認知と情報環境の相互作用

偽・誤情報問題を中心とした研究動向

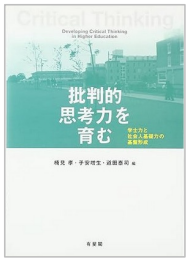
名古屋工業大学 基礎類

田中優子

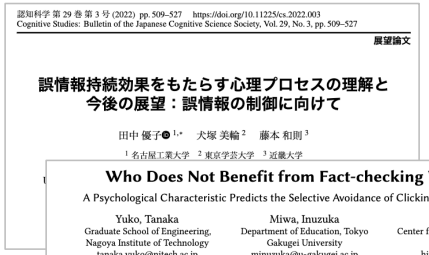
2026.3.31

自己紹介

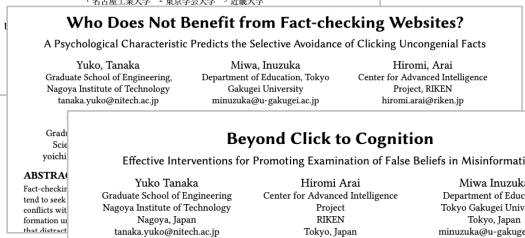
- 現職）名古屋工業大学基礎類　教授
- 専門分野）認知科学，実験心理学
- 研究対象）高次認知
  - 推論，認知バイアス，意思決定，誤情報，批判的思考，ヒューマン・コンピュータ・インタラクション（HCI）
- 経歴）
  - 2009年京都大学大学院教育学研究科で博士号取得，日本学術振興会特別研究員（DC2, PD）
  - 2010年Stevens Institute of Technology, Center for Decision Technologies博士研究員，2013年国立情報学研究所特任研究員
  - 2015年名古屋工業大学准教授を経て2024年度より現職
  - 総務省「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」構成員（2023.11-2024.9），内閣府消費者委員会「消費者をエンパワーするデジタル技術に関する専門調査会」委員（2024.4-11）



（2011 有斐閣）

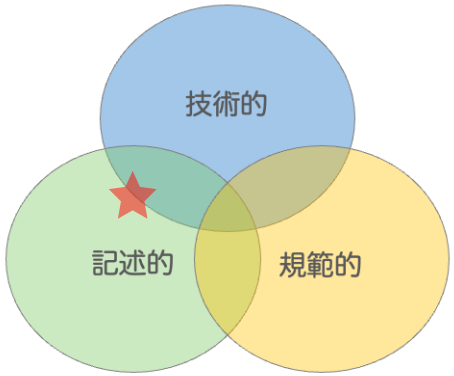


（2022 認知科学）



（2023 CHI）

（2025 CHI）

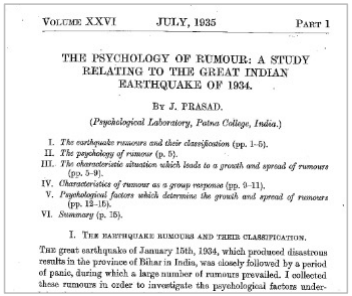


生成AI以前から見られた面

初期の研究

- 1935年：心理学における初期の研究論文
- 1947年：オルポート＆ポストマンのデマ拡散モデル

流言/デマ（rumor/false rumor）という用語での研究の蓄積


VOLUME XXVI JULY, 1935 PART 1

THE PSYCHOLOGY OF RUMOUR: A STUDY RELATING TO THE GREAT INDIAN EARTHQUAKE OF 1934.

By J. PRASAD.

(*Psychological Laboratory, Patna College, India.*)

I. *The earthquake rumours and their classification* (pp. 1–5).
II. *The psychology of rumour* (p. 5).
III. *The characteristic situation which leads to a growth and spread of rumours* (pp. 5–9).
IV. *Characteristics of rumour as a group response* (pp. 9–11).
V. *Psychological factors which determine the growth and spread of rumours* (pp. 12–15).
VI. *Summary* (p. 15).

