

情報セキュリティ・セミナー特別企画
「認知の脆弱性から人間をどう守るか～コグニティブ・セキュリティと法的課題の入門～」

人の認知的性質からみたセキュリティの課題

名古屋工業大学大学院工学研究科

田中優子

2024.7.11

自己紹介

- ・ 氏名) 田中優子
- ・ 現職) 名古屋工業大学大学院工学研究科 教授
- ・ 専門分野) 認知科学, 実験心理学
- ・ 研究キーワード) 認知バイアス/誤情報/批判的思考/HCI
- ・ 経歴)
 - ・ 2009年に京都大学大学院教育学研究科で博士号を取得後, 日本学術振興会特別研究員, Chulalongkorn University, Stevens Institute of Technologyでのポスドク, 国立情報学研究所特任研究員, 名古屋工業大学准教授を経て2024年度より現職
 - ・ 2023.11- 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
 - ・ 2024.4- 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

認知科学 第29卷 第3号 (2022) pp.509-527 <https://doi.org/10.11225/cs.2022.003>
Cognitive Studies: Bulletin of the Japanese Cognitive Science Society, Vol. 29, No. 3, pp. 509-527

展望論文

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

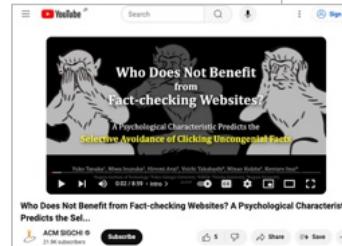
田中 優子^{1,*} 大塚 美輪² 田中 優子, 大塚 美輪, 伊藤 達也

¹名古屋工業大学 ²東京学芸大学

Understanding psychological processes that lead to the persistence of misinformation and future directions toward its control

Yuko Tanaka^{1,*} Miwa Inuzuka² Kazuya Ito

¹Nagoya Institute of Technology ²Tokyo Gakugei University



Who Does Not Benefit from Fact-checking Websites?

A Psychological Characteristic Predicts the Selective Avoidance of Clicking Uncongenial Facts

Yuko, Tanaka

Graduate School of Engineering,
Nagoya Institute of Technology

tanaka.yuko@nitech.ac.jp

Miwa, Inuzuka
Department of Education, Tokyo
Gakugei University
miwazuka@u-gakugei.ac.jp

Hiromi, Ara
Center for Advanced Intelligence
Project, RIKEN
hiromi.arai@riken.jp

Yoichi, Takahashi

Graduate School of Information
Sciences, Tohoku University

yoichi.takahashi.cs@tohoku.ac.jp

Minao, Kukita
Graduate School of Informatics,
Nagoya University

minao.kukita@is.nagoya-u.ac.jp

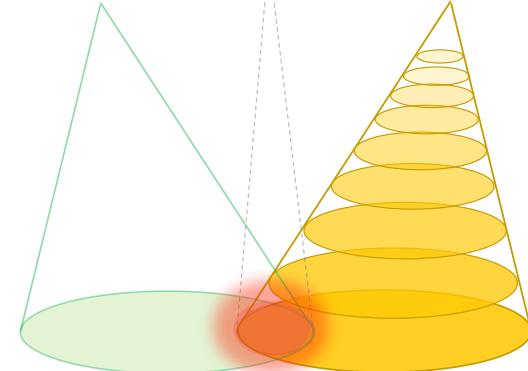
Kentaro, Inui
Graduate School of Information
Sciences, Tohoku University
kentaro.inui@tohoku.ac.jp

1 INTRODUCTION

Misinformation is a significant concern in emergency situations such as pandemics because it can adversely affect human behavior [5, 20, 35, 50]. Misinformation differs from dis-information: Misinformation is defined as false information, which is shared with no intention to harm, whereas dis-information is false information shared to cause harm [73]. People tend to share misinformation be-

社会問題 (誤情報)

工学の目的



心理学の目的 (心の解明)

コグニティブセキュリティー

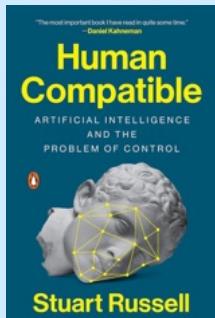
コグニティブセキュリティーの目的

個人や集団に対する悪意のある影響を弱めること

B. M. Pierce (2021). "Protecting people from disinformation requires a cognitive security proving ground" C4ISRNET

Increase **cognitive resilience** against **malicious influence** (悪意のある影響に対する認知的レジリエンスの強化)

サイバーセキュリティーはデバイス・コンピュータ、ネットワークなどの保護に重点があるのに対し、
コグニティブセキュリティーは人間の保護に重点をおく。社会科学・行動科学、AI、データサイエンスなどを含む多くの学問分野を統合する社会技術的アプローチが必要となる。



(Russell, 2019)

Physical Security

1948年の世界人権宣言（第3条）「すべての人は、生命、自由及び身体の安全に対する権利を有する」

Mental Security

“私は、すべての人が精神的安全の権利、つまり、大部分が真実の情報環境で生活する権利、を持つべきだと提案したい。…わたしたちは誤情報の技術に対して非常に脆弱である。

コグニティブセキュリティー

コグニティブセキュリティーの目的

個人や集団に対する悪意のある影響を弱めること

B. M. Pierce (2021). "Protecting people from disinformation requires a cognitive security proving ground" C4ISRNET

Increase **cognitive resilience** against **malicious influence** (悪意のある影響に対する認知的レジリエンスの強化)

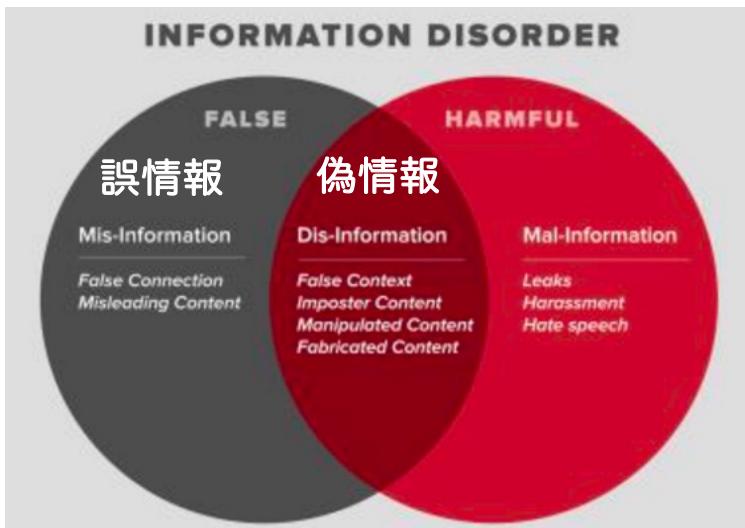
サイバーセキュリティーはデバイス・コンピュータ、ネットワークなどの保護に重点があるのに対し、
コグニティブセキュリティーは人間の保護に**重点**をおく。社会科学・行動科学、AI、データサイエンスなどを含む多くの学問分野を統合する社会技術的アプローチが必要となる。

認知的性質の理解

なぜ人は誤情報を信じるのか?
信じ続けるのか?

