「インフォデミック」（情報流行病）の調査 2025/2/20

調査概要・目的

新型コロナウイルス感染症（COVID-19）の流行とともに注目された「インフォデミック」（情報流行病）は、インターネット上で真偽不明な情報が爆発的に拡散する現象を指します。本調査の目的は、**AI 技術**がインフォデミックに与える影響を明らかにし、それを防ぐための**技術的手法**と**社会実装の事例**を分析することです。特に、大規模言語モデル（LLM）やディープラーニングといった AI 技術が誤情報拡散を加速させる要因、および同じ技術を用いた誤情報検出・ファクトチェックの可能性と課題に焦点を当てます。また、AI アルゴリズムのバイアス（目的の不一致や **misalignment**）や、生成 AI のハルシネーション（根拠のない回答生成）がもたらすリスクについて検討し、SNS プラットフォーム等における社会実装の実例も調査します。これにより、AI 時代におけるインフォデミック対策の現状と展望を総合的に明らかにします。

調査手法と検索戦略

本調査では、2020 年以降に発表された関連分野の学術文献を体系的に収集・分析しました。主な手順は以下の通りです。

1. **データベース検索**: Web of Science、Scopus、PubMed、IEEE Xplore、ACM Digital Library、arXiv、SSRN といった主要データベースを用い、以下のキーワードで文献検索を行いました。
    - “Infodemic” や “misinformation” と AI の関連: *“AI and misinformation”, “Infodemic AND AI”, “Fake news detection AI”*
    - **大規模言語モデルや生成 AI**: *“Generative AI”, “Large Language Models (LLMs)”, “Hallucination AI”, “LLM misinformation”*
    - **レコメンドアルゴリズム・フィルターバブル**: *“Recommendation algorithms”, “filter bubble”, “echo chamber”, “algorithmic amplification”*
    - **アルゴリズムのバイアス・目的不一致**: *“algorithmic bias”, “misalignment”, “AI ethics misinformation”*
    - **ファクトチェック・検出技術**: *“misinformation detection”, “AI fact-checking”, “fake news detection deep learning”*
2. **文献選定**: 検索結果から査読付きジャーナル論文、国際会議論文、プレプリント（arXiv 等）、システマティックレビュー・サーベイ論文を中心にピックアップしました。タイトルとアブストラクト段階でインフォデミックと AI の関連性が高いものを一次選別し、さらに出版年（主に 2020 年以降）、査読有無、引用数などを考慮して重要文献を精査しました。また必要に応じ、学術出版社のブックチャプターや信頼できるホワイトペーパー等も参照しました。
3. **分類と分析**: 収集した文献を内容に応じて以下のカテゴリに分類し、それぞれの要点を整理しました。
    - **(1) AI 技術がインフォデミックを加速・拡散させる要因**（例：生成 AI による偽情報大量生成、レコメンド機能によるエコーチェンバー化）

- **(2) AI を用いた誤情報検出・ファクトチェック手法**（ディープラーニングによるフェイクニュース検知モデルなど）
- **(3) 生成 AI（大規模言語モデル）のハルシネーションと誤情報生成リスク**
- **(4) アルゴリズム・バイアス（目的の不一致）がインフォデミックに及ぼす影響**
- **(5) 社会実装事例**（SNS プラットフォームや政府・民間での AI 活用による対策）

以上の戦略に基づき、関連文献をレビューし知見をまとめました。

文献の分類と要点整理

(1) AI 技術がインフォデミックを加速・拡散させる要因

生成 AI による偽情報の大量生成: 近年の大規模言語モデル（LLM）の発展により、真偽不明な情報を人間らしい文章で大量生成することが容易になりました (Frontiers | ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health) (Blessing or curse? A survey on the Impact of Generative AI on Fake News)。例えば、ChatGPT のようなモデルは短時間で膨大なテキストを生成できるため、悪意ある者がそれを利用して高品質なフェイクニュースを無数に作成・拡散することが可能です (Blessing or curse? A survey on the Impact of Generative AI on Fake News)。実際、*ChatGPT など LLM の出現は「AI 駆動のインフォデミック」という新たな脅威を生みうる*と指摘する研究者もいます (Frontiers | ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health)。De Angelis ら(2023)は、LLM が**未曾有の規模で誤情報拡散を加速させる可能性**を警告し、AI 生成テキストを検知する技術開発の緊急性を訴えています (Frontiers | ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health)。また、Loth ら(2024)のサーベイでは、生成 AI により**個人単位でカスタマイズされた虚偽情報**を自動生成することすら可能になっており、フェイクニュース拡散が新たな段階に入ったと述べられています (Blessing or curse? A survey on the Impact of Generative AI on Fake News)。