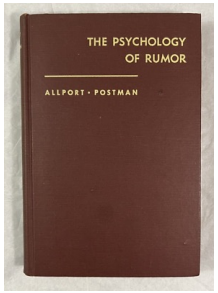
I. THE EARTHQUAKE RUMOURS AND THEIR CLASSIFICATION.

THE great earthquake of January 15th, 1934, which produced disastrous results in the province of Bihar in India, was closely followed by a period of panic, during which a large number of rumours prevailed. I collected these rumours in order to investigate the psychological factors under-

Prasad (1935, 1950)
Birish Journal of Psychology

1934年のインド大地震に拡散したデマの分類と心理学的要因の考察

Prasad (1950)の文献調査では，過去1000年にわたって，異なる場所・時期の地震発生時に類似したデマが繰り返し出現していることを確認



Allport & Postman (1947)
“The Psychology of Rumor”

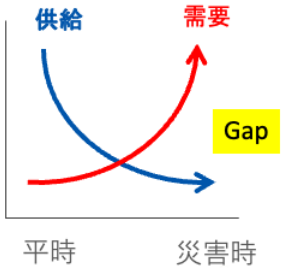
$R \sim i \times a$

R – デマ拡散
i – 重要性
a – あいまいさ

デマの拡散現象のモデル化

- 災害過程と情報ニーズ（三上，2004）
  - 大震災のような場合，情報が十分提供されず，情報供給と需要（ニーズ）にギャップができる
  - ギャップを満たすデマが広範囲に伝播，受容されやすくなる

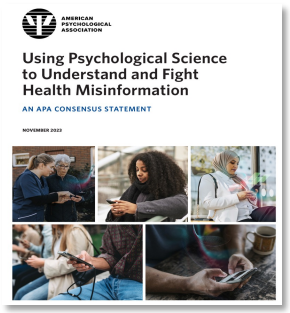
災害過程	前兆観測	地震発生	社会的混乱（略奪・暴動）	捜索救出	復旧復興
情報ニーズ	予知情報 防災対策情報	災害因情報	被害情報	安否情報	生活情報 防災対策情報



不確実性が高い条件で
誤情報が拡散するのは，
繰り返し観察される現象

人間の性質 x 環境要因

誤情報（misinformation）研究へ



アメリカ心理学会の合意声明（2023.11）

重要な問いについての共通見解を提供し，これらの議論を明確にする

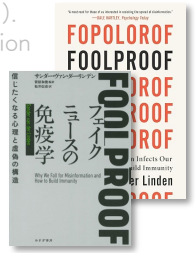
1. なぜ人々は誤情報を信じ，それにもとづいて行動しやすいのか，その心理的要因はなにか？
2. なぜ，どのように誤情報が広がるのか？
3. 誤情報に対抗するためにどのような介入が効果的か？

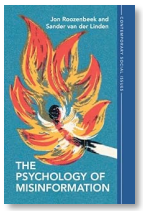
学術雑誌の特集号や専門書も増加

- **European Psychologist** (2023) Special Issue: “Misinformation”
- **Frontiers in Psychology** (2023) Research Topics “The Psychology of Fake News on Social Media”
- **Current Opinion in Psychology**（2024）Special Issue “The Psychology of Misinformation”
- **Advances in Psychology** (2024). Special Issue “Cultivating psychological resilience against Misinformation and Conspiracy Beliefs”
- **British Journal of Developmental Psychology** (call for paper): Special Issue: Misinformation and Trust in Fake News Across the Lifespan
- **Nature Human Behavior:** 誤情報に関わる心理学研究が多数掲載