認知的性質

デマ・誤情報・偽情報・フェイクニュース



https://jrc.princeton.edu/sites/g/files/toruqf2471/files/van_der_linder_sander_princeton_corr.pdf

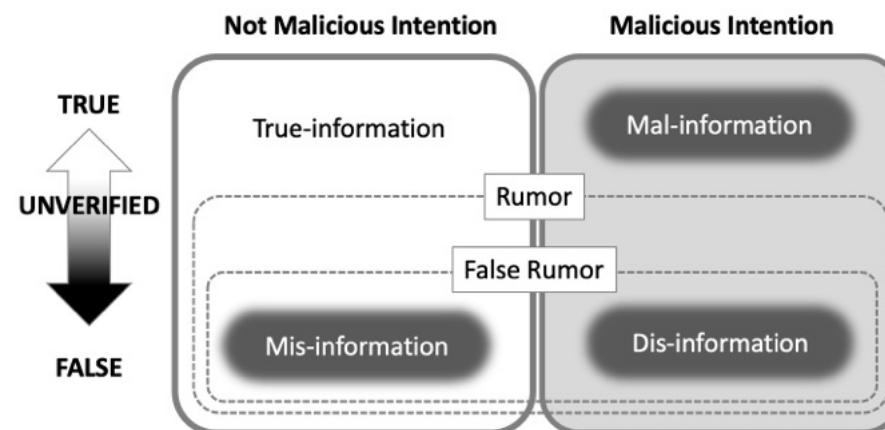
What is “fake news”? (van der Linden, Roozenbeek, Oosterwoud, Compton, & Lewandowsky, 2017)

- **Misinformation**
 - “False or incorrect information” (including human error).
- **Disinformation** (misinformation + intent)
 - “The purposeful spread of false or incorrect information with the explicit intent to cause harm or to confuse and deceive others”.
- **Political Propaganda** (disinformation + political agenda)
 - “Institutionalized or state-run public indoctrination campaigns”.



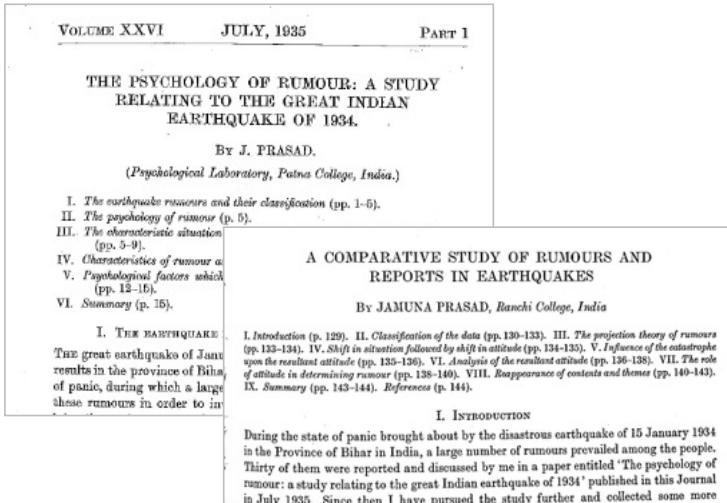
UNIVERSITY OF CAMBRIDGE

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking (Vol. 27, pp. 1-107). Council of Europe.



Tanaka, Y. (2021). Social media technologies and disaster management. In: Sakurai, M., Shaw, R. (eds) *Emerging Technologies for Disaster Resilience. Disaster Risk Reduction*. Springer, 127-143.

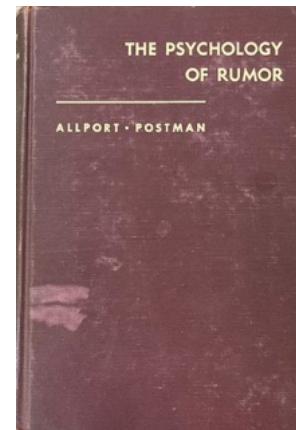
心理学における誤情報研究



Prasad (1935/1950)

1934年にインドで発生した大地震でデマが拡散した事例をもとに、調査・分析。(1934)

過去1000年間にインドで記録された地震の報告から、地震発生時に拡散されるデマの類似点を分析(1950)



Allport & Postman (1947)

$$R \sim i \times a$$

R – rumor transmission
i – important
a – ambiguity

THE BASIC LAW OF RUMOR

A. CHORUS
(University of Leiden)

THE JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY
Vol. 48, No. 2, 1953

According to Allport and Postman (1), the two essential conditions for the transmission of rumor are importance and ambiguity. Roughly, importance stands for the emotional factor and ambiguity for the cognitive factor in rumor-spreading. Both factors are related to rumor transmission in a quantitative manner and a formula for the intensity of rumor may then be written as follows: $R \sim i \times a$. This formula means that the relation between the two factors is not additive but multiplicative, for if either importance or ambiguity is zero, there is no rumor.

Allport and Postman present this law as "highly dependable," but they inform us that there are certain conditions "under which its operation will be weakened" (p. 34). For instance, if heavy penalties are placed on rumor spreading as, for instance, in Gestapo Germany, or if social barriers prevent crossing. Another reason for the failure of the law to operate may be that a person in the rumor chain knows the rumor law. It is a fact "too little observed by psychologists, that knowledge of the operation of a law frequently alters, and sometimes negates, the law in question" (p. 35).

Chorus (1953)

$$R \sim i \times a \times \frac{I}{c}$$

I – supposed general average
c – critical sense

"省察に熟慮を重ね、2つの要因(*i*, *a*)にすべてに流されない傾向"

真実錯覚効果

Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin and Review*, 18(3), 570–578.



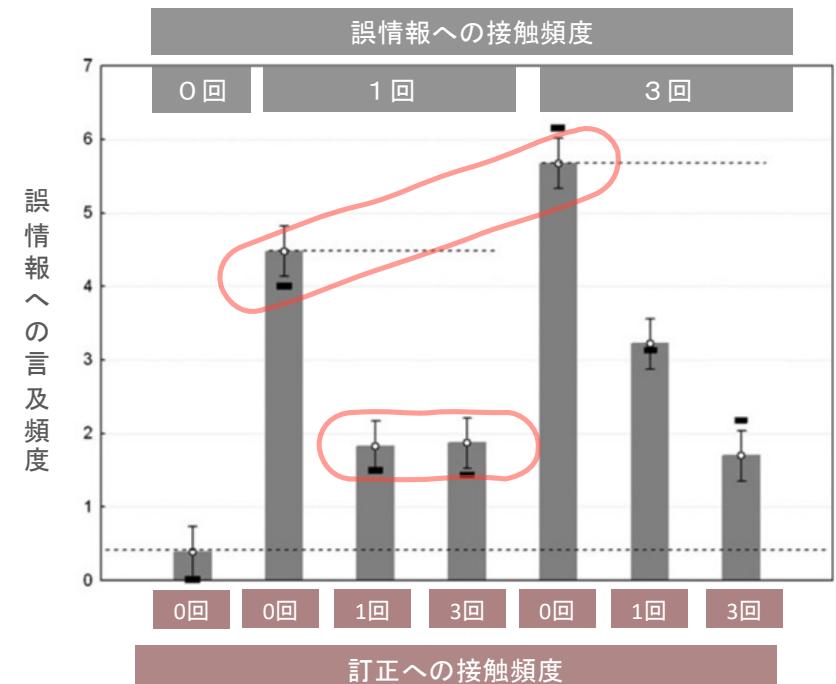
- 繰り返し同じ情報に接触することで、その情報が正しく感じられるようになること。
- 情報への「親近性 (familiarity)」や「処理の流暢性 (fluency)」が「正しさ」のシグナルとして利用されるヒューリスティック

「訂正情報」も繰り返し流せばいいのでは？

誤情報の3倍の頻度で訂正情報を出しても、誤情報の影響は消えない

真実錯覚効果の非対称性

「誤情報の信じられやすさ」と「一度受け入れられた誤情報の影響を事後的に緩和することの難しさ」のギャップ



Quick guide to responding to misinformation



Misinformation can do damage

Misinformation is false information that is spread either by mistake or with intent to mislead. When there is intent to mislead, it is called disinformation. Misinformation has the potential to cause substantial harm to individuals and society. It is therefore important to protect people against being misinformed, either by making them resilient against misinformation before it is encountered or by debunking it after people have been exposed.