ソーシャルメディアのアルゴリズム拡散効果: Facebook や YouTube、X（旧 Twitter）など SNS 上では、投稿の表示順序や拡散範囲を左右するレコメンドアルゴリズムが、インフォデミック拡大の一因とされています。アルゴリズムは多くの場合ユーザーの興味関心やエンゲージメント（反応）を最大化するよう最適化されており、その結果、**扇情的・極端な内容**が優先的に露出しやすくなります (The Intersection of AI, Fake News, & Racial Bias | Business Wire Blog)。例えば、近年の研究は「誤情報を含む投稿ほど道徳的憤慨（outrage）を誘発しやすく、それが拡散を促進する」ことを示しました (Our outrage over social media posts helps misinformation spread, study shows)。プラットフォーム側のアルゴリズムはユーザーの強い反応（いいねやシェア、コメント）を高いエンゲージメントとみなし、結果的に**憤慨を呼ぶ虚偽情報がアルゴリズムによって拡散増幅される**現象が確認されています (Our outrage over social media posts helps misinformation spread, study shows)。McLoughlin ら(2024)は、このような**アルゴリズムの誤情報増幅効果**について「プラットフォームの設計上の意図せざる結果であり、早急な対策が必要」と述べています (Our outrage over social media posts helps misinformation spread, study shows)。一方で、近年 YouTube に関する大規模実験研究では、短期的なレコメンドの偏り（フィルターバブル）の影響は限定的とする結果も報告されており、アルゴリズムの影響評価にはさらなる検証が必要との指摘もあります (New Study Challenges YouTube's Rabbit Hole Effect on Political Polarization | Computational Social Science Lab)。しかし総じて、**エコーチェンバー/フィルターバブル**現象（同質の情報ばかりが集まる環境）によりユーザーが偏った情報世界に閉じこもるリスクは広く認識されています

(The Intersection of AI, Fake News, & Racial Bias | Business Wire Blog)。AI アルゴリズムが意図せず形成するこれらの情報環境が、誤情報の信憑性を高めインフォデミックを加速させる要因となり得ます。

その他の AI 活用による拡散手法: 加えて、ソーシャルボット（自動投稿プログラム）も AI 技術で高度化しています。ディープラーニングを組み合わせたボットは人間らしい文章や振る舞いで SNS 上の議論に参加し、誤情報を広めたりトレンド操作したりすることが可能です。画像・動画分野でも GAN などの生成 AI により、**ディープフェイク**と呼ばれる偽動画が作成され、著名人が偽の発言をしている映像などが拡散して社会的混乱を招く例が出てきました (The Intersection of AI, Fake News, & Racial Bias | Business Wire Blog)。このように、多様な AI 技術が「情報の捏造と拡散」を容易にし、インフォデミックの火種を各所に生み出しているのが現状です。