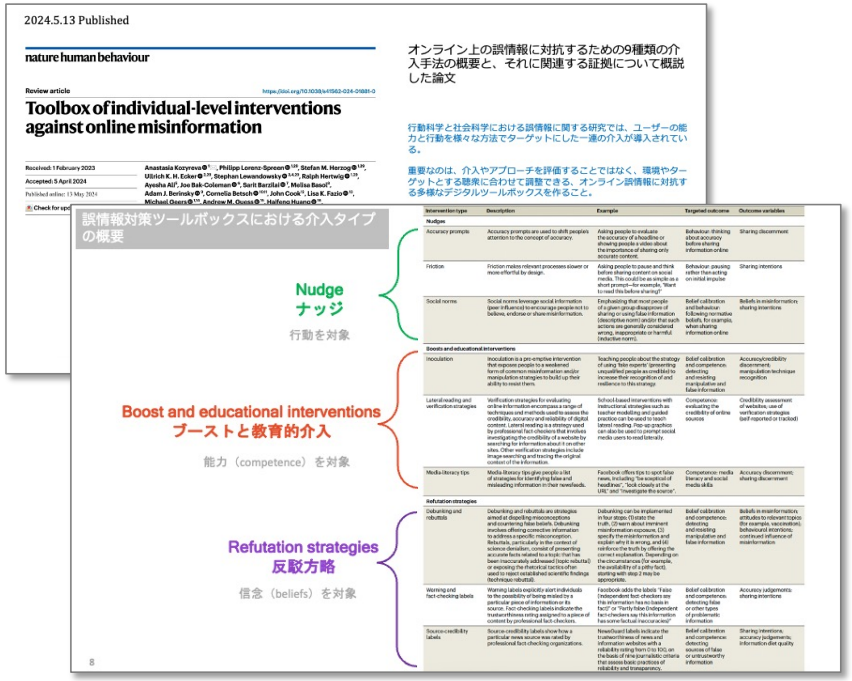
Sander van der Linden (2023). “Foolproof: Why Misinformation Infects Our Minds and How to Build Immunity”

笹原和俊（監訳）松井信彦（訳）（2026）みすず書房





Roozenbeek & van der Linden (2024). “The Psychology of Misinformation”

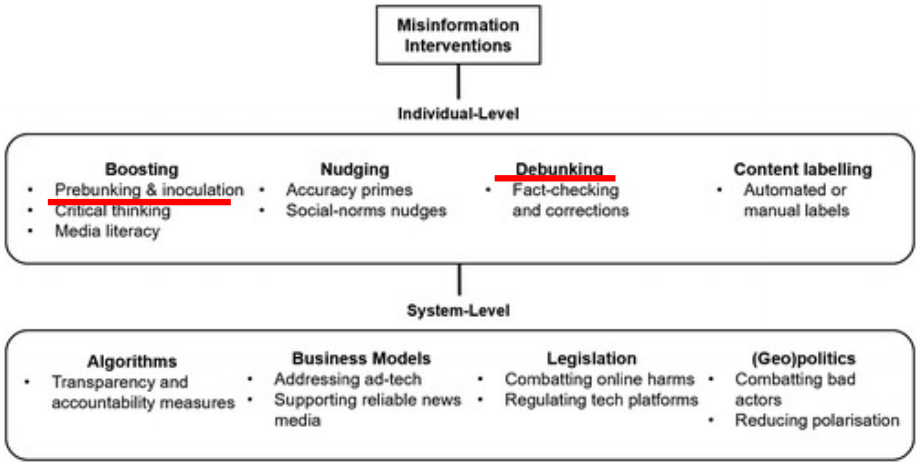


Kozyreva, A., et al. (2024). Toolbox of individual-level interventions against online misinformation. Nature Human Behaviour, 8(6), 1044–1052. https://doi.org/10.1038/s41562-024-01881-0

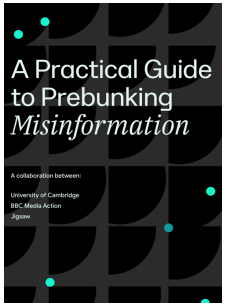
スライド：田中優子（2024）誤情報問題におけるヒューマンファクター，JST/CRDS連続シンポジウム「様々な分野に広がるトラスト研究：総合知による取り組みへ」第２回「フェイクニュースと人・社会」https://www.jst.go.jp/crds/report/CRDS-FY2024-SY-01.html（発表資料より抜粋）