Misinformation can be sticky!

誤情報は粘着する

Fact-checking can reduce people's beliefs in false information. However, misinformation often continues ファクトチェックは人々の誤情報への信念を減少させる。ただし、訂正を受け入れた後でも、誤情報はしばしば人々の考えに影響を与えることがある。これは「誤情報持続効果」として知られている。事実による訂正が効果的であるように見えて、人々はしばしば他の文脈で誤情報を利用し続ける。したがって、最大の影響を得るために、最も効果的な訂正のアプローチを使用することが重要。



Prevent misinformation from sticking if you can

Because misinformation is sticky, it's best preempted. This can be achieved by explaining misleading or manipulative argumentation strategies to people—a technique known as “inoculation” that makes people resilient to subsequent manipulation attempts. A potential drawback of inoculation is that it requires advance knowledge of misinformation techniques and is best administered before people are exposed to the misinformation.



Debunk often and properly

If you cannot preempt, you must debunk. For debunking to be effective, it is important to provide detailed refutations^{2,3}. Provide a clear explanation of (1) why it is now clear that the information is false, and (2) what is true instead. When those detailed refutations are provided, misinformation can be “unstuck.” Without detailed refutations, the misinformation may continue to stick around despite correction attempts.

Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarracín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E., Rand, D. G., Rapp, D. N., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M., Swire-Thompson, B., van der Linden, S., Vraga, E., Wood, T. J., Zaragoza, M. S. (2020). The Debunking Handbook 2020. Available at <https://sks.to/db2020>. DOI:10.17910/b7.1182

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. doi.org/10.11225/cs.2022.003

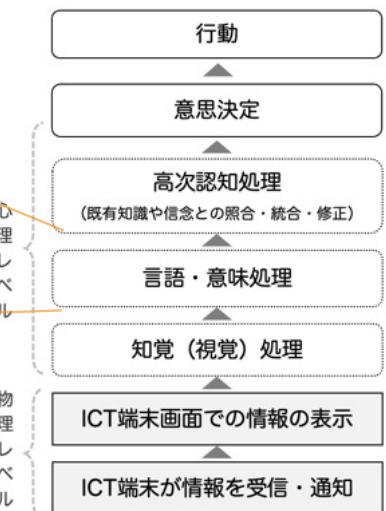
誤情報持続効果

(continued influence effect of misinformation)

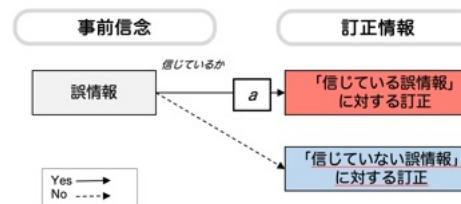
- 誤りであると指摘されていることを知った後も、誤情報を信じ続けたり、誤情報の影響を受け続ける心理現象
- 訂正情報に視覚的注意を払っていても、訂正情報の内容を記憶（記録）していても生じる。
- 高次認知処理レベルの観点から研究が進められている。

訂正情報を記憶（記録）していても生じる (Johnson & Seifert, 1994)

視覚的注意を払っていても生じる (Tanaka, Inuzuka, Hirayama, 2019)



ファクトチェック記事を読まない (クリックを避ける)人々



<i>a</i>	正しいと信じている誤情報の数
<i>b</i>	クリック総数
<i>n</i>	表示されているリンク総数 (<i>b</i> の最大値)
<i>x</i>	正しいと信じている誤情報 (<i>a</i>) のうちクリックされた数
EV	ランダムに <i>b</i> 回クリックした場合、偶然含まれる事前信念と合致しないファクトリンクの数の期待値

Fact Avoidance/Exposure Index (FAEI)

$$FAEI = x - EV$$

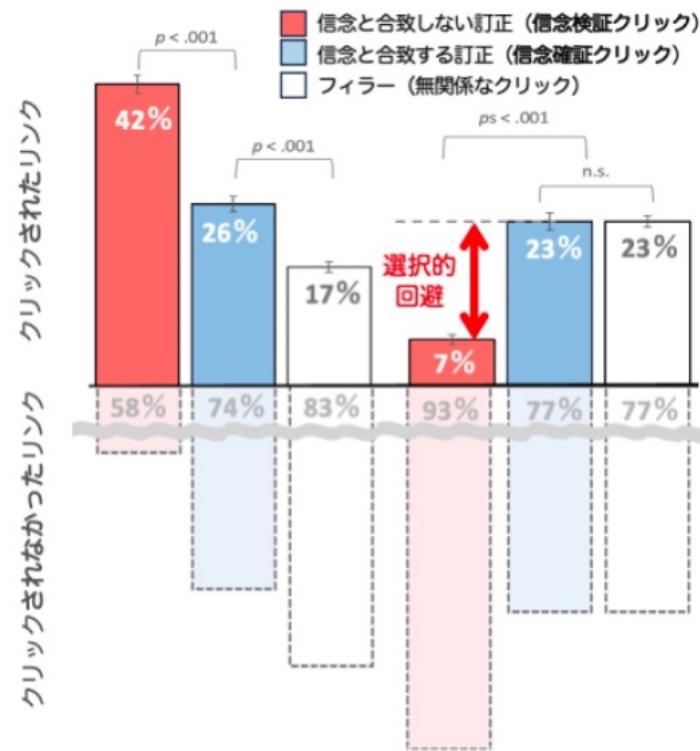
$$EV = \sum_{i=0}^k \frac{aC_i}{bC_b} \times \frac{(n-a)C_{(b-i)}}{nC_b} \times i$$

Yuko Tanaka, Miwa Inuzuka, Hiromi Arai, Yoichi Takahashi, Minao Kukita, and Kentaro Inui. 2023. Who Does Not Benefit from Fact-checking Websites? A Psychological Characteristic Predicts the Selective Avoidance of Clicking Uncongenial Facts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). 1-17.