(2) AI を用いた誤情報検出やファクトチェックの手法と精度

ディープラーニングによる自動フェイクニュース検知: 誤情報対策として、近年は AI を活用した自動検出技術が発展しています。従来はルールベースやメタデータ分析、人手のファクトチェックが主でしたが、それらは爆発的な情報量を前に**スケーラビリティに限界**がありました (Fact-checking information generated by a large language model can decrease news discernment)。これに対し、BERT など Transformer ベースのディープラーニングモデルによる文章分類がフェイクニュース検知に応用され、高い精度を報告する研究が増えています。実際、最新のレビューによれば、**大規模言語モデル(LLM)を用いたフェイクニュース検知は従来手法よりも高い精度と効率を示す**ケースが多いとされています (A Survey on the Use of Large Language Models (LLMs) in Fake News)。LLM は文脈理解や微妙な表現の違いを捉える能力が高いため、単純なキーワードマッチや従来モデルよりも誤情報を検知しやすい利点があります (A Survey on the Use of Large Language Models (LLMs) in Fake News)。例えば、ニュース記事の真偽判定タスクで従来モデルの精度を LLM が上回った報告もあり、SNS 上の偽アカウント検知（プロフィール解析・ネットワーク解析）にも LLM が応用されています (A Survey on the Use of Large Language Models (LLMs) in Fake News) (A Survey on the Use of Large Language Models (LLMs) in Fake News)。加えて、生成 AI 自体を活用し、ChatGPT のようなモデルに文章のファクトチェックを行わせる試みも行われています (Fact-checking information generated by a large language model can decrease news discernment)。総じて、AI は大規模なデータをリアルタイムに分析できる強みから、**インフォデミック対策の有力なツール**になると期待されています。

ファクトチェック AI の有効性と課題: 一方で、AI による自動ファクトチェックには課題も残ります。Indiana 大学の DeVerna ら(2023)による実験では、**LLM が生成したファクトチェック情報をユーザに提供しても、必ずしもユーザのニュース真偽の識別力向上につながらない**ことが報告されました (Fact-checking information generated by a large language model can decrease news discernment)。この研究では、ChatGPT に政治ニュース見出しの真偽判定と解説をさせ、その情報が被験者に与える影響を測定しました。その結果、LLM が**誤って「真のニュース」を偽情報とラベル付けした場合にユーザの正しいニュースへの信頼が低下**したり、AI が自信なさげな判断を示した「偽ニュース」に対して逆に信憑性が上がってしまうケースが確認されました (Fact-checking information generated by a large language model can decrease news discernment)。つまり、AI ファクトチェックは完璧ではなく、誤判断がかえって認知を混乱させるリスクがあります。また、ユーザが任意で AI ファクトチェックを見る設定では、**真偽を問わず情報を拡散しようとする傾向**が強まることも示され、場合によっては AI が意図せず誤情報拡散を助長する可能性も指摘されています (Fact-checking information generated by a large language model can decrease news discernment)。このように、AI 検出モデルの精度向上とともに、**誤検出の影響を最小化する設計**（例：明確な不確実性の提示や人間監督との協調）が求められています。加えて、検出モデル自体がブラックボックスになりがちで、なぜその判断に至ったかの説明性（Explainability）も課題です。これはファクトチェック結果の説得力にも関わるため、近年は AI の説明

可能な誤情報検出や、人間と AI の協調（human-AI collaboration）による検出フレームワークの研究も進められています (Combating Misinformation in the Age of LLMs: Opportunities and Challenges)。例えば専門家をループに入れて AI の検出結果を評価・改善する枠組みや、AI が提示した根拠を人間が検証するプロセスなどが提案されています (Combating Misinformation in the Age of LLMs: Opportunities and Challenges)。総じて、AI は大量の誤情報を**高精度にふるいにかけるフィルター**として有望ですが、過信は禁物であり、人間の監督と組み合わせて信頼性を確保する必要があると考えられます。

(3) 生成 AI のハルシネーションと誤情報生成リスク

ハルシネーション（幻覚）とは: 生成 AI、特に大規模言語モデルは高品質な文章を作成できますが、しばしば**事実無根の内容をもっともらしく生成してしまう**現象が報告されています。これを「ハルシネーション」と呼び、AI が自信ありげに**架空の事実やソースをでっち上げる**問題として注目されています (Frontiers | ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health)。LLM は次の単語を確率的に予測して文章を紡ぐため、知識の穴や訓練データに存在しない問合せに対して、それらしい回答を**でっち上げてしまう**ことがあります (Combating Misinformation in the Age of LLMs: Opportunities and Challenges)。Papageorgiou ら(2024)のサーベイによれば、この種のハルシネーションは特に細部情報（日付や人名、数字など）に関して起こりやすく、主要なメッセージは正しくても**細部が虚偽**というケースが生じ得るといいます (Combating Misinformation in the Age of LLMs: Opportunities and Challenges)。ユーザに悪意がなくとも、「短いニュース記事を書いて」といったプロンプトに対し、LLM はそれらしく見えるが完全な創作であるニュースを生み出してしまうことがある (Combating Misinformation in the Age of LLMs: Opportunities and Challenges)。つまり、**生成 AI は意図せず誤情報を生み出すリスク**を常に内包しているのです (Combating Misinformation in the Age of LLMs: Opportunities and Challenges)。

