Computer Security Symposium 2023
UWS企画① 招待講演

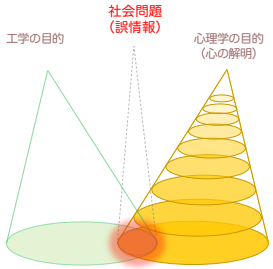
# 認知バイアスと「コグニティブセキュリティー」

名古屋工業大学
田中優子

2023.10.30

## 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　准教授
- 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
- 経歴）
  - 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 特任研究員を経て現職




## 概要

- 「コグニティブセキュリティー」に関する動向
  - 誤情報・偽情報
- 人の認知の特徴
- 今後の課題

## 動向「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

研究開発の俯瞰報告書
システム・情報科学技術分野
（2023年）

## CRDS 研究開発の俯瞰報告書システム・情報科学技術分野（2023年）

研究開発の俯瞰報告書 概要

**システム・情報科学技術分野（2023 年）**

CRDS 研究開発戦略センター

| 技術トレンド | 社会・経済の動向 | 俯瞰 | 重点的に取り組むべき研究開発領域 | 推進シナリオ | ビジョン |
|---|---|---|---|---|---|
| **社会的要請との符合**<br>研究開発活動や科学技術そのものに対する社会的要請の高まり。データに関するプライバシーの考慮やAI技術へのトラスト担保など。<br>**あらゆるもののスマート化・自律化**<br>機器のスマート化が進み、大量のデータの収集と解析が可能になった。ビッグデータと機械学習を組み合わせたサービスが多数生み出された。<br>**あらゆるもののデジタル化・コネクティッド化**<br>無線化・大容量化・グローバル化。ウェブ、スマートフォン、IoT、クラウドなど、社会基盤のデジタル化とコネクティッド化。 | **世界**<br>ロシアのウクライナ侵攻。新型感染症。経済のブロック化。格差、貧困、食料備在化、遺伝化、自然災害リスク。産業・労働構造の変化。<br>**日本**<br>DX推進。サプライチェーンリスクの顕在化。少子高齢化。経済低成長。 | | **デジタル安全保障に対する総合知による取り組み**<br>① デジタル社会におけるトラスト形成<br>② コグニティブセキュリティー<br>③ データ共有<br>**スマート化・自律化の根本である知能の原理探究**<br>④ 知能モデルの解明・探求／身体性に宿る知能<br>⑤ 人間中心インタラクション<br>⑥ バイオハイブリッドロボット<br>⑦ 最適化<br>**サステナブル社会のためのICT基盤**<br>⑧ 社会課題解決に向けたメタバースデザイン<br>⑨ ネットワークのスマート化<br>⑩ 社会デジタルツイン<br>⑪ 社会システムを支えるAIアーキテクチャー | | **社会課題解決と人間中心社会の実現**<br>経済発展と社会問題解決を両立し、誰もが快適で活力に満ちた質の高い生活を送れる社会の実現。ITは人間の判断や決定を補助する道具として働く。<br>**データ駆動型・知識集約型の価値創造**<br>知識・情報・データベース化と統合利活用を実現するプラットフォームやAIにより、データ駆動型・知識集約型の価値創造とDXが加速される。<br>**サイバー世界とフィジカル世界の高度な融合**<br>IoTやCPSが社会生活を支える基盤となる。オープンなサービスプラットフォームなどが実現し、多くの産業が効率化・省エネルギー化する。 |

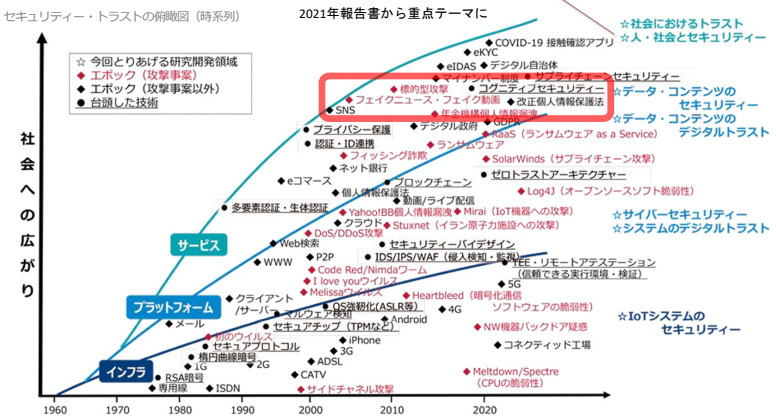
技術：強い技術を核とした弾丸化　産業：強い産業の発展・革新の推進　社会：社会問題の先行解決　基盤：社会基盤を支える横断技術確保