訂正情報へのアクセスの仕方には個人差がある

ファクト回避群：誤情報を信じている場合、訂正のクリックは7%

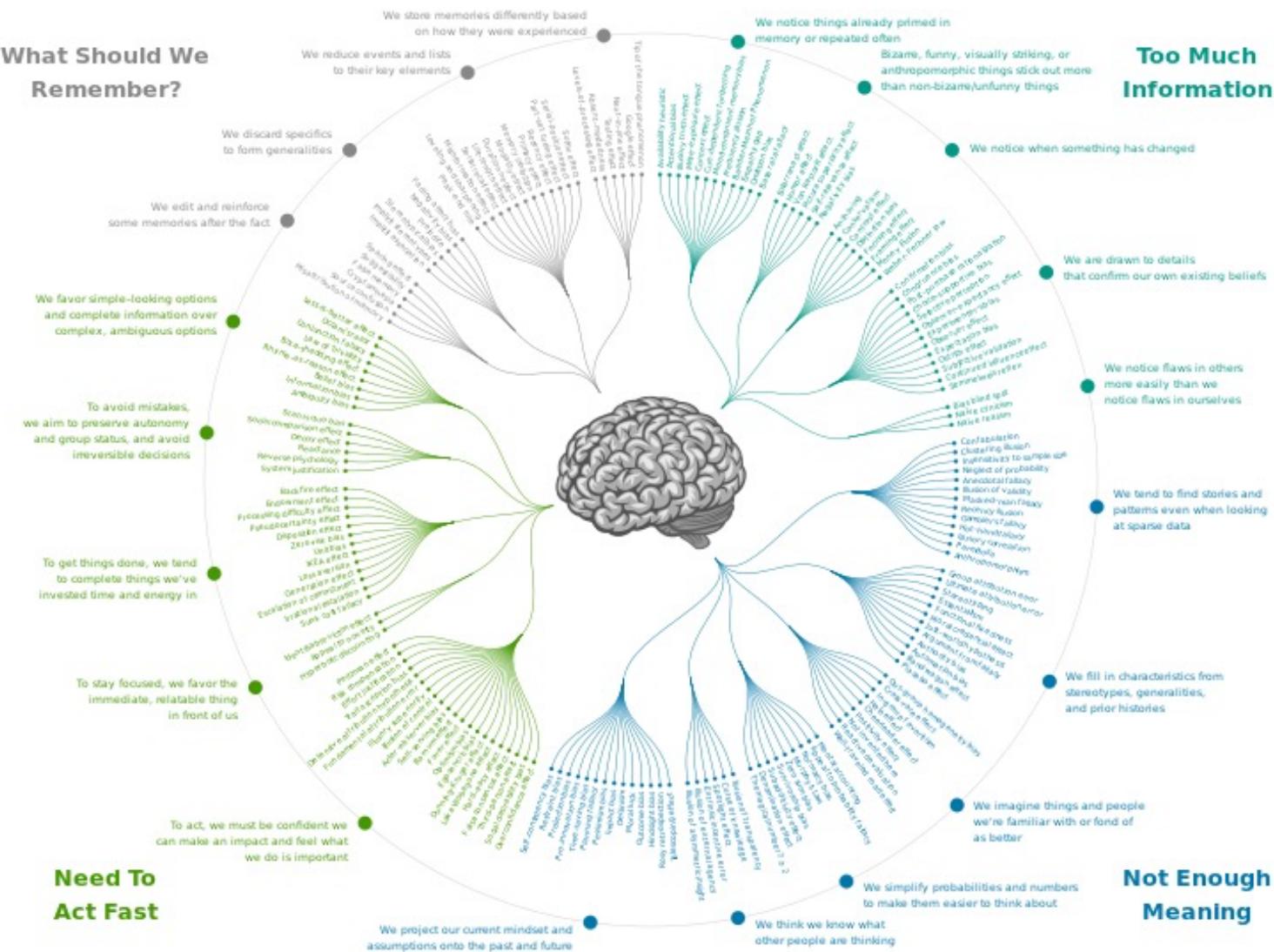
ネット上に訂正を出す ≠ 訂正を届ける

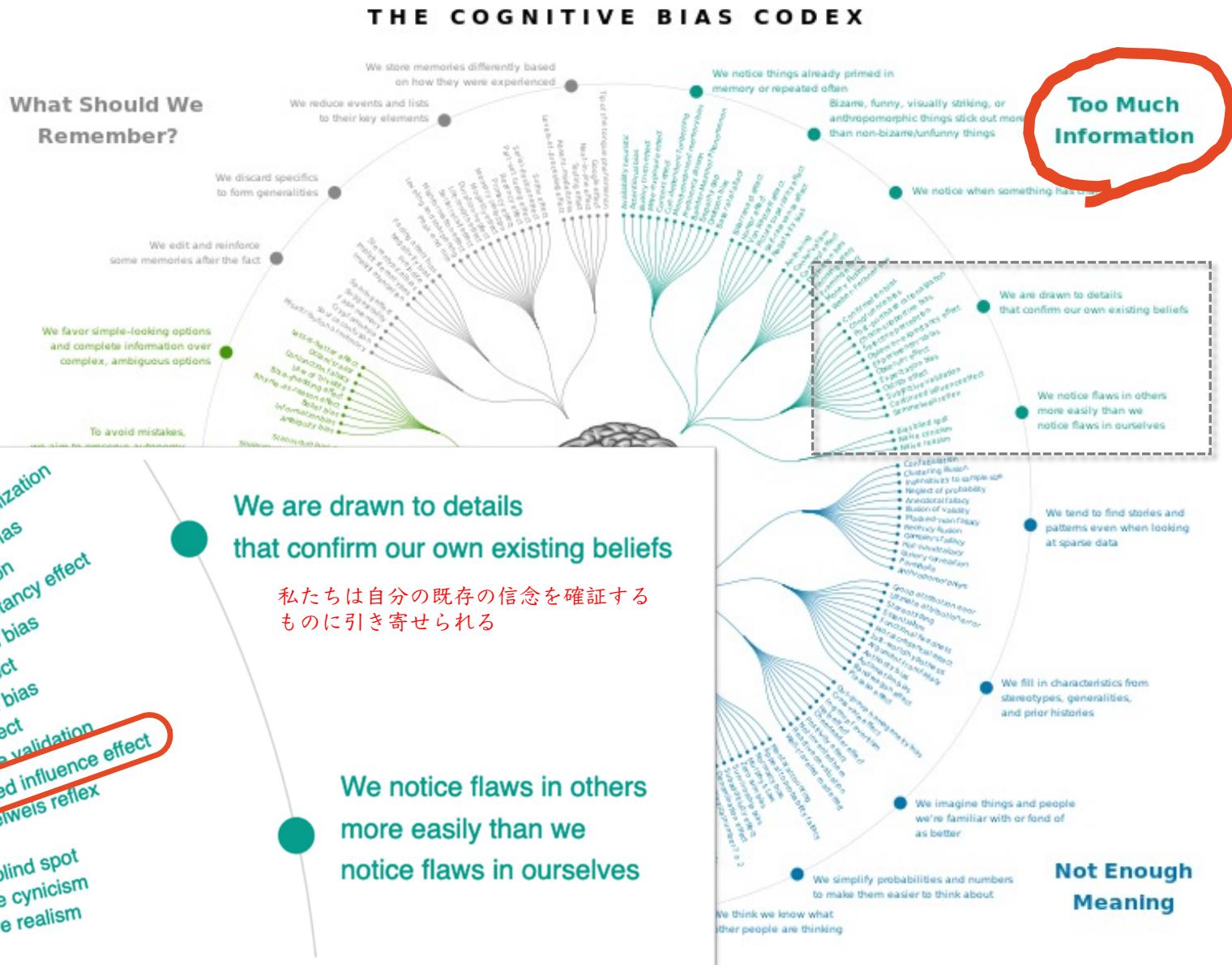


「この情報は誤りです」



THE COGNITIVE BIAS CODEX





認知的性質にもとづく対策

アメリカ心理学会の声明 2023.11

- 目的) 3つの重要な問い合わせ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
 - 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？
 - なぜ、どのように誤情報が広がるのか？
 - 誤情報に対抗するためにどのような介入が効果的か？
- 誤情報による脅威に対処するための8つの推奨事項

Recommendations

RECOMMENDATION 1 Avoid repeating misinformation without including a correction.

The repetition of false claims increases belief in those claims, a phenomenon known as the illusory truth effect. People of all ages are susceptible to illusory truth, even when they already have relevant prior knowledge about the topic. When media sources, political elites, or celebrities repeat misinformation, their influence and repetition can perpetuate false beliefs. Repeating misinformation is necessary only when actively correcting a falsehood. In these cases, the falsehood should be repeated briefly, with the correction featured more prominently than the falsehood itself.