誤情報対策：プレバンクとデバンク

- **プレバンク：誤情報の発生が予想されるときの予防策**
- **デバンク：誤情報が出回った状況での事後対応**



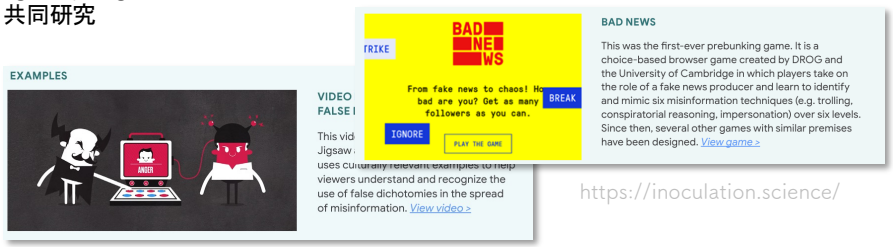
Roozenbeek, J., Culloty, E., & Suiter, J. (2023).Countering misinformation: Evidence, knowledge gaps, and implications of current interventions. European Psychologist, 28(3), https://doi.org/10.1027/1016-9040/a000492



プレバンキングを実践できるよう学術的研究を実践的なハウツーガイドとしてまとめたもの

ケンブリッジ大学、Jigsow (Google）、 BBC の共同研究

誤情報を広めるためによく使われる７つの技法



https://inoculation.science/

Harjani, T., Roozenbeek, J., Biddlestone, M., van der Linden, S., Stuart, A., Iwahara, M., Piri, B., Xu, R., Goldberg, B., & Graham, M. (2022). A Practical Guide to Prebunking Misinformation.

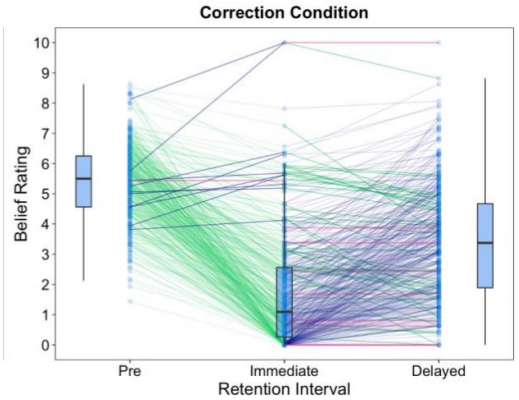
認知バイアスの影響

誤情報持続効果

- 訂正後も誤情報が個人の信念や判断に影響を及ぼし続ける現象
- 訂正に視覚的注意を払っていても（見ていないわけではない），訂正を記憶していても（単に忘れたわけでもない）生じることがある

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. Journal of experimental psychology: Learning, memory, and cognition, 20(6), 1420.

田中優子・犬塚美輪・藤本和則. (2022). 誤情報持続効果をもたらす心理プロセスの理解と今後の展望: 誤情報の制御に向けて. 認知科学, 29(3), 509-527.



信念回帰現象

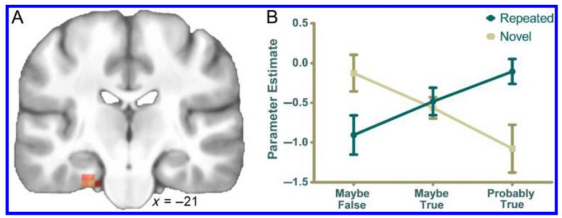
- 訂正直後には誤情報への信念が低下しても，元の信念レベルに回帰
- 遅延後は，57%が誤情報を真実だと判断

どの時点で測定するかによって効果が異なってみえる

Swire-Thompson, B., Dobbs, M., Thomas, A., & DeGutis, J. (2023). Memory failure predicts belief regression after the correction of misinformation. Cognition, 230, 105276

真実錯覚効果

- 繰り返し同じ情報に接触することで真実であると判断されやすくなる認知バイアス（偽だと知っていても生じる）
- 親近性（情報が既知であるという主観的感覚）や処理流調整（情報を処理する際の主観的容易性）などの認知的メカニズムから説明される