ハルシネーションがもたらす影響: 生成 AI の回答が権威あるものと受け止められる場合、そのハルシネーションは誤情報として広まり得ます。例えば ChatGPT が架空の論文やデータを引用した場合、それを鵜呑みにした人々によって二次拡散され、AI 発のデマが増幅される危険があります (Frontiers | ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health)。医学や公共政策などセンシティブな領域で誤った情報が広まれば、健康被害や社会的混乱を招く恐れも指摘されています。また、ハルシネーション問題は **AI への信頼性低下**にも繋がります (AI hallucinations can pose a risk to your cybersecurity - IBM)。IBM の報告では、「AI の幻覚は誤情報拡散につながる上、それが積み重なると人々が AI から得られる情報全般を信用しなくなる」可能性が指摘されています (AI hallucinations can pose a risk to your cybersecurity - IBM)。このように、生成 AI のハルシネーションは、一方でインフォデミックの火種となり（誤情報の新たな供給源となり）、他方で AI 技術への社会的信頼も揺るがす**二重のリスク**を孕んでいます。

悪意ある利用と高度化する偽情報: さらに深刻なのは、**悪意ある主体が生成 AI を意図的に利用して誤情報を量産するケース**です。LLM に対して巧妙なプロンプトを与えれば、特定の陰謀論やプロパガンダを裏付ける偽記事、偽レポートを次々と作成できます (Combating Misinformation in the Age of LLMs: Opportunities and Challenges)。ChatGPT などは利用規約上、有害なコンテンツ生成を抑制する仕組みがありますが、プロンプトを工夫して迂回する「プロンプトエンジニアリング」により禁止された出力を得ようとする試みも確認されています。また、LLM を使って既存の偽情報を**「言い換え（パラフレーズ）」**することで検出をすり抜ける手法も現れています。Das ら(2023)の研究では、ニュース記事の偽情報を LLM で言い換えると、既存のフェイクニュース検出器が見抜きにくくなることが実証されました (Fake News Detection After LLM Laundering: Measurement and Explanation)。すなわち、**LLM による誤情報のロンダリング**が可能であり、検出 AI とのいたちごっこを激化させる恐れがあります (Fake News Detection After LLM

Laundering: Measurement and Explanation)。以上のように、生成 AI のもたらす誤情報リスクは単なる偶発的なハルシネーションに留まらず、悪用による大規模キャンペーンや検出逃れなど多岐にわたります。研究コミュニティでは、これらに対抗するための**安全性向上策**（例えば、出力に対する検証プロセス組み込みや、検出しやすい特徴をあらかじめ盛り込んだ「検出容易型 LLM」の開発 (Frontiers | ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health)）が模索されていますが、決定打はまだありません。Chen & Shu (2024)も、**LLM 時代の新たな脅威に対抗する研究課題**として「ハルシネーションの低減」「LLM の安全性向上」「LLM 生成コンテンツの検知」を挙げており (Combating Misinformation in the Age of LLMs: Opportunities and Challenges)、今後の重要分野となっています。