サイバー攻撃の拡大に伴い、これまでのシステムの脆弱性を狙った攻撃に加えて、フィッシングメールによる**人への攻撃**や、フェイクニュースやデマによる**世論の誘導**などが社会に影響を及ぼしている。**情報サービスを利用するユーザー（人）の認識や行動**に着目して、セキュリティー技術単体に加えて、心理学、経済学などの人文・社会科学を含めた学際的アプローチによるセキュリティーを扱う。

研究開発の俯瞰報告書
システム・情報科学技術分野
（2023年）

図2-4-1 セキュリティー・トラストの俯瞰図（時系列）

2021年報告書から重点テーマに




# コグニティブセキュリティー

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

**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

Applied Research Laboratory for Intelligence and Security (ARLIS), University of Maryland

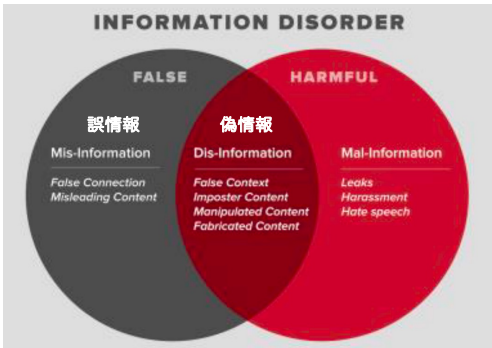
Increase cognitive resilience against malicious influence (悪意のある影響に対する認知的レジリエンスの強化)

批判的思考とメディア・リテラシーの育成
巧妙な悪意ある影響に遭遇した人や組織をリアルタイムで識別・防御するツールの開発

**誤情報・偽情報の影響**

**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。

# 誤情報・偽情報




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

総務省「プラットフォームサービスに関する研究会　第二次とりまとめ」（2022.8）

## 偽情報対策に関する今後の取組の方向性（第二次とりまとめ）

1 自主的スキームの尊重
・民間による自主的な取組を基本とした対策を進めていくとともに、総務省はモニタリングと検証等を継続的に行っていくことが必要

2 我が国における
・我が国における偽
とから、プラットフォーム事業者は、自らのサービス上で流通している我が国における偽情報の問題について適切に実態把握を行い、研究者が分析を行うために必要な情報の無償で情報提供が行われることが望ましい

3．多様なステークホルダーによる協力関係の構築
・「Disinformation対策フォーラム」「Innovation Nippon」等の産学官民の連携の場において継続的に議論・研究が行われることが望ましい

4 プラットフォーム事業者による適切な対応及び透明性・アカウンタビリティの確保
・プラットフォーム事業者は、リスク分析・評価に基づき、偽情報へのポリシーの設定とそれに基づく運用を適切に行い、それらの取組に関する透明性・アカウンタビリティ確保を進めていくことが求められる
・総務省は、これらの取組
行っていくことが必要。と
応が図られているかどう
が必要

5 利用者情報を活用した情報配信への対応
・広告の種類・対応に応じてリスクや問題の差異を分析したうえで、特に、偽情報を内容とする広告の配信やターゲティング技術の適用については、そのリスクを踏まえ、より注意深い対応と、それに伴う透明性・アカウンタビリティ確保が求められる

ファクトチェックの推進

ICTリテラシー向上の推進

6 ファクトチェックの推進
・プラットフォーム事業者・ファクトチェッカー・ファクトチェック推進団体・既存メディア等が連携したさらなる取組が期待される
・「Disinformation対策フォーラム」報告書を踏まえたファクトチェックを持続的かつ総合的に行う主体についての具体的検討が進められることが望ましい
・我が国におけるファクトチェック結果を積み重ねて分析を行うことにより、偽情報の傾向分析やそれを踏まえた対策の検討が行われることが望ましい