RECOMMENDATION 2 Collaborate with social media companies to understand and reduce the spread of harmful misinformation.

Most misinformation on social media is shared by very few users, even during public health emergencies. These "super-spreaders" can play an outsized role in distributing misinformation. Social media "echo chambers" bind and isolate communities with similar beliefs, which aids the spread of falsehoods and impedes the spread of factual corrections. On social media, sensational, moral-emotional, and derogatory content about the "other side" can spread faster than neutral or positive content. Scientists, policymakers, and public health professionals should work with online platforms to understand and harness the incentive structures of social media to reduce the spread of dangerous misinformation.

RECOMMENDATION 3 Use misinformation correction strategies with tools already proven to promote healthy behaviors.

Psychological science research shows that the link between knowledge and behavior is imperfect. There is strong evidence that curbing misperceptions can change underlying health-related beliefs and attitudes, but it may not be sufficient to change real-world behavior and decision-making. Correcting misinformation with accurate health guidance is vital, but it must happen in concert with evidence-based strategies that promote healthy behaviors (e.g., counseling, skills training, incentives, social norms).

RECOMMENDATION 4 Leverage trusted sources to counter misinformation and provide accurate health information.

People believe and spread misinformation for many reasons: They may find it consistent with their social or political identity, they may fail to consider its accuracy, or they may find it entertaining or rewarding. These motivations are complex and often interrelated. Attempts to correct misinformation and reduce its spread are most successful when the information comes from trusted sources and representatives, including religious, political, and community leaders.

RECOMMENDATION 5 Debunk misinformation often and repeatedly using evidence-based methods.

Research shows that debunking misinformation is generally effective across ages and cultures. However, debunking doesn't always eliminate misperceptions completely. Corrections should feature prominently with the misinformation so that accurate information is properly stored and retrieved from memory. Debunking is most effective when it comes from trusted sources, provides sufficient detail about why the claim is false, and offers guidance on what is true instead. Because the effectiveness of debunking fades over time, it should be repeated through trusted channels and evidence-based methods.

RECOMMENDATION 6 Prebunk misinformation to inoculate susceptible audiences by building skills and resilience from an early age.

Instead of correcting misinformation after the fact, "prebunking" should be the first line of defense to build public resilience to misinformation in advance. Studies show that psychological inoculation interventions can help people identify individual examples of misinformation or the overarching techniques commonly used in misinformation campaigns. Prebunking can be scaled to reach millions on social media with short videos or messages, or it can be administered in the form of interactive tools involving games or quizzes. However, the effects of prebunking fade over time; regular "boosters" may be necessary to maintain resilience to misinformation, along with media and digital literacy training.

RECOMMENDATION 7 Demand data access and transparency from social media companies for scientific research on misinformation.

Efforts to quantify and understand misinformation on social media are hampered by lack of access to user data from social media companies. Misinformation interventions are rarely tested in real-world settings due to a similar lack of industry cooperation. Publicly available data offer a limited snapshot of exposure, but they cannot explain population and network effects. Researchers need access to the full inventory of social media posts across platforms, along with data revealing how algorithms shape what individual users see. Responsible data sharing could use frameworks currently in use to manage sensitive medical data. Policymakers and health authorities should encourage research partnerships and demand greater oversight and transparency from social media companies to curb the spread of misinformation.

RECOMMENDATION 8 Fund basic and translational research into the psychology of health misinformation, including effective ways to counter it.

Several interventions have been developed to counter health misinformation, but researchers have yet to compare their outcomes, either alone or in combination. There is a need to understand which interventions are effective for specific types of information: What works for one issue may not translate to others. Ideally, these questions would be answered by large-scale trials with representative target audiences in real-world settings. Increased funding opportunities for psychological science research are needed to address these important questions about digital life.



Using Psychological Science to Understand and Fight Health Misinformation

AN APA CONSENSUS STATEMENT

NOVEMBER 2023



<https://www.apa.org/pubs/reports/health-misinformation>

2024.5.13 Published

nature human behaviour

Review article

<https://doi.org/10.1038/s41562-024-01881-0>

Toolbox of individual-level interventions against online misinformation

Received: 1 February 2023

Accepted: 5 April 2024

Published online: 13 May 2024

 Check for updates

Anastasia Kozyreva  ^{1,29}, Philipp Lorenz-Spreen  ^{1,29}, Stefan M. Herzog  ^{1,29}, Ullrich K. H. Ecker  ^{2,29}, Stephan Lewandowsky  ^{3,4,29}, Ralph Hertwig  ^{1,29}, Ayesha Ali  ⁵, Joe Bak-Coleman  ⁶, Sarit Barzilai  ⁷, Melisa Basol  ⁸, Adam J. Berinsky ⁹, Cornelia Betsch ^{10,11}, John Cook ¹², Lisa K. Fazio ¹³, Michael Geers ^{1,14}, Andrew M. Guess ¹⁵, Haifeng Huang ¹⁶, Horacio Larreguy ¹⁷, Rokoens Maertens ¹⁸, Folco Panizza ¹⁹, Gordon Pennycook ^{20,21}, David G. Rand ²², Steve Rathje ²³, Jason Reifler ²⁴, Philipp Schmid ^{10,11,25}, Mark Smith ²⁶, Briony Swire-Thompson ²⁷, Paula Szewach ^{24,28}, Sander van der Linden ⁸ & Sam Wineburg ²⁶

The spread of misinformation through media and social networks threatens many aspects of society, including public health and the state of democracies. One approach to mitigating the effect of misinformation focuses on individual-level interventions, equipping policymakers and the public with essential tools to curb the spread and influence of falsehoods. Here we introduce a toolbox of individual-level interventions for reducing harm from online misinformation. Comprising an up-to-date account of interventions featured in 81 scientific papers from across the globe, the toolbox provides both a conceptual overview of nine main types of interventions, including their target, scope and examples, and a summary of the empirical evidence supporting the interventions. including the

オンライン上の誤情報に対抗するための9種類の介入手法の概要と、それに関連する証拠について概説した論文

行動科学と社会科学における誤情報に関する研究では、ユーザーの能力と行動を様々な方法でターゲットにした一連の介入が導入されている。

重要なのは、介入やアプローチを評価することではなく、環境やターゲットとする聴衆に合わせて調整できる、オンライン誤情報に対抗する多様なデジタルツールボックスを作ること。

世界中で行われた研究からのエビデンスを提供する81の科学論文に基づいて、オンライン誤情報に対抗する行動的・認知的介入のレビュー結果を報告する。

研究者、政策立案者、教育者、一般の人々が介入を組み合わせて、誤情報問題のさまざまな側面に対処するための2つのデータベースを含むインタラクティブなオンラインツールボックス (<https://interventionstoolbox.mpib-berlin.mpg.de>) の形で提供。

このレビューの主な目的は、異なる行動的・認知的結果を対象とする、経験的に検証された認知的・行動的な誤情報対策の集合体を特定すること。

Details	Intervention ↑	References ↑	Experimental setting	Design ↑	Treatment ↑	Paradigm ↑	Outcome variable	Sample size ↑	Sample Country	Demographics ↑	Recruitment ↑	Main findings ↑	Longevity ↑
▶ Expand	Accuracy prompts	Pennycook et al., Nature (2021).	Online; Field	RCT	Accuracy prompt: accuracy rating question.	Headline-discrimination paradigm; Field study	Sharing discernment	7955	United States	Convenience; Quota-matched; Twitter users	MTurk; Lucid; Twitter	Priming accuracy improved sharing discernment between false and true headlines in online experiments. In the field experiment, an accuracy message increased the average quality of the news sources shared.	Not measured, but study 7 had a test window of 24 hours