(4) アルゴリズム・バイアス（misalignment）がインフォデミックに及ぼす影響

アルゴリズム・バイアスと目的の不一致: AI システムは訓練データや設計目的に起因する**バイアス（偏り）**を内包することがあります。インフォデミック文脈でのアルゴリズム・バイアスとは、AI が特定の種類の情報を過剰に（または過少に）拡散・検出してしまう偏りや、システムの目標設定が社会的な望ましさと食い違う（misalignment）ことを指します。例えば、前述の SNS 推薦アルゴリズムは「ユーザーエンゲージメント最大化」という目的関数に沿って最適化されていますが、これは**「真実性の最大化」ではない**ため、この目的の不一致が誤情報拡散を助長する構造的要因となります (Our outrage over social media posts helps misinformation spread, study shows) (The Intersection of AI, Fake News, & Racial Bias | Business Wire Blog)。プラットフォームにとって望ましい指標（閲覧時間やクリック率）がそのまま社会的に望ましい指標ではない典型例であり、AI の目標設定のミスマッチ（misalignment 問題）が表面化したものと言えます。結果として、**刺激的な虚偽情報ほどアルゴリズムに評価される**という偏りが生まれ、インフォデミックに拍車をかけることになります (The Intersection of AI, Fake News, & Racial Bias | Business Wire Blog)。研究コミュニティでは、この問題に対し「アルゴリズムの説明責任と透明性（Algorithmic accountability）」を高めることや、エンゲージメント以外の健全性指標（信頼できる情報源かどうか等）を組み込む提案がなされています。

データやモデルに内在するバイアス: また、AI モデル自体が学習に用いたデータセットの偏りを反映し、特定の誤情報に鈍感であったり、逆に無害な情報を誤って有害と分類したりするケースもあります。例えば、多言語対応が不十分なモデルは主要言語以外での誤情報を見落とす偏りがありえますし、人種・民族に関するステレオタイプを学習したモデルが**特定コミュニティに関する誤情報を強化**してしまう可能性も指摘されています ((PDF) The Influence of Social Media Algorithms on Racial and Ethnic Misinformation: Patterns and Impacts)。Lucas(2024)は**「SNS アルゴリズムが人種・民族的な誤情報を無意識のうちに拡散強化し、偏見を増幅する」**危険性を分析しており ((PDF) The Influence of Social Media Algorithms on Racial and Ethnic Misinformation: Patterns and Impacts)、アルゴリズム・バイアスが社会的少数派に不利益を与える問題にも注意が必要です。さらに、誤情報検出 AI 自体にもバイアスのリスクがあります。事実とみなす基準が西洋中心であれば、他文化圏の正当な表現を誤って偽情報扱いする恐れがあり、ファクトチェック AI の公平性も課題です。

対策の方向性: アルゴリズム・バイアス問題に対処するため、近年は AI 倫理の観点から**公正なアルゴリズム設計**や**バイアス検知・低減手法**が模索されています (Algorithmic bias detection and mitigation: Best practices and policies ...)。具体的には、トレーニングデータを多様化しバランスを取ること、モデルの判断根拠を人間が監査できるようにすること、フィードバックループによる偏りの増幅を防ぐ機構（例えば一定以上偏った情報ばかり推薦しないよう制御する）などが検討されています。また、プラットフォーム企業にアルゴリズムの影響評価を義務付ける動き（欧州のデジタルサービス法におけるリスク評価など）も始まっています。要するに、**AI システムと社会との価値のず**

れ（misalignment）を是正し、インフォデミックを助長しない設計・運用を目指す取り組みが徐々に進みつつあります。

(5) 社会実装事例（SNS プラットフォーム等における AI の活用）

世界の主要な SNS プラットフォームや各国政府・民間団体は、AI 技術を活用したインフォデミック対策を実装し始めています。その代表的な事例と傾向を以下にまとめます。

- **Facebook/Meta**: Facebook では AI による有害コンテンツ検出システムを大規模に導入しており、COVID-19 パンデミック時には誤情報に対する警告ラベル付与や削除を強化しました。Meta 社の発表によれば、2020 年～2021 年にかけて数千万件規模の COVID-19 関連の虚偽情報ポストを AI と人間のレビューの組み合わせで削除したとされています（※Meta 公式リリースより）。また、多言語対応の誤情報検出モデル（XLM-R など）を開発し、世界各国の選挙やワクチンに関するデマ対策に適用しています。もっとも、Facebook 上で活動する膨大なユーザコミュニティを網羅するのは困難であり、AI で検知しきれない誤情報の拡散（例：クローズドなグループ内でのデマ拡散）は依然課題です。