7 情報発信者側における信頼性確保方策の検討
・現代のメディア環境に対応した情報の信頼性の確保の在り方について、既存メディア・ネットメディア・プラットフォーム事業者など関係者の間で検討を深めていくことが望ましい
・ミドルメディアを中心とした偽情報の生成・拡散・流通メカニズムに関する実態把握と分析を踏まえ、検討を深めていくことが望ましい

8 ICTリテラシー向上の推進
・偽情報の特徴を踏まえながら引き続き、総務省が開発した啓発教育教材を活用することを含め、ICTリテラシー向上施策が効果的となるよう取り組むことが必要

9 研究開発の推進
・ディープフェイク等に対抗にするための研究開発や事業者の対応が進められることが望ましい

10 国際的な対話の深化
・偽情報に関する政策について国際的な対話の深化を深めていくことが望ましい

・総務省は、違法・有害情報となる偽情報に関するプラットフォーム事業者の取組状況について、先述の違法・有害情報対策に関する記載内容を踏まえて、偽情報への対応に関する透明性・アカウンタビリティの確保に向けて、行動規範の策定及び遵守の求めや法的枠組みの導入等の行政からの一定の関与を具体的に検討することが必要。また、流通状況に関する実態把握と取組に関するモニタリング手法を検討しつつ継続すること必要。

---

POLITIFACT
The Poynter Institute

"Credit card debt is above $1 trillion for the FIRST TIME EVER."

ウォール・ストリート・ジャーナル紙は、アメリカのミサイルがガザの病院爆発を引き起こしたとは報じていない

Social Media
stated on October 17, 2023 in a post on X:

『ウォール・ストリート・ジャーナル』紙は、10月17日のガザ病院の爆発はアメリカのミサイルによるものだと報じた。

The Wall Street Journal reported that the Oct. 17 Gaza hospital explosion was caused by an American missile.

「ウォール・ストリート・ジャーナル紙病院に投下された爆弾はアメリカのMK-84だ」「この爆弾は精密誘導式で、MKファミリーの中で最大のもので、重量は約950キロ」 この投稿は本読発行時点で37万5000回以上閲覧されている。

Facebook posts
stated on October 11, 2023 in una publicación en Facebook:

Benjamin Netanyahu dijo que lanzará una bomba nuclear a Gaza.

ウォール・ストリート・ジャーナル紙が、この爆発はアメリカの弾道ミサイルによるものだと報じたという主張を［FALSE］だと評価する。

The Wall Street Journal didn't report that an American missile caused deadly Gaza hospital blast

IF YOUR TIME IS SHORT

- The Wall Street Journal said it published no such report.
- We found no evidence that any other credible news source reported that the hospital explosion was caused by an American missile.

See the sources for this fact-check

An Oct. 17 attack on a Gaza hospital, estimated to have killed hundreds of civilians being treated and taking refuge there, sparked a blame game between Israel and Palestinian officials, with each accusing the other of responsibility.

Amid this back and forth, claims spread on social media that The Wall Street Journal reported that the blast at the Al-Ahli Arab Hospital was caused by an American missile.

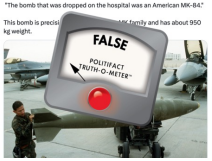
"The Wall Street Journal: "The bomb that was dropped on the hospital was an American MK-84,'" read an Oct. 17 post on X, formerly Twitter. "This bomb is precision-guided, largest in MK family and has about 950 kg weight." The post had been viewed more than 375,000 times as of this publication.

Iran Observer @IranObserverO

The Wall Street Journal:

"The bomb that was dropped on the hospital was an American MK-84."