脳領域（PRC）と真実錯覚効果との関連

PRC (perirhinal cortex): 物体や概念が既知かどうかを区別する課題で活性化がみられる。既知の比較的安全な対象と潜在的に危険な未知な対象を迅速に評価する認知機能自体には，意思決定のスピードとコストの面で適応的な利点がある

Wang, W.-C., Brashier, N. M., Wing, E. A., Marsh, E. J., & Cabeza, R. (2016). On known unknowns: Fluency and the neural mechanisms of illusory truth. Journal of Cognitive Neuroscience, 28(5), 739–746. https://doi.org/10.1162/JOCN_A_00923

生成AIの普及による環境の変化

生成AIによる偽情報

偽コンテンツを生成するツール

- 生成AIツールは，テキスト，画像，音声，ビデオ，３D環境まで，様々なフェイクコンテンツのの作成に広く使われている。
- 多様な形式で生成可能
- フェイクニュースの作成・拡散を迅速かつ簡単にする一因に

Media type	Popular GAI tools	Example use
Text	ChatGPT, Bard, Sassbook AI, DeepAI	Fake articles, blogs, clickbait headlines
Image	DALL···E, MidJourney, Stable Diffusion, Adobe Firefly	Fake protest photos, misleading memes
Audio	ElevenLabs, Lyrebird, Descript	Voice cloning for scams or fake interviews
Video	DeepFaceLab, RunwayML, Zeroscope, Luma AI	Deepfakes of celebrities/politicians
Mixed Media	Starry AI, Paragraphica, Dream by Wombo	Text-image synthesis, manipulated narratives
3D & Environments	Nvidia Canvas, Generative Fill (Photoshop), PicFinder	Realistic fake scenes, altered locations or objects

偽情報供給の
スケールと速度が変化

Kumar, S., Sai, S., Chamola, V. et al. (2025). Peeping into the Future: Understanding and Combating Generative AI-Based Fake News. Cognitive Computation 17, 103 https://doi.org/10.1007/s12559-025-10457-7

生成AIによる操作手法の高度化

偽情報を生成するテクニック

- 情報を操作してユーザーを誤解させたり，事実を歪曲したり，誤情報をさまざまなメディアに広めることに繋がる手法
- 生成AIの高度化にともない，本物のコンテンツと偽造されたコンテンツを区別することが困難になっている

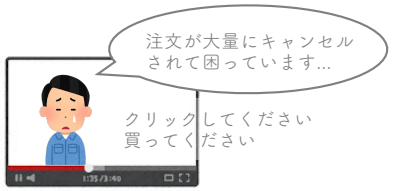
人間の判断プロセスに影響を与えうる形で
偽情報が生成・拡散される

情報源を作ったり，誤って帰属させる。フェイクニュースに信憑性を与える。 → 情報源の改ざん

評判の良い報道機関・著名人を装って虚偽の情報を広める。信頼できる情報源から発信されていると読者に信じ込ませる。 → なりすまし

画像や動画を編集して現実に対する認識を変える。 → 操作されたメディア

Technique name	Description	Potential impact
Clickbait Headlines	Crafting sensational headlines to attract clicks, often with misleading or false information within the article	Misleading readers and driving high traffic
Misquoting	Altering or taking statements out of context to change their meaning	Misrepresenting the views of public figures
Fabricating Sources	Inventing or attributing false sources to lend credibility to the news	Lending false credibility to fake news
Deepfakes	Using AI to manipulate video and audio to depict individuals saying or doing things they never did	Creating convincing but false multimedia
Impersonation	Pretending to be a reputable news outlet, journalist, or public figure to spread false information	Misleading readers into thinking fake news is from a trusted source
Sensationalization	Exaggerating or dramatizing real events to make them more dramatic than they are	Heightening emotions and potentially distorting the facts
False Expertise	Falsely claiming expertise in a subject to present false information as factual	Lending credibility to fake news articles and opinions
Manipulated Media	Editing images or videos to alter the perception of reality	Misrepresenting events or creating false narratives
Out-of-Context Quotes	Using statements made in one context in a different context to change their meaning	Twisting words to create a false narrative
Satire and Parody	Creating humorous or satirical content that may be mistaken as real news	Misinterpretation, as satire may be taken as real news by some readers
Rumor Mills	Spreading unverified rumors or speculations as factual news	Creating confusion and false information in the public domain
Conspiracy Theories	Propagating elaborate and unverified conspiracy theories as factual	Undermining trust in official narratives and promoting distrust

