第３回インフォメーション・ヘルスアワード『アイデア部門』　応募用紙

○アイデア名
インフォメーション・フードのリスク成分表示―情報添加物・汚染物質・アレルゲン―

○応募者氏名
**　または　グループ名　小林　千晃（東京大学大学院　工学系研究科　技術経営戦略学専攻　吉田塁研究室）**

○解決したい情報環境をめぐる課題

デジタル空間には、個人の認知をハッキングする情報が氾濫している。怒りを煽り、認知バイアスを悪用する情報によって、私たちは容易に説得されてしまう。**攻撃者の巧妙な手法は、認知の脆弱性を攻撃するため、気付きにくく自衛も困難である。**現行の対策はファクトチェックなど事後的な対応にとどまっているが、一度形成された信念を訂正することは難しい。したがって、情報を摂取するその場で危険性を示し、誤った信念の形成自体を防ぐことが重要である。**情報的健康の観点からは、人々が日々「その情報を食べてよいのか」を判断できる仕組みが求められる。**

○アイデアの具体的な内容（どんなもので、どんな人が、どう使うと、課題が解決できるのか）

情報操作テクニックを検知し、情報という食品に含まれる**情報添加物・汚染物質・アレルゲン表示ラベル**を提示する。具体的には、情報操作テクニックを検出する機械学習モデルと大規模言語モデル（LLM）を組み合わせたモデルによって、**過剰なレトリックの使用や、論理的誤謬、認知バイアスを誘発するトリガーを検出し、直感的に理解できる形で可視化する。**

1. <u>検知対象と具体例</u>

- **情報添加物（レトリックの過剰使用）**：煽情的な形容詞の乱用（「史上最悪」「完全無欠」「絶対に」）、二分法的表現（「善か悪か」「我々か彼らか」）など
- **情報汚染物質（論理的誤謬）**：因果と相関の誤り（「ある地域でWi-Fiの普及率が高いほど出生率が低下している。つまりWi-Fiは人間の生殖能力を下げている。」）、ストローマン論法（A「授業でデジタル教材をもう少し取り入れた方がいい。」B「つまり君は先生を全部ロボットに置き換えたいんだな。」）など
- **情報アレルゲン（認知バイアスを誘発するトリガー）**：感情刺激を狙った言葉や画像（怒り・嫌悪・誇りなど）、アンカリング（初期値や極端な数字を提示した印象操作）など

2. <u>実装形態</u>

ブラウザ拡張やSNSアプリ内機能として利用できるように設計する。ユーザーが投稿や記事を開くと、自動的に成分ラベルが表示され、過剰な添加物や汚染物質、アレルゲンの存在がアイコンやマーカーで直感的に示される。また、テキストや、記事・SNS投稿・動画などのURLを貼り付けると、その成分ラベルが表示されるWeb UIも実装する。

3. <u>利用イメージ</u>

- **受け手：**記事やSNS投稿の閲覧時に、拡張機能によるリアルタイム成分表示を確認し、直感的に「注意が必要な情報だ」と理解できる。情報に含まれる危険性に気づいた上で、「その情報を食べるのか」を判断できる。これは加工食品の成分表やアレルギー表示を見ることに対応する。
- **発信者**：投稿前に「有害物質が入っていないか」をチェックし、適切な修正を施すことで、情報の信頼性を高めることができる。これは食品検査に対応する。

この仕組みによって、受け手は直感的に情報の信頼性を判断できるようになり、発信者は質を高める努力を促される。結果として、情報流通全体の質が底上げされ、私たちの情報的健康が守られる。

○アイデアは未発表のものかどうか。すでに「試作」「試行」している場合は、新たに付け加えたいアイデア（ブラッシュアップするポイント）など

未発表である。論理的誤謬やプロパガンダ手法の検出研究は盛んであり、既存のデータセットを活用可能である [1] [2]。本アイデアでは、そうしたデータを用いてトレーニングした機械学習モデルとLLMを組み合わせることで、高精度の「情報成分検知」を実現する。

○アイデアを思いついたきっかけ

コグニティブセキュリティの教育的介入を研究する中で、情報的健康の概念を知ったことが出発点である。その中で情報=食物の比喩に着目した。 **情報的健康の提言では「情報成分表示」の必要性は指摘されてきたが、具体的に情報のどの要素を表示するかの議論は不十分であった**[3] [4]。私の研究分野である接種理論では、個別の偽情報を警告するのではなく、抽象的で汎用的な情報操作手法を事前に学習させる方向へと軸足を移しつつある [5]。この流れを受け、情報操作手法を**添加物・汚染物質・アレルゲン**に見立て、成分表示に用いる発想に至った。
私たちは普段、加工食品を口にするときに「体によさそうか」「食べ過ぎは避けた方がよさそうか」を直感的に判断できる。ファストフードやジャンクフードを摂り過ぎると健康を損なうと誰もが理解しているし、仮に詳しく知らなくとも、添加物や汚染物質は厚生労働省の規制の下で管理され、食品会社も安全性の確保に努めており、成分表示に明確に記載されているため、安心して食べることができる。
だが情報は違う。どれほど安全かを示す指標が存在せず、**魅力的に見える情報ほどレトリックという添加物が過剰に含まれたり、誤謬という有害物質が混入していたりするかもしれない。**しかも私たちは、マスメディアという品質チェックの機能を備えた「食品会社」からだけでなく、出所不明な情報を日常的に大量摂取している。
食品のように国家権威が情報の安全性を管理・規制することは、言論の自由の観点から極めて難しい。だからこそ、私たち一人ひとりが自ら情報の良し悪しを判断する必要があり、その大きな助けとなる仕組みが求められていると考えた。

[1] Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.
[2] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical Fallacy Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
[3] 鳥海不二夫・山本龍彦（代表共同執筆者）ほか. (2024). KGRI Working Papers No.4 共同提言「健全な言論プラットフォームに向けて ver2.1 ―情報的健康を、実装へ」. 慶應義塾大学KGRI. https://www.kgri.keio.ac.jp/docs/XDignity-WP20241029.pdf
[4] 鳥海 不二夫, 山本 龍彦. デジタル空間とどう向き合うか――情報的健康の実現をめざして. 日本経済新聞出版, 2022.
[5] Cook, John, Stephan Lewandowsky, and Ullrich KH Ecker. "Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence." *PloS one* 12.5 (2017): e0175799.

情報操作テクニックをアイコンで表示！

真実を語る者

Translate post

昨日の事件を忘れるな。泣き叫ぶ子ども、崩れた家庭、奪われた未来。データは明白だ。ここ数年で若者のSNS利用が急増し、同じ期間に犯罪件数も上がった。だから犯罪の原因はSNSだ。今こそ選べ。SNSを厳しく規制して子どもを守るか、何もしないで彼らを危険に差し出すか。中立は存在しない。規制に反対する人は子どもの安全より自分の快適さを優先している。

11:30 AM · Sep 29, 2025

View post engagements

成分表

感情刺激

因果と相関の誤り

過度な二分法