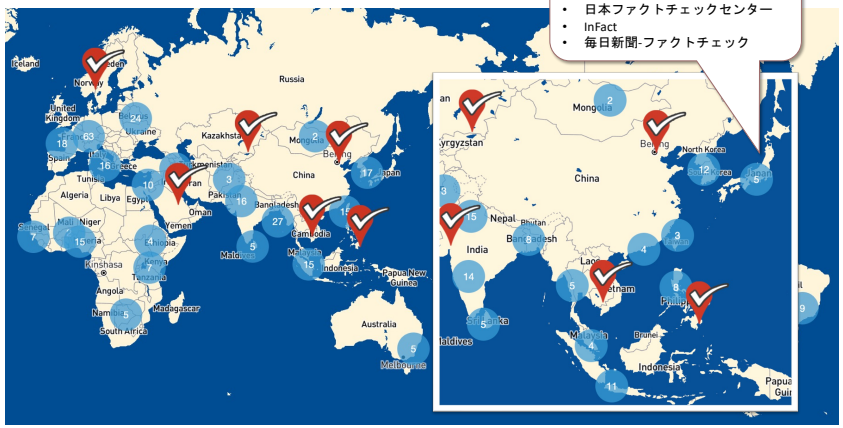
This bomb is precis... MK family and has about 950 kg weight.




---

Duke Reporters' LAB "Global Fact-checking Sites"

**Active (419)**
**Inactive (139)**




日本は５件
- ファクトチェックナビ
- リトマス
- 日本ファクトチェックセンター
- InFact
- 毎日新聞-ファクトチェック

---

# コグニティブセキュリティー

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

**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

Applied Research Laboratory for Intelligence and Security (ARLIS), University of Maryland

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)

批判的思考とメディア・リテラシーの育成
巧妙な悪意ある影響に遭遇した人や組織をリアルタイムで識別・防御するツールの開発

どのような攻撃にどのような影響を受けるか？

**誤情報に対する認知的脆弱性の理解**

**誤情報・偽情報の影響**

**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。

どのような状況で、どのような攻撃に、
どのような影響を受けるか？

なぜ脆弱性が生じるか？

その背後に、どのような認知的性質があるか？

Why Johnny can't …?

認知的脆弱性

Cognition

How human cognition works?

# 人の認知の特徴

白いチームの選手が何回ボールをパスするかを正確に数えてください


## 見えないゴリラ

Youtube
"The Monkey Business Illusion"

- 認知的処理（視覚的注意）の限界
  - すべてに同時に注意を払うことはできない
    - 白チームのパス回数
    - ゴリラの登場
    - カーテンの色の変化
    - 黒チームから１人退場

**Inattentional Blindness**
（不注意による見落とし）

視野の中には入っているものの、注意がむけられていないために見落とす現象

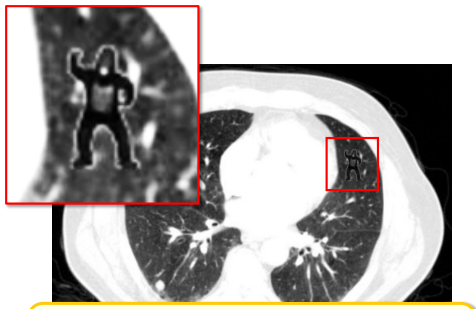
黒T３
白T３

黒T １人退場！
ゴリラ入場！

ゴリラ、胸をたたいてアピール！
カーテン赤色から金色に

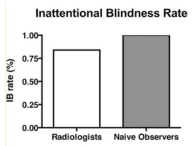
# 見えないゴリラ（ふたたび）

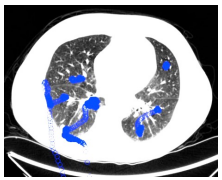
Drew T, Võ ML, Wolfe JM. The invisible gorilla strikes again: sustained inattentional blindness in expert observers. Psychol Sci. 2013 Sep;24(9):1848-53.



放射線科医によるCT画像診断
- 3分で５枚の画像をみて、肺がんの兆候（肺結節）を探す（肺結節の検出は専門医でも55%と難しい課題）
- １枚の画像は平均10個の肺結節を含む
- 肺結節の箇所をクリックするよう求められる
- 最後の画像にゴリラを表示

放射線科医の…
- 83%はゴリラに気づかない（非専門家は100%）
- 50%はゴリラの箇所を注視していても気づかなかった



「見る」という行為には通常 **予測** がある
予測していない刺激には気づかない

# 認知（Cognition）

スティーブン・ピンカー（2013）
心の仕組み＜上＞筑摩書房

心とは複数の演算器官からなる系であり，この系は，われわれの祖先が狩猟採集生活のなかで直面したさまざまな問題，とくに，物，動物，植物，他の人間を理解し，優位に立つために要求されたはずの課題を解決するなかで，自然淘汰によって設計されてきた。

