gross, "Loving-kindness meditation increases social connectedness," Emotion, vol. 8, no. 5, pp.720–724, 2008. https://doi.org/10.1037/a0013237
[302] B. J. Bushman, "Does venting anger feed or extinguish the flame? Catharsis, rumination, distraction, anger and aggressive responding," Personality and Social Psychology Bulletin, vol. 28, pp. 724–731, 2002. https://psycnet.apa.org/doi/10.1177/0146167202289002
[303] R. D. Ray, F. H. Wilhelm, and J. J. Gross, "All in the Mind's Eye Anger Rumination and Reappraisal," Journal of Personality and Social Psychology, vol. 94, pp. 133–145, 2008. https://doi.org/10.1037/0022-3514.94.1.133
[304] T. F Denson, M. L. Moulds, and J. R. Grisham, "The effects of analytical rumination, reappraisal, and distraction on anger experience," Behavior Therapy, vol. 43, pp. 355–364, 2012. https://doi.org/10.1016/j.beth.2011.08.001
[305] J. M. Richards, E. A. Butler, and J. J. Gross, "Emotion Regulation in Romantic Relationships: The Cognitive Consequences of Concealing Feelings," Journal of Social and Personal Relationships, vol. 20, pp. 599–620, 2003. https://doi.org/10.1177/02654075030205002
[306] K. N. Ochsner, R. D. Ray, J. C. Cooper, E. R. Robertson, S. Chopra, J. D. E. Gabrieli, and J. J. Gross, "For better or for worse: Neural systems supporting the cognitive down- and up-regulation of negative emotion," NeuroImage, vol. 23, pp. 483–499, 2004. https://doi.org/10.1016/j.neuroimage.2004.06.030
[307] K. McRae, K. N. Ochsner, I. B. Mauss, J. J. D. Gabrieli, and J. J. Gross, "Gender differences in emotion regulation: An fMRI study of cognitive reappraisal," Group Processes & Intergroup Relations, vol. 11, pp. 143–162, 2008. https://doi.org/10.1177/1368430207088035
[308] K. McRae, B. Hughes, S. Chopra, J. D. E. Gabrieli, J. J. Gross, and K. N. Ochsner, "The neural bases of distraction and reappraisal," Journal of Cognitive Neuroscience, vol. 22, pp. 248–262, 2010. https://doi.org/10.1162/jocn.2009.21243
[309] R. Kalisch, K. Wiech, H. D. Critchley, B. Seymour, J. P. O'Doherty, D. A. Oakley, P. Allen, and R. J. Dolan, "Anxiety reduction through detachment: Subjective, physiological, and neural effects," Journal of Cognitive Neuroscience, vol. 17, pp. 874–883, 2005. https://doi.org/10.1162/0898929054021184
[310] M. H. Davis, "A Multidimensional Approach to Individual Differences in Empathy," JSAS Catalog of Selected Documents in Psychology, vol. 10, pp. 85, 1980. https://cir.nii.ac.jp/crid/1571135650071470080
[311] M. H. Davis, "Measuring individual differences in empathy: Evidence for a multidimensional approach," Journal of Personality and Social Psychology, vol. 44, pp. 113–126, 1983. https://doi.org/10.1037/0022-3514.44.1.113
[312] A. D. Galinsky and G. B. Moskowitz, "Perspective-taking: Decreasing stereotype expression,