- **X（旧 Twitter）**: Twitter 社（現 X 社）も機械学習を用いたスパム・ボットアカウント検出や、有害な投稿の自動フラグ付けを行っています。2020 年米大統領選挙や COVID-19 に際しては、誤情報と判定されたツイートに閲覧注意のラベルを付け拡散を抑制する措置を実施しました。AI がポリシー違反の可能性がある投稿を検知すると、人間のモデレーターが精査し、必要に応じ削除・アカウント凍結を行う体制です。さらに Twitter は**「コミュニティノート（旧 Birdwatch）」**と呼ばれるユーザ参加型のファクトチェック機能も展開し、アルゴリズム＋人間コミュニティのハイブリッドで誤情報の拡散抑制に挑戦しています。ただし、近年の運営方針の変化により、コンテンツモデレーションの厳格さについては議論があるところです。

- **YouTube**: 世界最大の動画共有プラットフォームである YouTube も、アルゴリズムの調整によってインフォデミック対策を図っています。具体的には、2019 年頃より**陰謀論や医療デマなど「ボーダーライン」の動画を推薦アルゴリズムで拡散しにくくする**変更を加えたと報告されています。この結果、問題あるコンテンツの視聴時間が米国内で大幅に減少したとのデータも公開されました（YouTube 公式発表より）。加えて、信頼できる情報源（公式ニュースや公衆衛生機関）の動画を検索結果や推薦で上位表示する施策も取られています。もっとも、アルゴリズムの詳細は非公開であり、外部からその実効性を検証することは難しい状況です。一方、Pennsylvania 大学の研究 (New Study Challenges YouTube’s Rabbit Hole Effect on Political Polarization | Computational Social Science Lab)では、YouTube 風の実験環境を用いてアルゴリズムを操作しユーザの反応を測定する試みがなされました。それによると、短期的な偏向推薦はユーザの態度を急激には変えず、むしろユーザ自身の選好が強く影響する結果となりました (New Study Challenges YouTube’s Rabbit Hole Effect on Political Polarization | Computational Social Science Lab)。プラットフォーム上のインフォデミック対策には、アルゴリズム変更だけでなくユーザ教育との両輪が必要であることが示唆されます。

- **その他の取り組み**: **Google** は「Fact Check Explorer」や検索結果における信頼性インジケータ表示など、検索アルゴリズム側での誤情報対策を進めています。また **WHO（世界保健機関）**は各国政府と協調し、「インフォデミック・マネジメント」と称して AI を用いたオンライン上のデマ検知ダッシュボード構築や、ソーシャルリスニングによる早期警戒システムの導入を呼びかけています。民間では、**Fake News 対策のスタートアップ**や検証専門の NPO が AI 技術を活用し、疑わしいコンテンツの検出・通報を行うサービスも登場し

ています。例えば、不正な画像や動画を検出するための画像取引チェッカー、ニュース記事の信頼度をスコアリングするブラウザ拡張機能などが開発・提供されています。

これら社会実装事例から明らかになるのは、**AI 単独で完璧な解決策とはなっていないものの、人手では対処しきれないインフォデミックの規模に対抗する上で AI は不可欠な役割**を果たしつつあるという点です。プラットフォーム各社は試行錯誤を重ねながら、AI 検出と人間の判断を組み合わせた多層的アプローチを模索しています。

考察・分析

以上の調査結果を踏まえると、AI とインフォデミックの関係は**「諸刃の剣」**とも言える状況が浮かび上がります。一方では、生成 AI やパーソナライズアルゴリズムが誤情報拡散を前例のない速度と規模で加速させている現実があります。LLM が人間には不可能なペースで偽情報テキストを大量生産でき、SNS アルゴリズムが人々の関心を惹くがゆえに虚偽でも刺激的な情報を増幅してしまう——この構図自体は今後も続く可能性が高いでしょう。特に生成 AI の高度化により、テキストだけでなく画像・音声・動画の分野でもディープフェイク技術が進歩すれば、インフォデミックの素材となるコンテンツは質量ともに飛躍的に増大し得ます。**「AI 駆動のインフォデミック」**という言葉が示すように、AI 技術は新たな情報災害の火種ともなりうるのです (Frontiers | ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health)。