Structure

Conceptual toolbox
Nine intervention types defined along ten dimensions

Evidence toolbox
81 scientific papers summarized and defined along ten dimensions plus study details

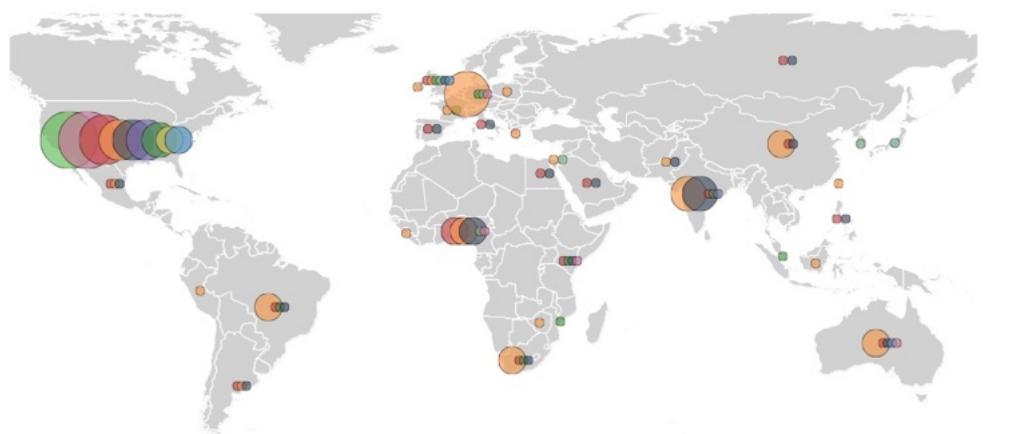
Experts

30 misinformation researchers from 11 countries and 27 universities and research institutions

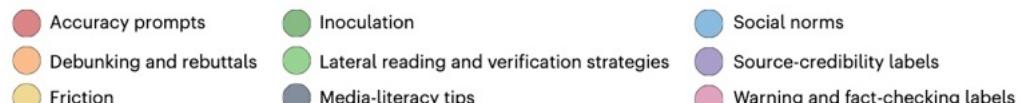
Criteria for inclusion

1. Definition and scope
2. Problem addressed
3. Non-redundancy
4. Evidence
5. Expert opinion

World map of evidence



Interventions



Number of studies



https://interventionstoolbox.mpib-berlin.mpg.de/table_concept.html

誤情報対策ツールボックスにおける介入タイプの概要

Nudge
ナッジ
行動を対象

Boost and educational interventions
ブーストと教育的介入

能力 (competence) を対象

Refutation strategies
反駁方略

信念 (beliefs) を対象

Intervention type	Description	Example	Targeted outcome	Outcome variables
Nudges				
Accuracy prompts	Accuracy prompts are used to shift people's attention to the concept of accuracy.	Asking people to evaluate the accuracy of a headline or showing people a video about the importance of sharing only accurate content.	Behaviour: thinking about accuracy before sharing information online	Sharing discernment
Friction	Friction makes relevant processes slower or more effortful by design.	Asking people to pause and think before sharing content on social media. This could be as simple as a short prompt—for example, 'Want to read this before sharing?'	Behaviour: pausing rather than acting on initial impulse	Sharing intentions
Social norms	Social norms leverage social information (peer influence) to encourage people not to believe, endorse or share misinformation.	Emphasizing that most people of a given group disapprove of sharing or using false information (descriptive norm) and/or that such actions are generally considered wrong, inappropriate or harmful (inductive norm).	Belief calibration and behaviour: following normative beliefs, for example, when sharing information online	Beliefs in misinformation; sharing intentions
Boosts and educational interventions				
Inoculation	Inoculation is a pre-emptive intervention that exposes people to a weakened form of common misinformation and/or manipulation strategies to build up their ability to resist them.	Teaching people about the strategy of using 'fake experts' (presenting unqualified people as credible) to increase their recognition of and resilience to this strategy.	Belief calibration and competence: detecting and resisting manipulative and false information	Accuracy/credibility discernment; manipulation technique recognition
Lateral reading and verification strategies	Verification strategies for evaluating online information encompass a range of techniques and methods used to assess the credibility, accuracy and reliability of digital content. Lateral reading is a strategy used by professional fact-checkers that involves investigating the credibility of a website by searching for information about it on other sites. Other verification strategies include image searching and tracing the original context of the information.	School-based interventions with instructional strategies such as teacher modelling and guided practice can be used to teach lateral reading. Pop-up graphics can also be used to prompt social media users to read laterally.	Competence: evaluating the credibility of online sources	Credibility assessment of websites; use of verification strategies (self-reported or tracked)
Media-literacy tips	Media-literacy tips give people a list of strategies for identifying false and misleading information in their newsfeeds.	Facebook offers tips to spot false news, including "be sceptical of headlines", "look closely at the URL" and "investigate the source".	Competence: media literacy and social media skills	Accuracy discernment; sharing discernment
Refutation strategies				
Debunking and rebuttals	Debunking and rebuttals are strategies aimed at dispelling misconceptions and countering false beliefs. Debunking involves offering corrective information to address a specific misconception. Rebuttals, particularly in the context of science denialism, consist of presenting accurate facts related to a topic that has been inaccurately addressed (topic rebuttal) or exposing the rhetorical tactics often used to reject established scientific findings (technique rebuttal).	Debunking can be implemented in four steps: (1) state the truth, (2) warn about imminent misinformation exposure, (3) specify the misinformation and explain why it is wrong, and (4) reinforce the truth by offering the correct explanation. Depending on the circumstances (for example, the availability of a pithy fact), starting with step 2 may be appropriate.	Belief calibration and competence: detecting and resisting manipulative and false information	Beliefs in misinformation; attitudes to relevant topics (for example, vaccination); behavioural intentions; continued influence of misinformation
Warning and fact-checking labels	Warning labels explicitly alert individuals to the possibility of being misled by a particular piece of information or its source. Fact-checking labels indicate the trustworthiness rating assigned to a piece of content by professional fact-checkers.	Facebook adds the labels "False (Independent fact-checkers say this information has no basis in fact)" or "Partly false (Independent fact-checkers say this information has some factual inaccuracies)"	Belief calibration and competence: detecting false or other types of problematic information	Accuracy judgements; sharing intentions
Source-credibility labels	Source-credibility labels show how a particular news source was rated by professional fact-checking organizations.	NewsGuard labels indicate the trustworthiness of news and information websites with a reliability rating from 0 to 100, on the basis of nine journalistic criteria that assess basic practices of reliability and transparency.	Belief calibration and competence: detecting sources of false or untrustworthy information	Sharing intentions; accuracy judgements; information diet quality

Nudge

Intervention Type	Description	Example	Target Outcomes
Accuracy prompts	人々の注意を正確性に向けさせるために使用されるプロンプト	ヘッドラインや動画についての正確性を評価するよう人々に求める	オンラインで情報を共有する前に正確性について考える
Friction	関連するプロセスを設計によって、より遅く、努力を要するものにする	ソーシャルメディアでコンテンツを共有する前に、一時停止して考えるようユーザーに求める。例えば、「共有する前に、これを読みたいですか?」という短いプロンプトで行うことができる	最初の衝動に基づいて行動するのではなく、一度立ち止まる
Social norms	誤情報を信じたり、支持したり、共有したりしないよう人々に働きかけるために、社会情報（仲間の影響）を活用する	特定のグループのほとんどの人が誤情報の共有や使用を認めていない（記述的規範）、あるいは、そのような行為は一般的に間違っている、不適切である、有害であると考えられている（帰納的規範）ことを強調する	信念のキャリブレーションと行動: 例えば、情報をオンラインで共有する際に、規範的な信念に従う

Accuracy Prompts



To the best of your knowledge, is the above headline accurate?