この見方に立つと，心理学はリバースエンジニアリングの一種にほかならない。通常のエンジニアリングにあっては，なにかの目的が先にあって，機械を設計する。逆に，機械が先にあって，なんのために設計されたのかを考えるのがリバースエンジニアリングである。

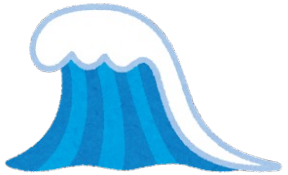
スティーブン・ピンカー
認知心理学，実験心理学

記述的（descriptive）アプローチ

認知容量は非常に限られている
外界をそのまま入力・保持しているわけではない

仮説 → 刺激・介入 →  → 反応を記述 → 設計図（認知モデル）を考察

## Information
膨大な外界情報



予測

す、すべてを受け止めきれない…！

認知的制約

## Cognition
知覚・注意・記憶・言語理解・思考・意思決定など

**認知心理学・認知科学**

**外界に対する人の情報処理プロセスを解明する分野**

## 二重過程理論（Dual Process Theory）

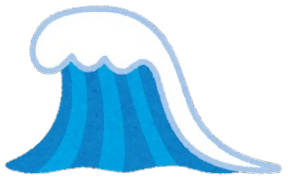
| System 1 | System 2 |
|---|---|
| 進化的に古い | 進化的に新しい |
| 自動的 | 制御的 |
| 速い | 遅い |
| 直観的 | 熟慮的 |
| 努力を要さない | 努力を要する |

真実錯覚効果

確証バイアス

Dual Processing


## Information
膨大な外界情報



す、すべてを受け止めきれない…！

認知的制約

## Cognition
知覚・注意・記憶・言語理解・思考・意思決定など

**認知心理学・認知科学**

**外界に対する人の情報処理プロセスを解明する分野**

# 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.

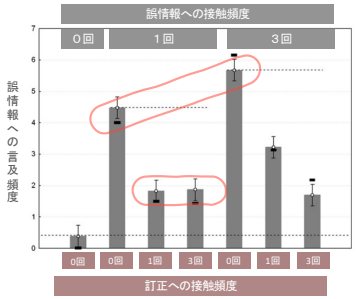
- 繰り返し同じ情報に接触することで、その情報が正しく感じれるようになること。
  - 情報への「親近性（familiarity）」や「処理の流暢性（fluency）」が「正しさ」のシグナルとして利用されるヒューリスティック

「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



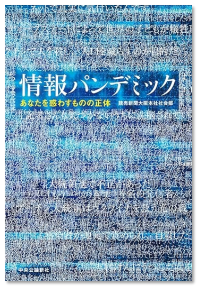

# 「調べれば調べるほどはまる罠」

積極的に調べればよいか？

- 『情報パンデミックーあなたを惑わすものの正体』（読売新聞大阪本社社会部，2022，中央公論新社）

自分で頑張って調べ，ネットの海の中からようやく探り当てた真実。そう信じて疑わなかった情報が，実はデタラメだったー。

取材していると，そのケースに当てはまる人は何人もいた。




「毎日，必死で調べれば，普通の人が知らない優れた治療法にたどり着けると信じていた」


# 確証バイアス

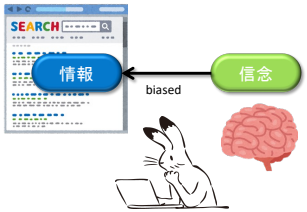
Shin, J., & Thorson, K. (2017). Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media. Journal of Communication, 67(2), 233–255.

- 既有の信念や期待を裏づける証拠を収集しようとする傾向
  - 合致する情報の探索、過大評価
  - 矛盾する情報を探索しない、無視、過小評価

- SNSでの確証バイアス（Shin & Thorson, 2017）
  - 2012年米国大統領選挙
  - ファクトチェックメッセージのリツイート・リプライ行動

| | 民主党 | 共和党 | 無党派 | |
|---|---|---|---|---|
| | Democrats | Republicans | Nonpartisans | $\chi^2$ |
| 民主党に有利 Fact checks advantageous to the Democratic Party | 32,701 (83.06%) | 591 (1.50%) | 6,076 (15.43%) | 22,402.0*** |
| 共和党に有利 Fact checks advantageous to the Republican Party | 2,501 (32.13%) | 3,602 (46.27%) | 1,682 (21.61%) | 360.27*** |

*Note:* Numbers in parentheses indicate row percentages.
***$p < .001$.