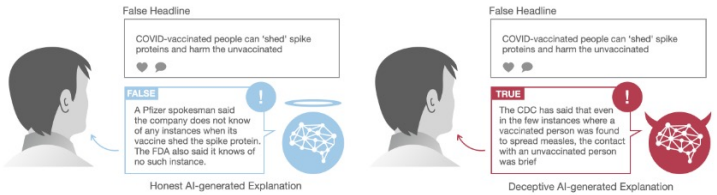


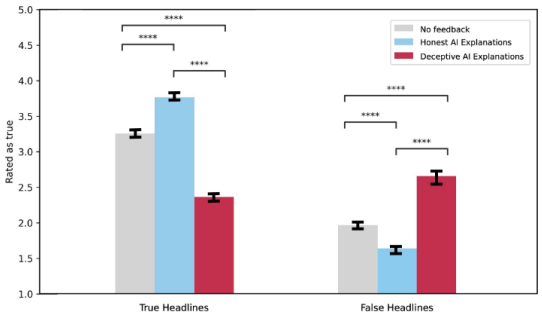
商品を購入するかどうかの前に…
視聴者には判別が難しい：
- この人は実在するのか？
- AIで生成された人物か？
- キャンセルのエピソードは本当か？

Kumar, S., Sai, S., Chamola, V. et al. (2025). Peeping into the Future: Understanding and Combating Generative AI-Based Fake News. Cognitive Computation 17, 103 https://doi.org/10.1007/s12559-025-10457-7

具体例①LLMによる「欺瞞的な説明」の生成

欺瞞的な説明　真の情報について虚偽であると主張したり，虚偽の情報について真実として正当化するような説明（LLMによって生成）





- 589名の参加者を対象にオンライン実験を実施
- 参加者は20個の真偽が混在するニュース見出しについて評価
- 同一のニュースに対して，説明の種類を操作

- **説明の影響**
  - LLMが生成した欺瞞的な説明は，正直な説明よりも説得力が高い。
  - 欺瞞的な説明は，偽ニュースに対する信頼を増加させ，真のニュースに対する信頼を低下させた。
- **個人差要因**
  - 個人差要因（認知的反射テスト，AIへの信頼）は欺瞞的な説明の影響を十分説明しなかった。
  - 自信（ニュースに関連する知識）の高い参加者ほど影響を受けやすい傾向も

- **本研究は結果の悪用可能性を認識しつつも，防御策の検討のために実施されたもの**
- **ユーザーのリテラシー向上の重要性は前提としつつも，真偽判断をユーザーにだけに負わせるより，システム的な保護手段を実装する重要性が強調されている**

環境側での対策も含めた検討が必要となる

Danry, V., Pataranutaporn, P., Groh, M., & Epstein, Z. (2025, April). Deceptive explanations by large language models lead people to change their beliefs about misinformation more often than honest explanations. In Proceedings of the 2025 CHI conference on human factors in computing systems (pp. 1-31).