他方で、AI はインフォデミックと戦うための有力な武器も提供しています。ディープラーニングによる自動検知システムは、人手では不可能な速度で膨大なコンテンツをモニタリングし、違反やデマをふるい落とすことを可能にしました (A Survey on the Use of Large Language Models (LLMs) in Fake News)。事実、主要プラットフォームは AI 無くして現在のコンテンツ監視を維持できないと言っても過言ではありません。また、ファクトチェック AI や言語モデルを活用した検証支援は、情報の真偽判定プロセスを効率化するポテンシャルがあります。ただし、現状の研究が示すように、**AI によるファクトチェックはそれ自体が完璧ではなく、人間の判断を補完するものでしかない**点には注意が必要です (Fact-checking information generated by a large language model can decrease news discernment)。AI の誤りは新たな誤情報や誤解を生むリスクがあり、完全自動化には限界があります。このため、**“AI + 人間”の協調**が重要との認識が広がっています (Combating Misinformation in the Age of LLMs: Opportunities and Challenges)。例えば AI が広範囲をスクリーニングし疑わしい情報を絞り込み、その上で人間のファクトチェッカーが精査する、といった二段構えの体制です。今後はこうした協調モデルの有効性を高める研究がますます求められるでしょう。

また、アルゴリズム・バイアスや AI の misalignment の問題は、技術的というより**社会的・倫理的な課題**として浮上しています。AI 開発者やプラットフォーム企業にとって、エンゲージメント至上主義からの脱却や、多様な価値観を尊重した設計が求められています。一部では「倫理的なデザイン」や「価値に整合した AI」の研究が進み、例えば RLHF（人間のフィードバックによる強化学習）で言語モデルを有害出力しにくく調整する、といった試みも行われています (Frontiers | ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health)。しかし、商業サービスにおけるアルゴリズム変更は収益や利用体験ともトレードオフになるため、実装には慎重さが必要です。プラットフォームが透明性を高め第三者の監査を受け入れることや、利用者側にアルゴリズム選択肢を与える（フィルターバブルから抜け出す）ような取り組みも議論されています。

さらに、生成 AI のハルシネーション対策としてはモデルアーキテクチャや学習プロセスの改良が模索されています。例えば、事実知識を明示的に参照する仕組み（外部知識ベースとの連携）や、モデルが自信度を出力し低信頼の回答には警告を付けるといった機能強化です。また、**水印技術**（生成 AI が生成したコンテンツに見分けがつくマ

ーカーを埋め込む）も提案されており (Frontiers | ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health)、AI 生成コンテンツを後から検知することで二次拡散を防ぐアイデアもあります。ただし、水印は完全ではなく変造や部分引用で失われる可能性もあり、万能策ではありません。

総合すると、**技術的手法と社会的アプローチの両面から対策を講じる必要**が明確です。技術面では「生成しない・広げない・見逃さない」を目標に、AI の安全な開発（有害出力を極力抑える）、アルゴリズムの健全化（拡散ロジックの見直し）、検出技術の高度化が進められるでしょう。社会面では、プラットフォーム規制やリテラシー教育、ファクトチェック機関の充実、そして国際的な協調が不可欠です。AI はグローバルに影響を与えるため、各国の枠を超えたガバナンスも議論が始まっています。例えば EU では包括的な AI 規制案（AI Act）で高リスク AI システムに対する透明性要求が盛り込まれ、インフォデミック対策にも寄与する可能性があります。日本でも総務省などが情報信頼性向上の検討を進めています。

本調査で浮かび上がった文献群は、**AI の進展がインフォデミック問題に新たな局面をもたらしている**ことを強調しています。一筋縄ではいかない複雑な状況ですが、同時に AI を上手く利活用することでこの問題に立ち向かう展望も示されています。

結論・今後の展望

AI 技術とインフォデミックの関係について、2020 年代以降の最新研究を調査・分析しました。大規模言語モデルやレコメンドアルゴリズムは、誤情報の生成・拡散を加速しうる強力なツールとなる一方で、それらを抑止・検出するための手段も提供しています。生成 AI によるフェイクニュース大量生産や、アルゴリズムのエコーチェンバー化といった負の側面が明らかになる一方、ディープラーニングを活用した自動検知システムの有望性や、AI と人間の協調によるファクトチェックの可能性も確認されました。