# 確証バイアス

- 既有の信念や期待を裏づける証拠を収集しようとする傾向

選択的接触　合致する情報の探索、過大評価
選択的回避　矛盾する情報を探索しない、無視、過小評価

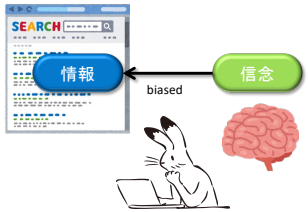
} 誤情報に関する認知的脆弱性としはどちらも重要

一度拡散した誤情報の影響をデバンクするには、訂正情報を避ける「選択的回避」という認知的脆弱性に対策する必要がある

「選択的回避」がそもそも観察できるか？

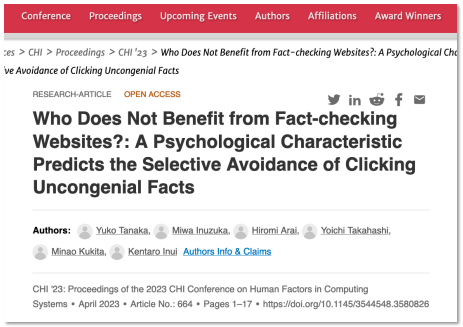
「選択的回避」傾向の個人差はあるか？

訂正情報はどの程度選択的に避けられるのか？



CHI23
Hamburg, Germany | Hybrid
April 23-28, 2023
reCHInnecting

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). Association for Computing Machinery, New York, NY, USA, Article 664, 1–17.


Conference Proceedings Upcoming Events Authors Affiliations Award Winners

:es > CHI > Proceedings > CHI '23 > Who Does Not Benefit from Fact-checking Websites?: A Psychological Cha
ve Avoidance of Clicking Uncongenial Facts

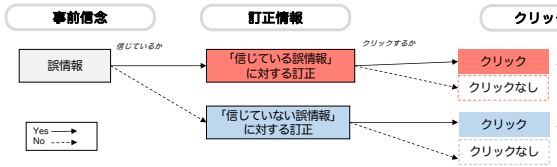
RESEARCH-ARTICLE OPEN ACCESS

Who Does Not Benefit from Fact-checking Websites?: A Psychological Characteristic Predicts the Selective Avoidance of Clicking Uncongenial Facts

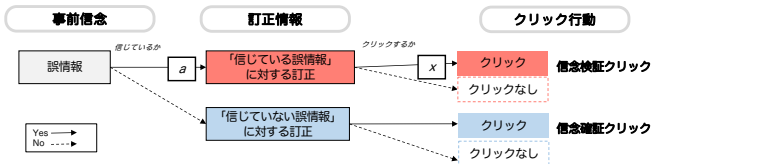
Authors: Yuko Tanaka, Miwa Inuzuka, Hiromi Arai, Yoichi Takahashi, Minao Kukita, Kentaro Inui Authors Info & Claims

CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems • April 2023 • Article No.: 664 • Pages 1–17 • https://doi.org/10.1145/3544548.3580826

田中優子（名古屋工業大　認知科学・実験心理学）
犬塚美輪（東京学芸大　教育心理学）
荒井ひろみ（理研、知能情報学・情報セキュリティ）
久木田水生（名古屋大　技術哲学、技術倫理）
乾健太郎（東北大・MBZUAI　自然言語処理）