stereotype accessibility, and in-group favoritism," Journal of Personality and Social Psychology, vol. 78, pp. 708–724, 2000. https://doi.org/10.1037/0022-3514.78.4.708
[313] G. Sheppes and N. Meiran, "Better late than never? On the dynamics of online regulation of sadness using distraction and cognitive reappraisal," Personality and Social Psychology Bulletin, vol. 33, pp. 1518–1532, 2007. https://doi.org/10.1177/0146167207305537
[314] E. Halperin, R. Pliskin, T. Saguy, V. Liberman, and J. J. Gross, "Emotion regulation and the cultivation of political tolerance: Searching for a new track for intervention," The Journal of Conflict Resolution, vol. 58, pp. 1110–1138, 2014. https://doi.org/10.1177/0022002713492636
[315] T. Roberton, M. Daffern, and R. S. Bucks, "Beyond anger control: Difficulty attending to emotions also predicts aggression in offenders," Psychology of Violence, vol. 5, pp. 74–83, 2015. https://doi.org/10.1037/a0037214
[316] Y. Nozaki and M. Mikolajczak, "Effectiveness of extrinsic emotion regulation strategies in text-based online communication," Emotion, vol. 23, pp. 1714–1725, 2023. https://doi.org/10.1037/emo0001186
[317] M. Groh, C. Ferguson, R. Lewis, and R. W. Picard, "Computational Empathy Counteracts the Negative Effects of Anger on Creative Problem Solving," in Proceedings of the 10th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 200–211, 2022. https://doi.org/10.1109/ACII55700.2022.9953869
[318] L. F. Barrett and J. J. Gross, "Emotional intelligence: A process model of emotion representation and regulation," in Emotions: Current issues and future directions, T. J. Mayne and G. A. Bonanno (Eds.), The Guilford Press, pp. 286–310, 2001.
[319] H. K. Lennarz, A. Lichtwarck-Aschoff, M. E. Timmerman, and I. Granic, "Emotion differentiation and its relation with emotional well-being in adolescents," Cognition and Emotion, vol. 32, pp. 651–657, 2018. https://doi.org/10.1080/02699931.2017.1338177
[320] T. L. Webb, I. S. Gallo, E. Miles, P. M. Gollwitzer, and P. Sheeran, "Effective regulation of affect: An action control perspective on emotion regulation," European Review of Social Psychology, vol. 23, pp. 143–186, 2012. https://doi.org/10.1080/10463283.2012.718134
[321] M. Tamir, "Effortful Emotion Regulation as a Unique Form of Cybernetic Control," Perspectives on Psychological Science, vol. 16, pp. 94–117, 2021. https://doi.org/10.1177/1745691620922199
[322] A. Acquisti, I. Adjerid, R. H. Balebako, L. Brandimarte, L. Cranor, S. Komanduri, P. G. Leon, N. M. Sadeh, F. Schaub, M. Sleeper, Y. Wang, and S. Wilson, "Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online," ACM Computing Surveys (CSUR), vol. 50, article no. 44, pp. 1–41, 2017. https://doi.org/10.1145/3054926
[323] P. Kuyer and B. Gordijn, "Nudge in perspective: A systematic literature review on the ethical issues with nudging," Rationality and Society, vol. 35, pp. 191–230, 2023. https://doi.org/10.1177/10434631231155005
[324] D. Kanev and V. Terziev, "Behavioral Economics: Development, Condition and Perspectives," Business Economics, vol. 52, pp. 387–410, 2017. https://ssrn.com/abstract=3145668
[325] W. Glod, "How Nudges Often Fail to Treat People According to Their Own Preferences: Social Theory and Practice," Social Theory and Practice, vol. 41, no. 4, pp. 599–617, 2015. https://www.jstor.org/stable/24575751
[326] P. G. Hansen and A. M. Jespersen, "Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy," European Journal of Risk Regulation, vol. 4, pp. 3–28, 2013. https://ssrn.com/abstract=2555337
[327] A. Caraban, E. Karapanos, D. Gonçalves, and P. Campos, "23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, paper no. 503, pp. 1–15, 2019. https://doi.org/10.1145/3290605.3300733
[328] K. Hartwig, F. Doell, and C. Reuter, "The Landscape of User-centered Misinformation Interventions – A Systematic Literature Review," ACM Computing Surveys, vol. 56, article no. 292, pp. 1–36, 2024. https://doi.org/10.1145/3674724
[329] A. V. Pandey, A. Manivannan, O. Nov, M. Satterthwaite, and E. Bertini, "The Persuasive Power of Data Visualization," IEEE Transactions on Visualization and Computer Graphics, vol.