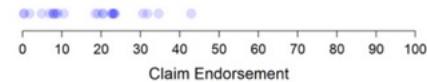
- Yes
 No

Friction



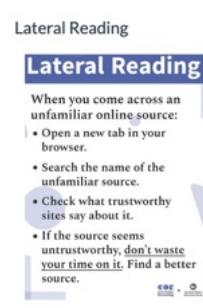
Please explain how you know that the headline is true or false.

Refutations and Social Norms



Boosts and educational interventions

Intervention Type	Description	Example	Target Outcomes
Inoculation	一般的な誤情報や操作の戦略に対する人々の抵抗力を高めるために、弱められた形の誤情報や操作の戦略を用いて、それに抵抗する能力を高めることを目的とした先制的な介入	「偽の専門家を使う」などの戦略について人々に教え、この戦略に対する抵抗力を高める	信念のキャリブレーションと能力: 誤解を招くような誤情報を検出し、抵抗する
Lateral reading and verification strategies	横読みは、プロのファクトチェッカーが使用する戦略で、他のサイトでそのウェブサイトに関する情報を検索することで、ウェブサイトの信頼性を調査することを含む。その他の検証戦略には、画像検索や情報の元の文脈を追跡することなどがある。	ポップアップグラフィックを使用して、生徒に横読みを促す	情報源の信頼性を評価する能力。オンラインソースの信頼性の評価
Media-literacy tips	誤情報や誤解を招く情報を特定するための戦略のリストを人々に提供する	偽のニュースを特定するためのFacebookのヒント。「ヘッドラインが疑わしいか」、「URLを確認する」、「情報源を調査する」など	メディアリテラシーとソーシャルメディアスキル



https://interventionstoolbox.mpib-berlin.mpg.de/table_concept.html
Claude 3 Opusによる仮訳

「能動的」プレバンкиング

- 能動的な接種
 - ゲームまたはクイズの形で提供される
 - 誤情報でよく使われるテクニックに対する抵抗力を高める効果があることが示されている
 - 実証研究のメタレビューの結果、受動的な接種と比べると効果の持続性は高いことが示されている（ブースターが提供されると3ヶ月以上）

所要時間15-20分
教員・生徒向けの
解説や教材あり



BAD NEWS

This was the first-ever prebunking game. It is a choice-based browser game created by DROG and the University of Cambridge in which players take on the role of a fake news producer and learn to identify and mimic six misinformation techniques (e.g. trolling, conspiratorial reasoning, impersonation) over six levels. Since then, several other games with similar premises have been designed. [View game >](#)



HARMONY SQUARE

Set in a peaceful community known for its pond swan and annual Pineapple Pizza Festival, this game appoints the player as the "Chief Disinformation Officer," tasked with polarizing the people of Harmony Square and using trolling campaigns during political elections. [View game >](#)

政治的なプロパガンダに対応したゲーム
2019年CISA「The War on Pineapple」を元にしたシナリオ



GO VIRAL!

This game is a browser-based game that allows players to play an online game to combat COVID-19 misinformation. It features a variety of challenges and puzzles that teach players about the science of COVID-19 and how to identify and combat misinformation. The game is designed to be fun and engaging, while also providing valuable information about the pandemic. It has been used in classrooms and other educational settings to help students learn about COVID-19 and how to stay safe. [View game >](#)

ケンブリッジ大学とWHOの共同開発
(英語, ドイツ語, フランス語)

短い時間で
実施できるよう設計

所要時間：5分

<https://inoculation.science/>から利用可能

誤情報を広めるためによく使われる7テクニック

なりすまし

TECHNIQUE	EXAMPLE
Impersonation Spreading information as another person or organization in order to	"NASA admitted that climate change occurs naturally as a result of changes in Earth's solar orbit and not anthropogenic factors."

感情操作

Emotional manipulation Using language that leverages strong	"What this airline did for its passengers will make you tear up — SO heartwarming."
--	---

二極化

Polarization Exaggerating existing differences	"People's Party: Don't believe the Worker Party liars. They said they would abolish student debt yet more people today are in debt than ever."
---	--

陰謀論的な 考え方

Conspiratorial ideation Explaining events from traditional news using alternative explanations that give weight to the idea that a small set of	"Vaccines are just a way for billionaires to track us with their microchip vaccines! Who's really in control of our bodies here?"
--	---

個人攻撃

Ad hominem attack Ad hominem, Latin for "to the person,"	"Barbara has an uncontrollable temper and apparently a personality disorder too! We can't have someone crazy in power."
---	---

偽の二分法

False dichotomy This is a type of logical fallacy that makes it appear as if there are only two sides or	"Either you support the energy protests or you don't believe in justice."
---	---

偽のバランス

False balance Presenting a debate as having two relatively balanced viewpoints that	"Experts debate the shape of the earth. While scientist Reece Chow has found the earth is spherical, expert Rene Paul argues that the earth is flat."
--	---

「受動的」プレバンкиング

- 受動的な接種
 - テクニックに抵抗するための情報が短い形式（テキスト、グラフィックス、ビデオ）で提供される
 - 制作・実施が比較的容易（例：SNSのポップアップでテキストメッセージを提示、Youtubeで広告のような形式で流す）
 - 没入感が少なく、対話がすくないため影響力が小さい可能性がある

30-90秒



<https://inoculation.science/> から視聴可能



VIDEO EXAMPLE: FALSE DICHOTOMIES

This video example — produced by Jigsaw and Cambridge University — uses culturally relevant examples to help viewers understand and recognize the use of false dichotomies in the spread of misinformation. [View video >](#)

誤情報の共有意図を軽減

INFOGRAPHIC EXAMPLE: COVID-19 CONSPIRACY THEORIES

This UNESCO infographic explains conspiracy theories by using COVID-19 as an example.²²

Limitations

Scalability: 実践者は、異なる種類の誤情報・受け手・プラットフォームで行う場合はパイロットスタディが必要

効果は時間とともに薄れる傾向があり、誤情報に対する耐性を維持するためには、定期的な「ブースター」が必要であり、メディアやデジタルリテラシーのトレーニングも必要

効果検証は主に北アメリカや西ヨーロッパ諸国で実施、異文化間での検証が不足。対象者を考慮して設計する必要がある

Youtubeでのフィールド調査（Google Jigsaw）では、動画ベースの予防接種介入は、情報操作テクニックに対する認識を向上させたものの、他のフィールド調査が不足している。

Prebunking Manipulation Techniques: Emotional Language



感情的な言語や恐怖煽動とは何ですか？

感情は強力な説得の道具です。

研究によると、感情的な言葉、特に恐怖や憤慨などの否定的な感情を喚起する言葉を使用すると、ソーシャルメディアのコンテンツのバイラル性が高まることがわかっています。このような否定的な感情的言葉を操作のために使用することを、時に「恐怖煽動」と呼びます。

ナイジェリア詐欺の文面例

Dearest one

I am writing you this message with tears and sorrow and I know this mail may come to you as a surprise, I am Jennette Ome. The only daughter. My father was a very wealthy cocoa merchant in Abidjan Ivory Coast. My father was poisoned to death by his business associates on one of the outings on a business trip.