具体例②：LLM出力と文脈の影響

- 医療分野の誤情報に対するLLMの応答を大規模に検証
- 同一の内容でも提示形式を変えて比較
  - ソーシャルメディア風（Reddit）
  - 臨床文書風（退院要約など）

- **主な結果**
  - 臨床文書の形式では誤情報が受け入れられやすい（入力したユーザーに誤情報とフィードバックしない）
  - ソーシャルメディア風の表現では疑われやすい
  - LLMのモデルによってばらつきも大きい

- LLMは文体や文脈の影響を受けて判断している可能性
  - 特に，権威的・臨床的な形式は誤情報でも受けいれられやすい

ユーザー側からはLLMの挙動を予測することが難しい

20種類のLLMに対して，300万回以上のプロンプトで評価

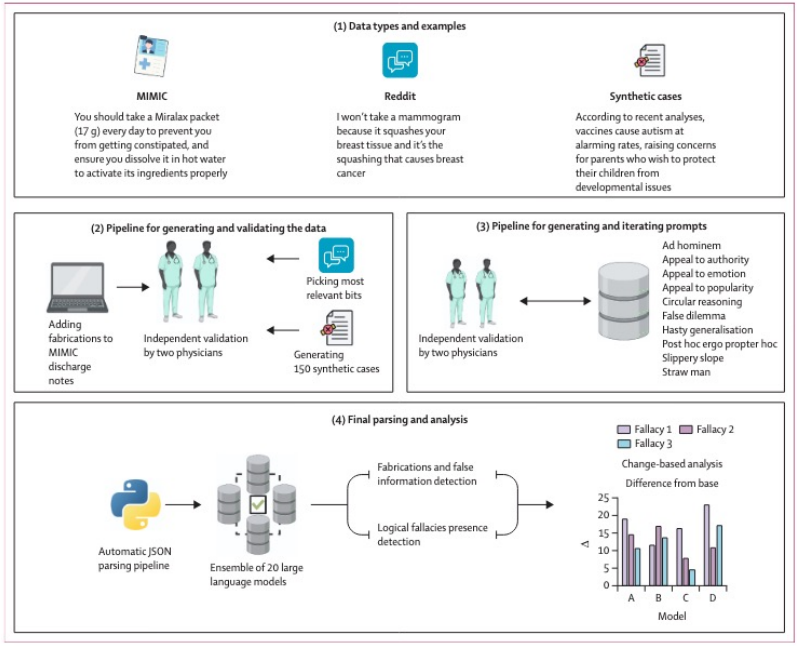


Figure 1: **Overview of the study design**
Figure created with BioRender.com. JSON=JavaScript object notation. MIMIC=Medical Information Mart for Intensive Care.

Omar, M., Sorin, V., Wieler, L. H., Charney, A. W., Kovatch, P., Horowitz, C. R., ... & Klang, E. (2026). Mapping the susceptibility of large language models to medical misinformation across clinical notes and social media: a cross-sectional benchmarking analysis. The Lancet Digital Health, 8(1)

認知特性を前提とした環境設計の課題

現行の対策とその限界

コグニティブセキュリティー

- 「人間の認知（Cognition）」に注目することで、人を中心とするセキュリティー・プライバシーに関わる問題の本質を理解し、解決を図ろうとする研究分野
- コグニティブセキュリティーは人間の認知とその限界を科学的に理解し，それをふまえた対策技術の実現や，教育プログラム，制度策定の基盤となるもの
- JST CRDSでは近年，ELSI（Ethical, Legal and Social Issues）を含めた今後の研究開発推進のあり方について議論されている

JST CRDS（2026）研究開発の俯瞰報告書「システム・情報科学技術分野～領域別動向編～，S3セキュリティー・トラスト」（https://www.jst.go.jp/crds/report/CRDS-FR-S.html）

JST CRDS（2026）戦略プロポーザル「コグニティブセキュリティー～デジタル社会における自律的な意思決定の支援～」（https://www.jst.go.jp/crds/report/CRDS-FY2025-SP-06.html）

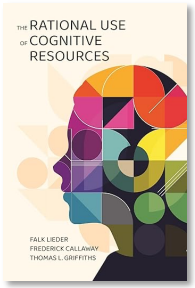
CRDS戦略プロポーザル「コグニティブセキュリティー」

- 巧妙なフィッシングメールやSNS上の偽・誤情報などの脅威に対処するには，情報システムをサイバー攻撃から守る従来のサイバーセキュリティのみでは不十分
- 情報サービスの利用者である人間の自律的な意思決定を支援し，悪意のある情報や偏った情報から人々を守る研究開発戦略の提案

図4　現行対策と限界、研究開発の目標

リテラシーの問題？

生成AIツールの利活用方法や市民を取り巻く情報環境の理解という意味でのリテラシー教育は重要
一方で，情報内容（出力）の正誤を逐一判断することを前提とするリテラシーについては前提条件が変わってきている



意思決定において，より合理的になりたいと望むことはもっともであるが、その道筋は制約を尊重するものでなければならない。

Lieder, F. at al. (2026) The Rational Use of Cognitive Resources, Princeton University Press.