20, pp. 2211–2220, 2014. https://doi.org/10.1109/TVCG.2014.2346419
[330] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor, "How does your password measure up? The effect of strength meters on password creation," in Proceedings of the 21st USENIX conference on Security symposium, pp. 1–16, 2012. https://dl.acm.org/doi/10.5555/2362793.2362798
[331] H. Liu, H. Lieberman, and T. Selker, "Automatic affective feedback in an email browser," MIT Media Lab Software Agents Group', 2002. https://agents.media.mit.edu/projects/emotusponens/empathybuddy.pdf
[332] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4040–4054, 2020. https://doi.org/10.18653/v1/2020.acl-main.372, https://research.google/blog/goemotions-a-dataset-for-fine-grained-emotion-classification/
[333] X. A. Li and D. Parikh, "Lemotif: An Affective Visual Journal Using Deep Neural Networks," in Proceedings of the International Conference on Computational Creativity (ICCC), pp. 453–460, 2019. https://ai.meta.com/research/publications/lemotif-an-affective-visual-journal-using-deep-neural-networks/
[334] J. M. B. Fugate and C. L. Franco, "What Color Is Your Anger? Assessing Color-Emotion Pairings in English Speakers," Frontiers in Psychology, vol. 10, article no. 206, 2019. https://doi.org/10.3389/fpsyg.2019.00206
[335] B. Ur, F. Alfieri, M. Aung, L. Bauer, N. Christin, J. Colnago, L. F. Cranor, H. Dixon, P. E. Naeini, H. Habib, N. Johnson, and W. Melicher, "Design and Evaluation of a Data-Driven Password Meter," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 3775–3786, 2017. https://doi.org/10.1145/3025453.3026050
[336] S. Domgaard and M. Park, "Combating misinformation: The effects of infographics in verifying false vaccine news," Health Education Journal, vol. 80, pp. 974–986, 2021. https://doi.org/10.1177/00178969211038750
[337] キャス・サンスティーン，ルチア・ライシュ, "データで見る行動経済学 全世界大規模調査で見えてきた「ナッジ（NUDGES）の真実」," 大竹文雄（監修），遠藤真美（訳），日経 BP, 2020.
[338] Testimoniun Ltd., "Deceptive Patterns," 2023. https://www.deceptive.design/types
[339] N. Levy, "Nudges in a post-truth world," Journal of Medical Ethics, vol. 43, pp. 495–500, 2017. https://doi.org/10.1136/medethics-2017-104153
[340] 山根承子, "ナッジ研究における諸課題—倫理的観点から—," 第 29 回日本健康教育学会学術大会, 30 巻, 1 号, pp. 68–72, 2022. https://doi.org/10.11260/kenkokyoiku.30.68
[341] 日本版ナッジ・ユニット BEST, "ナッジ等の行動インサイトの活用に係る倫理チェックリスト ①調査・研究編," 2020. https://www.env.go.jp/content/900447984.pdf
[342] M. K. Lee, S. Kiesler, and J. Forlizzi, "Mining behavioral economics to design persuasive technology for healthy choices," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 325–334, 2011. https://doi.org/10.1145/1978942.1978989
[343] C. R. Sunstein, "Nudges that fail," Behavioural Public Policy, vol. 1, pp. 4–25, 2017. https://doi.org/10.1017/bpp.2016.3
[344] H. Allcott and T. Rogers, "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation," American Economic Review, vol. 104, no. 10, pp. 3003–3037, 2014. https://doi.org/10.1257/aer.104.10.3003
[345] S. DellaVigna and E. Linos, "RCTs to Scale: Comprehensive Evidence from Two Nudge Units," Econometrica, vol. 90, no. 1, pp. 81–116, 2022. https://doi.org/10.3982/ECTA18709
[346] A. Lohn, "Disinformation At Scale: Using GPT-3 Maliciously for Information Operations," BLACK HAT USA 2021. https://i.blackhat.com/USA21/Wednesday-Handouts/us-21-Disinformation-At-Scale-Using-Gpt-3-Maliciously-For-Information-Operations.pdf
[347] Innovation Nippon, "フェイクニュース with コロナ時代の情報環境と社会的対処," 2021. http://www.innovation-nippon.jp/?p=840