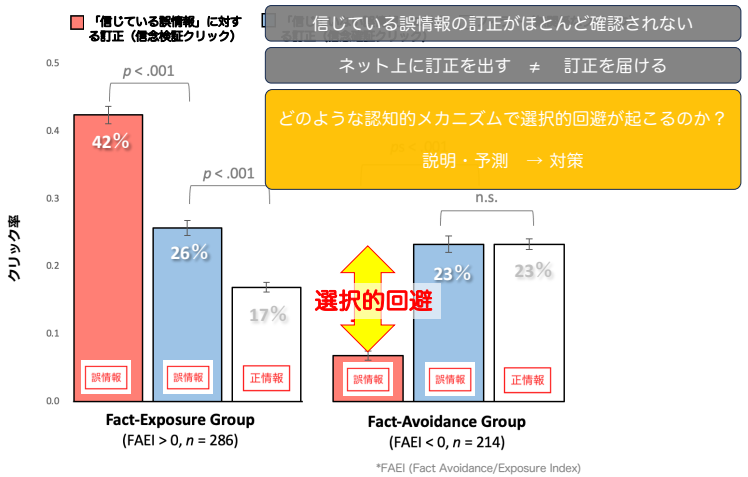

個人ごと…

| | |
|---|---|
| $a$ | 正しいと信じている誤情報の数 |
| $b$ | クリック総数 |
| $n$ | 表示されていリンク総数（$b$の最大値） |
| $x$ | 正しいと信じている誤情報（a）のうちクリックされた数 |
| $EV$ | ランダムに$b$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 |

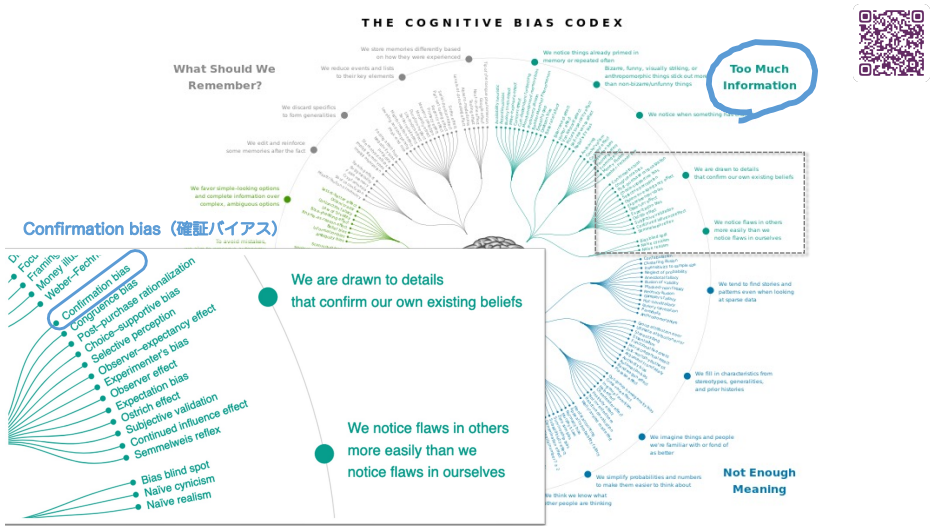
**Fact Avoidance/Exposure Index (FAEI)**

$$FAEI = x - EV$$

$$EV = \sum_{i=0}^{k} \frac{{}_aC_i \times {}_{(n-a)}C_{(b-i)}}{{}_nC_b} \times i$$



# 今後の課題




## 効果的なコグニティブセキュリティーに向けて

1. これまで蓄積されてきた認知に関する知見の共有
2. 情報セキュリティー環境におけるヒューマン・ファクターの仮説・検証・説明・予測
   - 情報環境における認知的な脆弱性は何とトレードオフ関係にあるのか？
     - 「正確な認知処理」と「効率的な認知処理」の両立を可能にする方法はあるか？

Why Johnny can't …?


How human cognition works?

**Cognition**

## 効果的なコグニティブセキュリティーに向けて

1. これまで蓄積されてきた認知に関する知見の共有
2. 情報セキュリティー環境におけるヒューマン・ファクターの仮説・検証・説明・予測
   - 情報環境における認知的な脆弱性は何とトレードオフ関係にあるのか？
     - 「正確な認知処理」と「効率的な認知処理」の両立を可能にする方法はあるか？
3. セキュリティー対策におけるfeasibilityの検討
   - 「気をつけて！」（warning, alart）はどこまで効果的か？
     - ユーザーの認知的特徴（制約）を考慮した対策が必要

**CHI 2023 Opening Keynote**
"The Human Element in Cybercrime and Cybersecurity"


放射線科医はゴリラも見落とさないで！

ご清聴ありがとうございました