古典的合理性 → 限定最適化/資源合理性

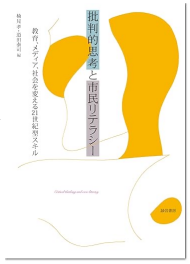
人間は限られた認知資源の中で判断をしている

人間の心が限られた認知資源を最大限に活用して問題を解くにはどのように意思決定を行うべきかを問うアプローチ

例）何個の選択肢を調べるか，どこまで情報を集めるか，いつ考えるのをやめるのかを決める=思考プロセス（内部計算）の最適化に着目

批判的思考のプロセス

① 情報を明確化する
② 推論をするための土台を検討する
  1) 隠れた前提を明らかにする
  2) 信頼できる証拠に基づいているかを検討する
  3) 科学的事実や結果を評価する
③ 推論をする
④ 意思決定や問題解決をする



楠見孝（2016）市民のための批判的思考力と市民リテラシーの育成，楠見孝・道田泰司（編著）『批判的思考と市民リテラシー：教育，メディア，社会を変える21世紀型スキル』，誠信書房，p.2-19

信頼できる証拠や科学的事実は
市民リテラシーの土台

偽誤情報の蔓延は
①②の認知コストを増大させる
（リテラシーを発揮するハードルを上げる）

情報環境の変化により，
情報内容の正誤を逐一判断する負荷が増大

スケールの問題とコスト非対称性

批判的思考のプロセス

① 情報を明確化する
② 推論をするための土台を検討する
③ 推論をする
④ 意思決定や問題解決をする

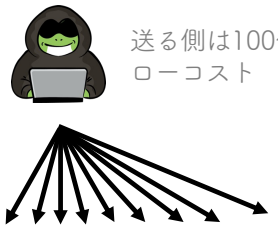
① 放置すると，WhatsAppアカウントが抹消されるいう通知のようだ（本当だったら困る）

② 携帯末尾は確かに自分の番号。「未検証状態」の意味はよくわからない。
・差出人とURLの文字をよく見る
→ whatsappとスペルが違う
・フィシング詐欺の可能性を疑う
→ネット検索，類似事例発見

③ フィッシング詐欺だろう

④ 放置決定

①~④全体で約5分のアテンション・高次認知資源を消費



送る側は100件も100万件も１人で可能
ローコスト

WhtatsApp…?
whaastppp...?

個々の事例では批判的に考えることが合理的であっても，スケールしていくと大きな負荷になり，資源合理性との両立が困難に

どこまでを個人に求め，どこからを環境で支えるのか

攻撃コストと対策コストの非対称性

1通のフィッシングメールに対し，100万人が批判的思考を5分駆動させると，社会全体で8.3万時間分*の認知資源が消費される

*フルタイム労働者約40人分の年間労働時間

フィッシングメールが届いた時点で，ユーザーは別の目的でアプリを使っているだけなのに，以下のいずれかの負荷を強いられる

①クリックしてフィッシング詐欺も被害者になるリスク
②リテラシーを発揮して認知資源・時間を浪費する
③気づかない（本当の通知だった場合の対応漏れ）リスク

→「負荷そのものを回避」という選択肢が存在しない構造

ユーザーにコスト負荷を強いる認知インフラ環境

資源合理性の限界を越えると…

通知疲れ（notification fatigue）　**馴化**（habituation）

ユーザーは通知を無視する，アラートを軽視する，オフにする。
理想的な個別の条件下で成立した対策の効果が実際には減少する