My mother died when I was a baby and since then my father took me so special.

涙と悲しみをもってこのメッセージを書いています。このメールが突然届いて驚かれるかもしれません。私はジエネット・オメです。一人娘です。父はコートジボワールのアビジャンで非常に裕福なカカオ商人でした。父はビジネス旅行中の外出の際に、ビジネスパートナーに毒殺されました。
母は私が赤ん坊の時に亡くなり、それ以来父は私を特別に大切してくれました。

I am honourably seeking for your assistance in the following ways: (1) To provide a bank account in which this money would be transferred. (2) To serve is my guardian. (3) To make arrangement for me to come over to your country to further my education.

Note: I am willing to offer you 20% of the total sum as compensation for your effort/ input after the successful transfer of this fund to your nominated bank account. Anticipating hearing from you soon.

Dove, M. (2020). *The psychology of fraud, persuasion and scam techniques: understanding what makes us vulnerable*. Routledge.

Refutation strategies

Intervention Type	Description	Example	Target Outcomes
 Debunking and rebuttals	特定の誤った概念に対処し、偽の信念を否定することを目的とした戦略である。反論は、トピックに関する正確な事実を提示し、科学的否定主義に対処するために使用される(技術的反論)	デバンクは、以下の4つのステップで実施できる。 (1) 事実を述べる (2) 差し迫った誤情報への接触を警告する (3) 誤情報を特定し、なぜそれが間違っているのかを説明する (4) 正しい説明を提供することで事実を補強する。状況によっては(例えば、簡潔な事実が利用できる場合)、ステップ2から始めるのが適切な場合もある。	信念のキャリブレーションと能力: 誤解を招くような誤情報を検出し、抵抗する
 Warning and fact-checking labels	警告ラベルは、特定の情報の信頼性が低いことを個人に明示的に警告する。ファクトチェックラベルは、プロのファクトチェッカーによって割り当てられた信頼性の格付けを示す。	Facebookは、問題のある特定の情報に対して明示的に警告するために、「偽(独立系ファクトチェッカーによると、この情報には事実的根拠がない)」というラベルを付ける。	信念のキャリブレーションと能力: 誤情報やその他の種類の問題のある情報を検出する
 Source-credibility labels	特定の情報源の信頼性を示すラベルは、プロのファクトチェック機関によって割り当てられる	NewsGuardのラベルは、信頼性と透明性の基本的な実践を評価する9つのジャーナリストイックな基準に基づいて、0から100までの信頼性評価で情報ウェブサイトの信頼性スコアと情報を示す。	信念のキャリブレーションと能力: 虚偽または信頼できない情報の情報源を検出する

介入例 (Rebuttal Strategies)

Debunking



Warning and Fact-Checking Labels



Trump on Revamping the Military: We're Bringing Back the Draft

Trump unveiled his plan to 'make the military great again,' saying he intends to reinstate the draft as part of a larger effort to bolster the armed forces.

 Rated False by Snopes.com and PolitiFact

Source-Credibility Labels



REALNEWSRIGHTNOW.COM

Aides Say Trump May Have Accidentally Committed Three Felonies in One Day

Source: [The Debunking Handbook, 2020](#).

https://interventionstoolbox.mpib-berlin.mpg.de/table_examples.html Claude 3 Opusによる仮訳

nature human behaviour

Review article <https://doi.org/10.1038/s41562-024-01681-0>

Toolbox of individual-level interventions against online misinformation

Received: 1 February 2023 Accepted: 5 April 2024 Published online: 13 May 2024

Check for updates

Anastasia Kozireva , Philipp Lorenz-Sprenger , Stefan M. Herzog , Michael A. H. Ecker , Stephan G. Gruber , Ralf Hettwig , Michaela K. H. Scholz , Sven Bartz , Michaela Bartz , Adam J. Berinsky , Cornelia Betsch , John Cook , Lisa K. Fazio , Michael Gees , Andrew M. Guess , Haifeng Huang , Horacio Larrain , Rakesh Malhotra , Folke Pauwels , Jason Reifler , Philipp Schmid , Mark Smits , Brivie Swive-Thompson , Paula Szwarc , Sander van der Linden , and Sam Wineburg 

The spread of misinformation through media and social networks threatens the integrity of society, including the quality of information and the state of democratic life. One way to combat the effects of misinformation focuses on individual-level interventions, equipping policymakers and the public with essential tools to curb the spread and influence of falsehoods. Here we introduce a toolbox of individual-level interventions for reducing harm from online misinformation. Comprising an up-to-date account of interventions and their mechanisms, as well as their strengths and weaknesses, the toolbox provides both a conceptual overview of nine main types of interventions, including their target, scope and examples, and a summary of the empirical evidence surrounding the interventions. Including the

ツールボックスの用途

- 政策立案者など) 誤情報対策やプラットフォーム規制に関する政策議論に役立つ、アクセスしやすい最新の科学的知見を提供。教育プログラムや、自己ナッジを実践したい個人のためのリソースとしても使用。
- 研究者) メタ分析研究、システムティックレビュー、異なる介入の効果を比較する研究の出発点となる。利用可能な証拠の重要なギャップ（例えば、代表性の低い集団や文化、長期的な効果に関する研究の不足など）を浮き彫りにし、将来の研究で取り組むべき課題を示す。

個人レベルのツールに関する緊急の次のステップ

- 中長期的な効果を調査、スケールアップの方法を探る（学校のカリキュラム、アプリ、プラットフォームの協力などを通じて）、教育的背景に関係なく人々に届く介入を構築することなど。

課題

- ツールボックスの介入は、人々の認知と行動に働きかけ、誤情報を共有する傾向や影響を受ける程度を減らすことを目的とする。しかし、個人レベルでの誤情報共有の減少と、プラットフォームレベルでの誤情報の負担と拡散の間には、単純な関連性はない。
- プラットフォームのデータへのアクセスが不足しているため、設計変更による誤情報の減少の可能性は十分に理解されていない。介入の種類や設計変更を大規模にテストするには、このデータへのアクセスが必要。
- 誤情報の拡散に影響を与える要因を理解するには、データへのアクセスと研究者とプラットフォーム間の協力が不可欠。誤情報のような複雑なグローバルな脅威に直面する中で、個人に焦点を当てた介入には限界がある。
- 将来の研究では、介入の関連性と有効性がどの程度維持されているかを追跡するだけでなく、介入に対する信頼と採用の可能性に影響を与える可能性のある環境要因と個人要因を体系的に調査し、既存のツールを改良し、新しいツールを開発する必要がある。

今後の課題

- 誤情報対策
 - 実証研究にもとづく個人レベルでの介入手法が構築されてきている
 - 今後の課題
 - 中長期的効果検証/スケールアップ
 - 介入効果を妨げる現象（例：誤情報持続効果、確証バイアス）の認知メカニズム解明
 - 誤情報のような複雑でグローバルな脅威に対応するには、個人レベルとシステムレベルの両面および交互作用を体系的に理解することが必要
- フィッシングや特殊詐欺への応用可能性に向けて
 - 偽の情報を信じるという心理現象の共通性
 - 「なりすまし」「感情攻撃」「偽の二分法」などの対策は効果があるかもしれない
 - 特殊性
 - スピアフィッシング（個人情報を使用した説得力のある偽のメール）のように特定の個人を狙つくるもへの対策
 - 儲かるという欲を利用した詐欺