[348] TrustLab, “Code of Practice on Disinformation,” 2023. https://www.trustlab.com/resources/codeofpractice-disinformation#report
[349] 総務省, “新型コロナウイルス感染症に関する情報流通調査,” 2020. https://www.soumu.go.jp/main_content/000693280.pdf
[350] House Permanent Select Committee on Intelligence (HPSCI), “Social Media Advertisements,” Permanent Select Committee on Intelligence Democrats, n.d. https://democrats-intelligence.house.gov/social-media-content/social-media-advertisements.htm
[351] 高橋幸市, 村田ひろ子, “社会の関心が低い人々の特徴〜「社会と生活に関する世論調査」から〜,” 放送研究と調査（月報）, 8 月号, 2011. https://www.nhk.or.jp/bunken/summary/yoron/social/052.html
[352] R. Plutchik, “The Nature of Emotions: Clinical Implications,” In: Emotions and Psychopathology, M. Clynes and J. Panksepp (Eds.) pp. 1–20. Springer, Boston, MA, 1988.
[353] K. Solovev and N. Pröllochs, “Moral Emotions Shape the Virality of COVID-19 Misinformation on social media,” in Proceedings of ACM Web Conference 2022, pp. 3706–3717, 2022. https://doi.org/10.1145/3485447.3512266
[354] H. Lee and H. J. Oh, “Normative Mechanism of Rumor Dissemination on Twitter,” Cyberpsychology, Behavior, and Social Networking, vol. 20, no. 3, pp. 164–171, 2017. https://doi.org/10.1089/cyber.2016.0447
[355] L. Muradova, “Seeing the Other Side? Perspective-Taking and Reflective Political Judgements in Interpersonal Deliberation,” Political Studies, vol. 69, pp. 644–664, 2021. https://doi.org/10.1177/0032321720916605
[356] M. Komori, A. Miura, N. Matsumura, K. Hiraishi, and K. Maeda, “Spread of risk information through microblogs: Twitter users with more mutual connections relay news that is more dreadful,” Japanese Psychological Research, vol. 63, pp. 1–12, 2019. https://doi.org/10.1111/jpr.12272
[357] S. Mishra, “Social Media Information Extraction: multi-task, multi-lingual, & multi-contextual,” The Data Science Conference, 2022. https://shubhanshu.com/assets/talks/data_science_conf2022.pdf
[358] J. Garson, “How to analyze the sentiment of your own Tweets,” X Developer Platform, 2020. https://developer.x.com/en/blog/community/2020/how-to-analyze-the-sentiment-of-your-own-posts
[359] M. Uniyal, “Twitter Sentiment Analysis Project,” 2025. https://www.appliedaicourse.com/blog/twitter-sentiment-analysis/
[360] M. Lewandowski, “Real-Time Twitter Sentiment Analysis and Prediction App with Pathway,” 2022. https://pathway.com/developers/templates/etl/twitter/
[361] drisskhattabi6, “Real-Time-Twitter-Sentiment-Analysis,” 2024. https://github.com/drisskhattabi6/Real-Time-Twitter-Sentiment-Analysis
[362] Yahoo! JAPAN Tech Blog, “ポジティブ？ネガティブ？ツイートの感情分析に BERT を活用した事例紹介 〜 学習データのラベル偏りに対する取り組み,” 2021. https://techblog.yahoo.co.jp/entry/2021051730150930/
[363] ikegami-yukino, “pymlask,” 2024. https://github.com/ikegami-yukino/pymlask
[364] Jigsaw, “API について,” 2017. https://developers.perspectiveapi.com/s/about-the-api?language=ja
[365] S. Perez, “Actually, X sees 500M posts per day — not 100M-200M as Musk recently said,” TechCrunch, 2023. https://techcrunch.com/2023/10/04/actually-x-sees-500m-posts-per-day-not-100m-200m-as-musk-recently-said/
[366] F. Duarte, “X (Formerly Twitter) User Age, Gender, & Demographic Stats (2025),” Exploding Topics, 2025. https://explodingtopics.com/blog/x-user-stats
[367] dentsu promotion plus, “SNS の投稿から感情を読む、テキスト感情分析技術の実力,” 2020. https://www.dentsu-pmp.co.jp/contents-bae/cinc
[368] X ヘルプセンター, “X プレミアムについて,” 2023. https://help.x.com/ja/using-x/x-premium

[369] 日本経済新聞, “コロナ禍で「女性不況」　男女共同参画白書,” 2021 年 6 月 11 日. https://www.nikkei.com/article/DGXZQOUA101LF0Q1A610C2000000/

[370] 株式会社ラークポート, “＜働くみんなのホンネ調査＞ 「ジェンダーハラスメント」について調査を実施,” PR TIMES, 2020 年 2 月 27 日. https://prtimes.jp/main/html/rd/p/000000053.000039106.html （掲載終了）

[371] Yahoo！JAPAN ニュース, “「もっと政治家を利用して」福岡市長が“挑戦したい若者”に送る言葉,” 2021 年 6 月 15 日, bizSPA！フレッシュより転載. https://news.yahoo.co.jp/articles/2d018f0459115612e4c29a4b738865153fc940f3 （掲載終了のため転載元の bizSPA！フレッシュを参照 https://bizspa.jp/post-469080/）

[372] 山田佳奈, 中島嘉克, 堀篭俊材, 杉山歩, 江口悟, “コロナ禍、デジタル格差に悩む高齢者　国がスマホ講習会,” 朝日新聞デジタル, 2021 年 6 月 11 日. https://www.asahi.com/articles/ASP6B74N1P6BULFA026.html