結論として、**AI はインフォデミック問題の「原因にして解決策の一部」**であり、その影響力の大きさゆえに適切なガバナンスが極めて重要です。AI モデルの安全設計やアルゴリズムの説明責任確保、そして国際的な標準策定が急務となっています。また、インフォデミック対策は技術だけで完結しないため、教育や政策との連携が欠かせません。今後の展望としては、より高度な **AI 安全技術（ハルシネーション抑制や水印埋込など）の研究開発**や、**マルチモーダルな誤情報検出**（画像・動画の偽造検知）への AI 活用拡大が挙げられます。さらに、生成 AI を逆に活用した**「偽情報の自動収集・要約・反駁生成」**といったソリューションも考えられ、すでに初期的な研究が始まっています。社会実装面では、プラットフォーム企業による透明性レポートの公開や第三者検証、国際協調によるデマ拡散ネットワークの解明などが進むでしょう。

インフォデミックは 21 世紀の情報社会が直面する重大な課題であり、AI 技術の発展とともにその様相を変え続けています。本調査で得られた知見が、今後の研究・政策立案に資する基礎として役立つことを期待します。AI を「凶器」ではなく「盾」として活用し、健全な情報環境を守るための努力が今まさに求められていると言えるでしょう。

参考文献リスト

[1] De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). **ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health.** *Frontiers in Public Health, 11*, 1166120

(Frontiers | ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health) (Frontiers | ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health).

[2] Deverna, M. R., Yan, H. Y., Yang, K.-C., & Menczer, F. (2023). **Fact-checking information generated by a large language model can decrease news discernment.** *arXiv preprint* arXiv:2308.10800 [Preprint]. (Later published in *PNAS*, 2024) (Fact-checking information generated by a large language model can decrease news discernment) (Fact-checking information generated by a large language model can decrease news discernment).

[3] Papageorgiou, E., Chronis, C., Varlamis, I., & Himeur, Y. (2024). **A Survey on the Use of Large Language Models (LLMs) in Fake News.** *Future Internet, 16*(8), 298 (A Survey on the Use of Large Language Models (LLMs) in Fake News) (A Survey on the Use of Large Language Models (LLMs) in Fake News).

[4] Chen, C., & Shu, K. (2024). **Combating misinformation in the age of LLMs: Opportunities and challenges.** *AI Magazine, 45*(3), 354-368 (Combating Misinformation in the Age of LLMs: Opportunities and Challenges) (Combating Misinformation in the Age of LLMs: Opportunities and Challenges).

[5] Loth, A., Kappes, M., & Pahl, M.-O. (2024). **Blessing or curse? A survey on the Impact of Generative AI on Fake News.** *arXiv preprint* arXiv:2404.03021 (Blessing or curse? A survey on the Impact of Generative AI on Fake News).

[6] McLoughlin, K. L., et al. (2024). **Misinformation exploits outrage to spread online.** *Science, 366*(6466), eadl2829 (Our outrage over social media posts helps misinformation spread, study shows) (Our outrage over social media posts helps misinformation spread, study shows).

[7] Das, R. K., & Dodge, J. (2023). **Fake News Detection After LLM Laundering: Measurement and Explanation.** *arXiv preprint* arXiv:2501.18649 (Fake News Detection After LLM Laundering: Measurement and Explanation).

[8] Lucas, W. (2024). **The Influence of Social Media Algorithms on Racial and Ethnic Misinformation: Patterns and Impacts.** (Research Gate Preprint) ((PDF) The Influence of Social Media Algorithms on Racial and Ethnic Misinformation: Patterns and Impacts).

[9] Business Wire. **AI's Influence on Spreading Misinformation and Disinformation.** *BusinessWire Blog*, Oct 5, 2023 (The Intersection of AI, Fake News, & Racial Bias | Business Wire Blog).

[10] Hu, E., Knox, D., Liu, N., Savaş, Y., et al. (2023). **Short-term exposure to filter-bubble recommendation systems has limited polarization effects: Naturalistic experiments on YouTube.** *PNAS, 120*(50), e2318127122 (New Study Challenges YouTube's Rabbit Hole Effect on Political Polarization | Computational Social Science Lab).